GÖRSEL SORU CEVAPLAMA PROBLEMİNDE BAĞLAMSAL VEKTÖRLERİN PERFORMANS ANALİZİ

Özlem HAKDAĞLI



T.C.
BURSA ULUDAĞ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GÖRSEL SORU CEVAPLAMA PROBLEMİNDE BAĞLAMSAL VEKTÖRLERİN PERFORMANS ANALİZİ

Özlem HAKDAĞLI

0000-0002-3637-4309

Doç. Dr. Metin BİLGİN
(Danışman)

YÜKSEK LİSANS TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

BURSA – 2022

Her Hakkı Saklıdır

ÖZET

Yüksek Lisans Tezi

GÖRSEL SORU CEVAPLAMA PROBLEMİNDE BAĞLAMSAL VEKTÖRLERİN PERFORMANS ANALİZİ

Özlem HAKDAĞLI

Bursa Uludağ Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç. Dr. Metin BİLGİN

Görsel soru cevaplama (GSC) çalışmaları, görsel imgeleri anlamlandırmanın yanında tutarlılık sağlamayı hedeflemektedir. GSC problemi, görsel bir imge ile bu imgeye sorulan soru arasındaki bağlantıyı ele almaktadır. Ele alınan bağlantının yorumlanması ve çözümlenmesi, sorulan soruya beklenen cevabın görsel içerisinden elde edilmesini sağlar. Çözümleme işlemini gerçekleştirmek için görsel imgelerin matematiksel düzlemde temsil edilmesi gereklidir. Bu temsiller vektör olarak adlandırılır.

Görsel vektörlerin elde edinimi aşamasında, bu çalışmada ImageNet verisi ile eğitilmiş olan Xception ve Inception-Resnet-V2 modelleri kullanılmıştır. Modeller derin konvolüsyonel ağlara ve tekrarlayan katman yapısı sayesinden görsel veriden yüksek doğruluk ile vektör temsili elde edilmektedir. Görsel vektör temsili, GSC problemi için yeterli değildir. Görsele sorulan sorunun matematiksel düzlemde temsili gerekmektedir. Metinsel verilerin temsili diğer adı ile kelime gömmeleri, ön eğitimli modeller olan Word2Vec, Kelime Temsili için Global Vektörler (Global Vectors for Word Representation, GloVe) ve FastText yöntemleri ile anlamsal bağlamdan bağımsız şekilde elde edilmektedir. Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (Bi-directional Encoder Representations from Transformers, BERT), inşa edilmiş olduğu çok başlı ilgi yapısı ile kelimelerin arasındaki alt bağlamı öğrenmekte ve temsil etmektedir. Bu çalışma ile sorulan sorunun anlamsal bütünlüğünü güçlendirmek için BERT bağlamsal vektörleri uyarlanmıştır. Çalışmanın sonuçları değerlendirildiğinde BERT yöntemi; Word2Vec, GloVe ve FastText yöntemlerinden daha yüksek doğruluk oranlarına ulaştığı görüldü. Böylelikle, literatüre yeni girmiş olan BERT bağlamsal vektörleri yönteminin GSC problemindeki başarısı gösterilmiştir.

Anahtar Kelimeler: Görsel Soru Cevaplama, Derin Öğrenme, Doğal Dil İşleme, Kelime Gömmeleri, Bağlamsal Kelime Vektörleri

2022, viii + 60 sayfa.

ABSTRACT

MSc Thesis

PERFORMANCE ANALYSIS OF CONTEXTUAL VECTORS IN VISUAL QUESTION ANSWERING PROBLEM

Özlem HAKDAĞLI

Bursa Uludağ University

Graduate School of Natural and Applied Sciences

Department of Computer Engineering

Supervisor: Assoc. Prof. Dr. Metin BİLGİN

Visual question answering (VQA) studies aim to provide consistency as well as to make sense of visual images. The VQA problem deals with the connection between a visual image and the question asked to that image. The interpretation and analysis of the discussed link ensures that the expected answer to the question asked is obtained from within the picture. In order to perform the analysis process, it is necessary to represent the visual images on the mathematical plane. These representations are called vectors.

In the acquisition phase of visual vectors, Xception and Inception-Resnet-V2 models which are trained with ImageNet data were used. The models obtain vector representation from visual data with high accuracy due to deep convolutional networks and residual layer structure. Visual vector representation is not sufficient for the VQA problem. The mathematical representation of the question asked to the image is required. Representation of textual data, also known as word embeddings, can be obtained independently of the semantic context with the pre-trained models Word2Vec, Global Vectors for Word Representation (GloVe) and FastText, Bi-directional Encoder Representations from Transformers (BERT)learns and represents the sub-context between words with the multi-headed attention structure it is built on. BERT contextual vectors were adapted to strengthen the semantic integrity of the question asked in this study. When the results of the study were evaluated, it was seen that the BERT method achieved higher accuracy rates than the Word2Vec, GloVe and FastText methods. Thus, the success of the BERT contextual vectors method, which has just entered the literature, in the GSC problem has been demonstrated.

Key words: Visual Question Answering, Deep Learning, Natural Language Processing, Word Embedding, Contextual Word Vectors

2022, viii + 60 pages.

TEŞEKKÜR

Tez çalışmasını yaparken zorlu süreçteki yönlendirmeleri, bilgi ve tecrübeleri ile destek olan değerli danışman hocam Sayın Doç. Dr. Metin BİLGİN'e teşekkürlerimi sunarım.

Tez çalışması sırasında bilgileri ve yardımlarıyla yanımda olan doktora öğrencisi İskender Ülgen OĞUL'a teşekkür ederim.

Tez çalışması sırasında desteğini esirgemeyen Teracity Yazılım'a teşekkür ederim. Tez süresi boyunca manevi olarak desteklerini esirgemeyen aileme teşekkür ederim.

Özlem HAKDAĞLI

…/…/…….

İÇİNDEKİLER

Sayfa

ÖZET....i
ABSTRACT....ii
TEŞEKKÜR....iii
İÇİNDEKİLER....iv
SİMGELER ve KISALTMALAR DİZİNİ....v
ŞEKİLLER DİZİNİ....vii
ÇİZELGELER DİZİNİ....viii
1. GİRİŞ....1
2. KAYNAK ARAŞTIRMASI....5
3. MATERYAL ve YÖNTEM....18
3.1. Öğrenme Tabanlı Yöntemler İçin Kelime Vektörleri Temsilleri....18
3.1.1. Word2Vec....19
3.1.2. GloVe....20
3.1.3. FastText....22
3.1.4. BERT....24
3.2. Görsel Vektörlerin Elde Edinimi İçin Konvolüsyonel Ağlar....27
3.2.1. Xception....27
3.2.2 Inception-ResNet v2....29
3.3. Görsel Soru Cevaplama Veri Seti VQA....30
3.4. Keşifsel Veri Analizi....32
3.5. Yöntem Tasarımı....35
3.6. Word2Vec, GloVe, FastText modelleri Hiper parametreler ve Keras Tuner....38
3.7. BERT modeli Hiper parametreler ve Keras Tuner....38
3.8. Doğruluk Metriği....39
4. BULGULAR ve TARTIŞMA....40
5. SONUÇ....55
KAYNAKLAR....57
ÖZGEÇMİŞ....60

SİMGELER ve KISALTMALAR DİZİNİ

Simgeler	Açıklama
T	Cümlenin toplam uzunluğu
A	Dikkat
H	Gizli
L	Katman
v	Sözlük boyutu
ck	Oluşturulan aday cevap kümesi

Kısaltmalar	Açıklama
BERT	Bi-directional Encoder Representations from Transformers
Bi-LSTM	Bi - long Short-Term Memory
CAG	Candidate Answer Generation
DAQUAR	DAtaset for QUestion Answering on Real-world images
DOC-VEC	Document Visual Question Answering
ELMO	Embedding from Language Models
FAP	Final Answer Prediction
FVTA	Focal Visual-Text Attention
GRU	Gated Recurrent Unit
GPT	Generative Pre-trained Transformer
GloVe	Global Vectors for Word Representation
GSC	Görsel Soru Cevaplama
LSTM	Long Short-Term Memory
MS COCO	Microsoft Common Object in Context
ReLU	Rectified Linear Unit
RCNN	Region Convolutional Neural Network
ST-VQA	Scene Text Visual Question Answering
VGG	Visual Geometry Group
VQA	Visual Question Answering

Çeviriler	Açıklama
Candidate Answer Generation	Aday Cevap Oluşturma
Smart-Shuffle	Akıllı Karıştırma
Region Convolutional Neural Network	Bölge Tabanlı Evrişimli Sinir Ağı
Bi - Long Short-Term Memory	Çift yönlü Uzun Kısa Süreli Bellek
Embedding from Language Models	Dil Modellerinden Gömmeler
Validation	Doğrulama
Document Visual Question Answering	Doküman Görsel Soru Cevaplama
Epoch	Döngü
Train	Eğitim
Early Stopping	Erken Durdurma
Convolutional Neural Networks	Evrişimsel Sinir Ağı
DAtaset for QUestion Answering on Real-world images	Gerçek Dünya Görüntüleri Üzerinde Soru Cevaplama İçin Veri Seti
Latent	Gizli
Visual Geometry Group	Görsel Geometri Grubu
Visual Question Answering	Görsel Soru Cevaplama
Fine-tuning	İnce Ayar
Cascade Answering Model	Kademeli Yanıtlama Modeli
Gated Recurrent Unit	Kapı Özyinelemeli Geçitler
Loss	Kayıp
Global Vectors for Word Representation	Kelime Temsili için Global Vektörler
Bag-of-Words	Kelime Torbası Yöntemi
Focal Visual-Text Attention	Odak Görsel-Metin Dikkat
Learning Rate	Öğrenme Oranı
Rectified Linear Unit	Rektifiye Edilmiş Lineer Birimleri
İmage Captioning	Resimden Metin Oluşturma
Scene Text Visual Question Answering	Sahne Metin Görsel Soru Cevaplama
Dropout	Seyreltme
Final Answer Prediction	Son Cevap Tahmini
Bias	Taraf
Bi-directional Encoder Representations from Transformers	Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri
Long Short-Term Memory	Uzun Kısa Süreli Bellek
Generative Pre-trained Transformer	Üretken Önceden Eğitilmiş Transformatör
Batch	Yığın

ŞEKİLLER DİZİNİ

Sayfa

Şekil 1.1. Görsel Soru Cevaplama problemine örnek 1
Şekil 3.1. Bert ve diğer transformatör temelli mimarilerin karşılaştırması 24
Şekil 3.2. Bert ön eğitim ve ince ayar mimarileri temsili 25
Şekil 3.3. BERT mimarisi kelime girdi gösterimi 26
Şekil 3.4. Detaylı Xception mimarisi 28
Şekil 3.5. Residual Inception blok mimarisi 29
Şekil 3.6. Inception-Resnet-V2 mimarisi sıkıştırılmış gösterimi 30
Şekil 3.7. VQA-2 veri seti dengelenmiş soru ve cevap analizi 32
Şekil 3.8. VQA-2 veri setinden alınan örnek görsel ve soru cevap ikilisi 32
Şekil 3.9. VQA-2 veri içindeki resimlerin ortalama boyut bilgisi 33
Şekil 3.10. VQA-2 veri setindeki soruların kelime bazında dağılım grafiği 34
Şekil 3.11. Çift yönlü LSTM mimarisi yapısı gösterimi 37
Şekil 4.1. Xception-Word2Vec modeli eğitim A) Doğruluk B) Kayıp grafikleri 47
Şekil 4.2. InceptionRV2-Word2Vec modeli eğitim A) Doğruluk B) Kayıp grafikleri 47
Şekil 4.3. Xception-GloVe modeli eğitim A) Doğruluk B) Kayıp grafikleri 48
Şekil 4.4. InceptionRV2-GloVe modeli eğitim A) Doğruluk B) Kayıp grafikleri 48
Şekil 4.5 Xception-FastText modeli eğitim A) Doğruluk B) Kayıp grafikleri 49
Şekil 4.6. InceptionRV2-FastText modeli eğitim A) Doğruluk B) Kayıp grafikleri 49
Şekil 4.7. Xception-BERT modeli eğitim A) Doğruluk B) Kayıp grafikleri 50
Şekil 4.8. InceptionRV2-BERT modeli eğitim A) Doğruluk B) Kayıp grafikleri 50
Şekil 4.9. Modelin doğru cevap verdiği örnek 53
Şekil 4.10. Modelin yanlış cevap verdiği örnek 53

ÇİZELGELER DİZİNİ

Sayfa

Çizelge 3.1. VQA-2 veri seti soru tümcelerinin sayı ve dağılımları 34
Çizelge 3.2. VQA-2 veri seti cevapların sayı ve dağılımı..... 35
Çizelge 3.3. Kelime Vektörleri modelleri derin öğrenme modeli hiper parametreler ... 38
Çizelge 3.4. BERT derin öğrenme modeli hiper parametreler 39
Çizelge 4.1. Xception-Word2Vec döngü başına model performans değerleri 40
Çizelge 4.2. InceptionRV2-Word2Vec döngü başına model performans değerleri 41
Çizelge 4.3. Xception-GloVe döngü başına model performans değerleri 42
Çizelge 4.4. InceptionRV2-GloVe döngü başına model performans değerleri 43
Çizelge 4.5. Xception-FastText döngü başına model performans değerleri..... 44
Çizelge 4.6. InceptionRV2-FastText döngü başına model performans değerleri..... 45
Çizelge 4.7. Xception-BERT döngü başına model performans değerleri..... 46
Çizelge 4.8. InceptionRV2-BERT döngü başına model performans değerleri..... 46
Çizelge 4.9. Xception ve kelime vektörlerine dayalı doğruluk çıktıları. 51
Çizelge 4.10. Inception-Resnet-V2 ve kelime vektörlerine dayalı doğruluk çıktıları. ... 51

1. GİRİŞ

Günümüz dünyasında teknoloji katlanan bir hızda gelişmekte ve ilerlemektedir. Gerek özgür yazılım gerekse özel girişimlerin teknolojiye gerçekleştirdiği katkılar daha da hızlı ilerleme olanağı sunmaktadır. Teknolojiye erişebilen insan sayısı ile teknolojinin gelişim hızı doğru bir orantıya sahiptir. Bu orantı aynı şekilde teknolojiye erişen insan sayısı ile oluşturdukları verilerin katlanarak artmasıda gözlemlenmektedir. Gelişmeler göz önüne alındığında veri depolama sistemleri ve veri ambarlarında elde edilen güncel başarılar, veri depolama ve dizinleme sorununu ortadan kaldırmıştır. Alan ve bağlantı sınırlamaları olmadan günlük olarak oluşturulan veriler, bireyin mahremiyetini koruma kuralları çerçevesinde depolanmaktadır. Depolanmış olan ham görsel ve metinsel veriler gerek kamusal gerekse özel kâr amacı gütmeyen oluşumlar için büyük bir önem taşımaktadır. Ham verilerin uygun yöntemler ile işlendiği süreçler sonucunda ortaya çıkan sonuçlar birçok problem için çözüm kaynağı olmaktadır.

İnsanlar; görme, duyma, hissetme gibi aktiviteleri aynı anda yapabiliyor ve aynı anda birçok farklı reflektörden gelen uyarıları işleyebilme yetenekleri bulunmaktadır. Bunlara bağlı olarak çeşitli karmaşık sorulara cevaplar üretebiliyorlar. Yapay zeka, insan beynini taklit ettiği bilindiğine göre, makinelerinde bunları yapabilmesi sağlanabilir prensibinden yola çıkarak Görsel Soru Cevaplama (GSC) problemi ortaya çıkmıştır. Son yıllarda bu konu üzerinde birçok araştırma yapılmaktadır. Akademik araştırma problemi olarak ortaya çıkan bu problem, kalite denetleme bölümünde ürünün kontrol edilmesine çözüm olarak kullanılabilir. Bu çalışma kapsamında, görsel ve metinsel verileri ele alan çalışmaların kesişim noktası olan GSC problemini ele almıştır. GSC problemine Şekil 1.1'de örnek verilmiştir.



Soru: Yakınlarda orman var mı?
Cevap: evet

Şekil 1.1. Görsel Soru Cevaplama problemine örnek

GSC çalışmaları, teorik olarak düz bir akış fikri olmasına rağmen oldukça karmaşık ve çoklu modellerin bir arada kullanıldığı sürecin ortak paydada birleşmesi gerektirmektedir. GSC, bir görseldeki imgelerin bir soru ile ilişkilendirilip, hedeflenen cevabın ortak yöntem ile elde edildiği problemdir. Gerçekleştirilen tez çalışmasında GSC problemindeki metin temsillerinin kalitesini ele almaktadır. Geleneksel çözümler, GSC probleminde Kelime Temsili için Global Vektörler (Global Vectors for Word Representation, GloVe) (Pennington , Socher , & Manning, 2014) kelime temsillerini temel alarak çözmüştür. Ancak 2017 yılında sunulan yeni bir ilgi mekanizması (Vaswani , et al., 2017) sayesinde kelime vektörlerinin bağlamsal bütünlüğünün en iyi şekilde yansıtılmasının önü açılmıştır. Bu gelişme aynı dönemde Google beyin takımının yeni bir dil modeli olan Transformatörlerden Çift Yönlü Kodlayıcı Temsilleri (Bi-directional Encoder Representations from Transformers, BERT) (Devlin , Chang , Lee , & Toutanova, 2019) yaklaşımını ortaya çıkarmıştır.

Öncül çalışmalar göstermiştir ki GSC alanındaki en büyük sıkıntılardan biri modellerin eğitildiği veri seti eksikliği olmuştur. Bu bağlamda ilk olarak Gerçek Dünya Görüntüleri Üzerinde Soru Cevaplama İçin Veri Seti (DAtaset for QUestion Answering on Real-world images, DAQUAR) (Malinowski & Fritz, 2014) sunulmuştur. Veri seti içerisindeki taraf sorunu ve dengesizlik fark edilmiş ve DAQUAR-uzlaşma adı verilen yeni bir veri seti türetilmiştir. Öncül veri setleri soru, cevap ve görsel arasındaki ilişki bakımından oldukça dengesiz bir yapıdadır. Bu durum elde edilen sonuçlardan, modellerin davranış biçimine kadar bir etki alanına sahiptir. Yash Goyal ve ekibi VQA-1 veri setini neredeyse yeniden inşa ederek set içindeki dengesizliği ortadan kaldırıp, taraf sorununa çözüm sunmuştur. Sunulan veri seti, gelişmeler neticesinde VQA-2 olarak adlandırılmıştır (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017). Devam eden çalışmalarda model ve veri seti özelinden çıkılarak, öğrenme modellerinin performansı ve hangi durumlarda en yüksek sonuca ulaştıkları ele alınmıştır. 3 000 grafik işlemci saati eşdeğerindeki çalışma ve kapsamlı model-vektör araştırmaları sonucunda elde edilen en yüksek hiper parametreler sunulmuştur (Teney , Anderson , He , & Hengel, 2018). Yapılan hiper parametre ve mimari özelindeki keşifsel çalışma, halen güncelliğini korumakta ve birçok

çözümün temelini oluşturmaktadır. GSC çalışmalarının dikkat çekmesi ve incelikle ele alınması yan alanlarıda oluşturmuştur.

İncelenen çalışmalarda gözlemlendiği üzere yaklaşımları iki ana başlık altında toplamak mümkündür. İlk olarak, çalışmaların büyük çoğunluğu görsel bilgi çıkarımı ve bilgi çıkarımı esnasında ilgi metotlarına ağırlık vermiştir. Standart konvolüsyonel ağlarla başlayan süreç, gelişen modeller ile ilgi mekanizmaları ve ardından çift yönlü ilgi tabanlı yaklaşımlara evrilmiştir. İkinci başlık ise yeni bir problem ve bu probleme ait veri seti sunulması olarak ele alınabilir. Bunlar soyut sahnelerde GSC ya da doküman görselleri üzerinde GSC çalışmaları örnek olarak verilebilir. Üçüncü bir sınıf olarak ele alınmasada Damien Teney'in çalışması (Teney , Anderson , He , & Hengel, 2018) keşifsel analiz bulgularını sunmuştur ve kendisinden sonraki birçok çalışma için mihenk taşı olmuştur.

İncelemeler sonucunda, literatürdeki metinsel verilerin keşifsel analizi eksikliği dikkat çekmiştir. Gerçekleştirilen çalışmaların büyük çoğunluğunda, ön eğitimli kelime vektörü olan GloVe (Pennington , Socher , & Manning, 2014) kullanılmıştır. Ön eğitimli kelime vektörleri büyük bir kelime külliyatı içerisinde, kelimelerin birbiri ile olan komşuluk ilişkilerine göre elde edilmektedir. Bu sayede, her bir kelime için yoğun olarak adlandırılan vektörel bir temsil üretilmektedir. Ancak 2017 yılında Vaswani'nin (Vaswani , et al., 2017) sunmuş olduğu ilgi mekanizması, kelime temsilleri alanında yeni bir çağ açmıştır. Yeni ilgi modeli, Google beyin takımının gerçekleştirdiği çalışma sonucu ile yeni bir dil modeli olan BERT'in ortaya çıkmasını sağlamıştır. BERT (Devlin , Chang , Lee , & Toutanova, 2019) ön eğitimli kelime vektörlerinden farklı olarak her bir kelimenin diğer kelimeler ile olan ilişkisini tek tek ele alarak cümle içerisindeki anlamsal bağlamı en iyi şekilde sunabilmektedir.

Bu tez çalışması, GSC problemini geleneksel kelime vektörleri ve bağlamsal kelime vektörlerinin performans değerlendirmesi yapılmıştır. Yapılan tez çalışmasından GSC probleminin ele alınmasında, metinsel vektörlerin değerlendirilmesi eksikliğine dikkat çekilmiştir. Çalışmanın ilk aşamasında geleneksel olarak kullanılan Görsel Geometri Grubu (Visual Geometry Group, VGG) mimarisinin dışına çıkılarak artık bloklar ve konvolüsyonel ilgi mekanizmanlarına dayanan vektörler Xception (Chollet, 2017) ve Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) modelleri ele

alınmıştır. İkinci aşamada ise ön eğitimli kelime vektörleri olan Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013), GloVe (Pennington , Socher , & Manning, 2014) ve FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) kelime vektörleri ile her bir model eğitilmiştir. Elde edilen sonuçların görsel özelliklerden bağımsızlığını göstermek için Xception (Chollet, 2017) ve Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) modellerini resimlerden özellik çıkarımı için kullanılmıştır. Sonra aynı düzen tekrar edilerek BERT mimarisine dayanan kelime vektörlü GSC modelleri elde edilmiştir. Modeller VQA-2 veri seti üzerinde eğitilmiştir. Veri seti birleştirilerek literatürde kabul görmüş olan 80 – 10 – 10 oranların bölünerek eğitime alınmıştır. Elde edilen sonuçlar grafik ve çizelgeler halinde sunulmuştur.

2. KAYNAK ARAŞTIRMASI

Bu bölümde, tez alanında daha önce yapılmış ve literatürde yer edinmiş çalışmalar ele alınmıştır. Her bir çalışmanın anlatılış ve ele alınış sırası kronolojik sıra baz alınarak yapılmıştır.

Malinowski, önceki çalışmalara nazaran bu çalışmada görsel ve dilbilimsel özelliklerin beraber kullanıldığı bir yaklaşımı benimsemektedir. Yaklaşımlarına göre her modalite yorumlanmalı ve ortak bir şekilde temsil edilmelidir. Bu durum, cevabın aslında görsel ve metnin beraber yorumlanmasına bağlı olduğunu gerekçe ederek açıklanmaktadır. Çalışmasında kendisinin ortaya sunmuş olduğu görsel sahne veri seti olan DAQUAR (Malinowski & Fritz, A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input, 2014) veri setini genişleterek bir ortak fikirlilik temeline dayandırılan DAQUAR-uzlaşma veri setini oluşturmuştur. Gerçekleştirdiği modelde uçtan uca sunduğu çözüm ile görsel sahnelerin anlamlandırılmasında metin sekanslarının yorumlanmasını temel almıştır. Bu durum, verilen bir soru ve resim üzerinden cevabın tahmin edilmesi temeline dayandırılması demektir. Bu senaryoda sorular birden çok cevaba sahip olabilirler dolayısıyla çalışma devamlı olarak problemi bir cevap seti olan kelimeler özeline bölmektedir. Çoklu kelimeleri tahmin edebilmek için, problemi sözlükten sıralı kelime tahmini şeklinde ele almaktadır. Çalışmaları konvolüsyonel ağlar ve Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) (Hochreiter & Schmidhuber, 1997) yapısı üzerine inşa edilmiştir. LSTM (Hochreiter & Schmidhuber, 1997) değişken uzunluktaki sekansları öğrenmede oldukça büyük başarı elde etmektedir. Çalışmada soru ve cevaplar vektör ile temsil edilmiş ve sonraki aşamada bu temsiller daha düşük boyutlu uzaya indirgenmiştir. Bu gömülme işlemi ortak lineer gizli öğrenme görmeleri ile sağlanmıştır. Eğitim zamanında, görsel soruya dayalı bilgiyi, önceden tahmin edilmiş olan cevap kelimeleri ile güçlendirerek devam eder. Açıklamak gerekirse, her LSTM (Hochreiter & Schmidhuber, 1997) adımında bir önceki adımda elde edilen cevap bilgisi, bir sonraki zaman adımına entegre edilmektedir. Gerçekleştirdikleri testlerin daha anlamlı sonuç vermesi için insan tarafından yapılan tahmin skorları ile karşılaştırmışlardır. Sonuçlarına göre kendi modelleri %89 değerinde başarı yakalamıştır (Malinowski , Rohrbach , & Fritz, 2015).

Akira Fukuki ve ekibi GSC problemini, bu alanın eksikliklerine odaklanarak ele almışlardır. Görsel ve metinsel bilgileri içeren problemlerin büyük veri setleri ile çözülmesi gerekliliği artık kabul görmüştür. Ancak bu problem, bir noktada görsel ve metinsel bilgilerin bir araya getirilerek kesiştirilmesi yöntemini gerektirmekteydi. Geleneksel olarak kullanılan yöntemlerde her bir görüntü ve metin vektörleri standart bir nokta çarpımı, aynı düzlemde toplanması ya da aynı düzlemde ortalama alınması ile birleştirilmekteydi. Fukuki ve ekibine göre geleneksel yöntemler çoklu modalite bilgilerini ve problemin genelleştirilmiş çözümünü göz ardı etmektedir. Sundukları çözümde metinsel ve görsel bilgilerin çoklu modalite sıkıştırılmış çift doğrusal havuzlama yöntemi ile birleştirilmesi ele alınmıştır. Öncüllerinde bu problem kesişen modalite havuzlaması ile çözülmekteydi. Kesişen havuzlama, iki modalite arasındaki benzerlikleri yakalamayı ele alır. Lakin benzerlik tabanlı sistemler, aradaki farkı yakalasa bile oldukça yüksek bir hesaplama gerektirmektedir. Bu hesaplama problemdeki vektör boyutları ile doğru orantılı olarak artmaktadır. Sundukları çoklu modalite kompakt çift doğrusal havuzlama ise iki vektörün dışsal çarpımını ele alarak bir model öğrenir. Bu sayede standart çift doğrusal havuzlama yöntemlerinden farklı olarak yüksek boyutlu veriler ve uzun sekanslar için problemin öğrenilmesine olanak sağlar. Sundukları yöntem, vektörlerin çarpımı esnasında elde edilen çarpım sonuçlarından öğrenen bir model ile verileri daha küçük bir boyuta haritalandırma (yansıtma) yapabilmektedir. Bu aynı zamanda konvolüsyon teoreminde zaman alanına göre konvolüsyon ve frekans domaininde eleman tabanlı çarpıma eşittir. Bu sayede konvolüsyon bir frekans domaini şeklinde yazılabilir. Sundukları yöntemi Resnet152 modeli ile 448 x 448 boyutlarındaki resimlerin 2 048 boyutunda temsil edildiği vektörler ve kategorik değişkenlerin ikili olarak temsil edilmesi (one-hot-encoding) yöntemi ile 1 024 boyutunda temsil edilen metinsel veriler üzerinde test etmişlerdir. Ortaya çıkan çözüm, GSC problemlerinde çoklu modalite kompakt çift doğrusal havuzlama yönteminin, çift ilgi yöntemi ile kullanıldığında en iyi performansı verdiğini göstermiştir (Fukui , et al., 2016).

Peng Zang ve ekibi GSC araştırma alanını resimsel sahnelerden diğer adı ile soyut sahnelerden, bilgi çıkarımı olarak ele alıyor. Resimsel, soyut sahneler iki şekilde önem katmaktadır. İlk olarak yüksek seviyede görsel metinsel soru cevaplama görevine

odaklanmayı teşvik eder. İkincil olarak, metinsel bilgilerin kontrol edilerek görsel temsillerin daha iyi anlaşılması için dengelenmiş bir veri seti sunar. Çalışmalarında sundukları fikre göre bazı GSC problemlerinin oldukça kolay olduğunun, kurulan ağların veriden ezberleyerek soruna çözüm sunduğunu iddia etmişlerdir. Örnek olarak, Bağlamda Microsoft Ortak Nesneleri (Microsoft Common Objects in Context, MS COCO) (Lin , et al., 2014) veri setinde zürafaların genel olarak çimlerin üzerinde olduğunu ve kurulan ağların çim ve zürafa bilgisini ezberlediğini öne sürmüşlerdir. Bir diğer örnekte, '*hangi renk?*' sorusuna, kurulan ağların verdiği cevap %23 oranında '*Beyaz'* olmaktadır. Bu duruma ikili sorular yani cevabı doğru veya yanlış olan durumlarında katıldığını öne sürülmüştür. Öyle ki bir kişi tüm sorulara herhangi bir bilgi çıkarımı yapmadan '*Evet'* cevabı vermesi durumunda %68 oranında başarı elde ettiği belirtilmiştir. Bu bilgiler ışığında, ideal olarak dil bağlamında zorlayıcı olmayı hedeflerken algoritmanın resim bilgisinide kullanmasını teşvik edecek bir çözüm sunmaya çalışmıştır. Bu hedef çerçevesinde, tarafsız bir veri seti oluşturmuşlardır. Oluşturulan veri seti, evet ya da hayır cevapları yerine açık uçlu olarak davranmaktadır. Soyut sahnelere dayalı veri setinin oluşturulması için resim kütüphanesi oluşturulmuştur. Bu kütüphanede 20 adet insan modeli, 8 farklı ifade, 99 adet nesne, 31 adet hayvan ve bunların kombinasyonunu içermektedir. Veri seri oluşturulması esnasında 50 000 adet soyut sahne üretilmiştir. Her bir sahne üç adet soru ile ilişkilendirilmiştir. Bu sorular alanında uzman kişiler tarafından topluluk yardımıyla doğruluğu kesin cevaplar ile bütünleştirilmiştir. Son aşamada veri seti üç farklı kategoriye bölünmüştür; cevabı evet olanlar, hayır olanlar ve numerik olanlar. Sunulan veri seti ile modelin doğru cevaba ulaşması için görsel öğeleri çok iyi öğrenmesi amacına ulaşılmıştır. Bu çalışmanın ana katkısı var olan soyut ikili GSC veri seti dengelenerek birbirini tamamlayan sahneler oluşturulmuştur. İkincil olarak sorunun içeriğini özetleyen demet yaklaşımını kullanmıştır. Bu sayede sorunun cevabının varlığı resim içinde doğrulanabilmektedir (Zhang , Goyal , Summers-Stay, Batra , & Parikh, 2016).

Yash Goyal ve ekibi öncül çalışmalardan çok farklı bir yol izlemiştir. Sundukları çalışma, ön tanımlı dil bilgisini kullanan GSC çalışmalarına bir karşılık niteliğindedir. Çalışmanın amacı GSC çalışmalarında geleneksel olan metinsel bilgi yerine görsel bilginin ön plana

çıkartılıp, modelin karar verme mekanizmasında etkileyici rol oynaması hedeflenmiştir. Dilin karmaşık yapısı görsellik ve dilbilim arasında kesişen araştırmaları oldukça zorlaştırmaktadır. Ancak son dönemde yapılan çalışmalar, dilbilimin kendisi tek başına çok güçlü bir ön bilgi sunmaktadır. Bazı durumlarda ön tanımlı dil bilgisi özellikleri modeli yanlış sonuçlara yönlendirebilmektedir. Bu durum, makine öğrenmesi modellerinin resimdeki veriyi çok iyi şekilde anlayarak gerekli görevi yerine getirdiği algısını oluşmaktadır. Ancak sadece dilbilimde bulunan bilgilerin ortaya çıkarılmasından başka bir şey değildir. Bu çalışmada ön tanımlı dil bilgisine karşı bir yaklaşım sunulmuştur ve GSC çalışmalarındaki resmi anlamanın önemi arttırılmıştır. Hedeflenen amaca ulaşmak için GSC setindeki verilerden dengelenmiş bir alt veri seti oluşturulmuştur. Oluşturulan alt veri setinde dilin içerisindeki taraf durumu en aza indirilmiştir. Yeni veri setinin oluşturulma aşamasında, katılımcılardan her bir resim için birden çok cevap ataması istenmiştir. Sundukları hipotez ise bu dengelenmiş veri seti görsel soru cevaplı modellerinin, görsel bilgiler ve anlamlara daha fazla odaklanması gerektiğini ulaşması hedeflenmiştir. Sonuçta bir soru ve iki farklı cevaba sahipse aynı zamanda bu cevaplar iki farklı resimden geliyorsa doğru cevaba ulaşmanın tek yolu resimde bulunan alt anlamı ortaya çıkartmaktır. Sunulan yeni GSC veri seti daha zor olduğu görülmektedir. Çünkü eşlenik resimler orijinal resimlere oldukça yakın bir şekilde seçilmiştir. Elde edilen veri seti yaklaşık olarak 1,1 milyon resim ve soru ikilisine sahiptir, yaklaşık olarak orijinal GSC veri setinin neredeyse iki katıdır (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017).

Damien Teney ve ekibi 2017 bilgisayarlı görü yarışması için, GSC problemini yeni bir model ya da veri seti üretme kapsamının dışında tutarak, problemin en iyi şekilde hangi parametreler ve mimariler ile çözülebileceğine dair keşifsel bir çalışma gerçekleştirmiştir. GSC, derin öğrenme modellerinin performansı, seçilen mimari ve hiper-parametre araştırmasına oldukça bağlıdır. Yaklaşık olarak 3 000 grafik işlemci ünitesi saati eşdeğerinde mimarilerin ve hiper-parametrelerin derin bir araştırma ile ele alındığı bir çalışma gerçekleştirilmiştir. Çalışmada başarıya giden noktaların ve önem arz eden yerlerin tanımlaması yapılmıştır. Bulgular ele alındığında sigmoid çıktılarının kullanılması, genel yaklaşım olan tek etiketli softmax çıktısına göre çoklu doğru sonuç

elde etmeyi sağlamaktadır. Kelimelerin öncül skorları yani doğruluğu kesin olan hedeflerin yumuşak skorlarının kullanılması, aday cevapların skorlarını bir regresyon görevine çevirebilmektedir. Bu durum geleneksel sınıflandırmadan daha farklıdır. Tüm doğrusal olmayan katmanlarda, kapı yapısına sahip tanh aktivasyon fonksiyonları kullanılmıştır. Genel mimaride tüm doğrusal olmayan katmanlar kapılı tanh olarak inşa edilmiştir. Bu seçim ile kapılı Rektifiye Edilmiş Lineer Birimleri (Rectified Linear Unit, ReLU) yapısından daha çok başarı göstermiştir. Resimler için aşağıdan yukarıya ilgi mekanizması barındıran görsel özellik çıkarma yöntemleri, bölgesel bazda özellik çıkarımı sağlar. Bu yaklaşım, geleneksel kafes benzeri yaklaşımından daha farklı ve daha çok bilgi içeren sonuçlar sunmaktadır. Aday cevaplar için ön eğitimli temsilciler kullanılması önerilmiştir. Verinin eğitimi esnasında geniş yığın eğitim pencerelerinin, veriyi her bir döngüde karıştıran (shuffle) yöntemi ile kullanılması önerilmiştir. Problemin metinsel tarafında ise soru cümleleri kelimelerine ayrılır. Hesaplama kolaylığı için sorular en fazla 14 kelime olacak şekilde ayarlanır. Her bir kelime eşleştirme tablosu kullanılarak 300 boyutlu vektörlere çevrilir. Bu vektörler GloVe (Pennington , Socher , & Manning, 2014) üzerinden elde edilmektedir. Sonuç olarak elde edilen 14 x 300 boyutundaki vektörler kendini tekrarlayan kapılı ünite olan Kapı Özyinelemeli Geçitler (Gated Recurrent Unit, GRU) (Chung, Gulcehre, Cho, & Bengio, 2014) kısmına gönderilir. Kendini tekrarlayan üniteler, 512 ünite boyutuna sahiptir. Görsel tarafta ise girdi resmi direkt olarak konvolüsyonel ağlara gönderilir ve bu sayede K x 2 048 boyutunda vektör temsilleri elde edilir. K burada resim bölgelerini temsil etmektedir. Bu sayede resmin o bölgesindeki bilgiler 2 048 boyutlu vektörler ile temsil edilmektedir. Ayrıca ek olarak aşağıdan yukarıya ilgi ağları ile görsel özellik çıkarımı da yapmışlardır. Görsel özellikler ile metinsel özelliklerin birleştirilmesi, çoklu model füzyon çarpımı ile elde edilir. Çapraz doğrulama yöntemi ile yapılan çalışmalarda 512 boyutu optimal değerleri veren mantıklı bir gizli durum boyutu olmuştur. Sundukları keşifsel analizleri ile GSC çalışmaları için büyük katkı sağlayan en iyi sonuç parametre uzayını sunmuşlardır (Teney , Anderson , He , & Hengel, 2018).

Junwei Liang ve ekibi sıralı resim ve metinler üzerinde gerçek dünya bilgisini dahil ederek GSC yapmayı hedeflemiştir. Çalışmalarında kendilerinin geliştirdiği özgün bir ağ

yapısı olan odaksal görsel metinsel ilgi (Focal Visual-Text Attention, FVTA) kullanmışlardır. Geliştirilen yapı GSC içerisinde kolektif anlamlandırma yapmaya odaklanmıştır. GSC çalışmalarından ayrılarak tekil resim ve soru üzerinden değil toplu resimler üzerinde çalışılmıştır. Örnek vermek gerekirse, bir kullanıcının resim ve videoları sıralı bir şekilde organize edilmiştir. Bu organizasyon yaratılış tarihinden başlayarak devam etmektedir. Bazı resim ve videolar Küresel Konumlandırma Sistemi gibi alt bilgiler ile ilişkilendirilmiştir. Çalışmanın odak noktası bu tarz resim ve metinlerin sorularına cevap verebilecek bir model üretmektir. Çalışmada iki tane zorlayıcı nokta bulunmaktadır. Bunlardan ilki, sunulan girdiler yapılandırılmamış bir formdadır. Bir soru birden çok sekans ile ilişkilendirilmiş ve her bir sekans birden çok zaman adımına sahiptir. Bu tarz formatlara görsel metinsel sekans adı verilmektedir. İkinci zorlayıcı kısım ise bu sıralı veriye sağlanabilecek tekil cevapların yanında yorumlanabilir gerekçeler sunmaktır. Diğer bir ifadeyle cevabı destekleyecek bir kanıt belirtmektir. Bu bahsedilen iki zorlayıcı noktaya dikkat çekebilmek için FTVA modeli doğrulama işlemlerini insanlardan örnek alarak yapmaktadır. Bir soruya cevap verebilmek için insan, ilk olarak girdi bilgileri üzerinden geçer ve küçük detaylara odaklanır ve bu sayede cevabı elde etmeye çalışır. Model bu gidişatı takip eden bir yöntem üzerine kurulmuştur. FVTA ilk olarak alakalı bilgiyi yerelleştirerek öğrenmeye çalışır. Bunu girdi bilgileri üzerinden yapar ardından bir cevabı bulabilmek için çapraz model istatistik yöntemini kullanarak bu alakalı bölgelerin havuzlama verilerini alır. Ardından model, yeni bir çekirdek oluşturarak ilgi temelli veri yapısını ve gizli (latent) bilgiyi ortak bir şekilde işler. Sunulan model, resim ve metin içerisinden elde edilen az, küçük ve alt sekans denebilecek bilgilerini kullanarak toplu akıl yürütme yapılmasına olanak sağlar. Araştırmanın gerçekleştirdiği katkıları özetlemek gerekirse bunu üç aşamada ele almak doğru olacaktır. İlk olarak bu alanda yeni bir ilgi çekirdek yapısı sunuyorlar. Yaptıkları araştırmalarda standart ilgi mekanizmalarından daha iyi sonuç verdiklerini söylenmektedir. İkinci olarak sunulan ilgi sensörleri akıl yürütme sürecini açıklamak için gerekli kalıtsal bilgileri yerelleştirmede kullanabilir. Son olarak, sundukları model ile iki farklı GSC kalite testi çalışmalarında çok iyi performans gösterdiği söylenmektedir (Liang , Jiang , Cao , Li , & Hauptmann, 2018).

Yukarıdan aşağı görsel ilgi mekanizması son dönemlerde resimden metin oluşturma (image captioning) başta olmak üzere birçok görsel temele dayanan çalışmada kullanılmaktadır. Bu yöntem resim üzerindeki aşağıdan yukarıya ilgi mekanizmasını kombine ederek nesne seviyesinde ve diğer göze çarpan resim bölgelerindeki bilgileri ortaya çıkartmak için geliştirilmiştir. Aşağıdan yukarıya ilgi mekanizması Faster-RCNN (Ren , He , Girshick , & Sun, 2015) yapısına dayanmakta ve her bir resim bölgesini özellik vektörü ile ilişkilendirmektedir. Aşağıdan yukarı ilgi mekanizması, uzamsal görüntü özelliklerinin genel bir tanımıdır. Ancak bu çalışmada bu uzamsal bölgeler, bağlı kutular olarak tanımlanmış ve bunun üzerinden aşağıdan yukarıya ilgi mekanizması uygulanmıştır. Faster-RCNN (Ren , He , Girshick , & Sun, 2015) nesneleri iki aşamada tanımlamaktadır. İlk aşamada bölge önerisi ağları ile aday nesneleri tanımlar. İkinci aşamada ise konvolüsyonel ağlar seviyesinde küçük bir ara yapısı ile tanımlanmış olan özellikler üzerinde kayarak ilerler. Çalışmalarında aşağıdan yukarı ilgi modelini eğitirken Visual Genome veri setindeki nesne ve özellik verileri kullanılmıştır. Bazı görsel öğeler MS COCO (Lin , et al., 2014) veri setinde de bulunduğundan dolayı veri kirliliğini önleme amacı ile benzer resimler çıkartılmıştır. Nesneler ve özellikler özgür metin formatında olduğu için metinsel temizlik gerçekleştirmişlerdir. GSC modelini eğitirken daha iyi performans almak için VQA-2 veri setini Visual Genome ile desteklemişlerdir. Sundukları ilgi yapısı, nesne seviyesinde ve nesnenin bağlı olduğu bilgileri kullanarak daha doğal bir çıkarım yapabilmektedir. Sunulan tekli modelde %63,2 performans elde etmişlerdir. Gelişmiş testlerde ise 30 modelin ortak bir şekilde çalıştığı durumda %70,34 performans görmüşlerdir (Anderson , et al., 2018).

Sahneye dayalı GSC çalışmaları, resimlerin içinde bulunan metinlerin yüksek seviye bağlamsal temsillerinin önemini ortaya koymayı amaçlar. Sorulan soruya doğru ve uygun bir cevap üretilmesi için resim içerisindeki metnin okunması ve anlamlandırılmasını temel alan bir dizi artan zorlukta görev tanımlanmıştır. Bu görev için hem nedensellendirme hatalarını ele alan hem de metin tanıma modellerinin eksik yönlerini temel alan yeni bir doğrulama yöntemi sunulmuştur. Bundan dolayı bu metinsel ifadelerden yararlanacak modeller oluşturmak büyük bir ihtiyaçtır. Mevcut GSC çalışmaları ve veri setleri görseldeki metinleri yok sayması gibi ciddi kısıtlara sahiptir.

Resimdeki metinsel ifadelerin çıkartımı konusu karmaşık bir yapıya sahiptir. Bu karmaşık yapı konum, renk, nesne bilgisi ve anlamsal(semantik) bilgileri, lokalize etme, tanımlama ve en son olarak tanımlanmış metnin yorumlanmasını gerektirmektedir. Diğer yandan genel GSC yaklaşımları klasik ve operasyonel şartlanma ile çalışmaktadır. Anlatılanların ışığında bu çalışma yeni bir veri seti Sahne Metin Görsel Soru Cevaplama (Scene Text Visual Question Answering, ST-VQA) (Biten , Tito, Mafla, & Rusiol, 2019) setini ortaya koymuştur. ST-VQA çoklu kaynaklardan beslenerek veri setlerinde görülen, seçim, yakalama, olumsuz etki, taraf gibi etmenleri azaltmıştır. Devamında ise resimde en az iki adet metin içeren görseller seçilmiştir. Bu sayede verilen görsel sorunun en az iki adet olası cevabı olduğundan emin olunmuştur. ST-VQA setine uygun bir şekilde; güçlü bağlamsal, zayıf bağlamsal ve açık sözlük olarak zorlukları artan üç farklı görev tanımlanmıştır. Güçlü bağlamsal görevde, uçtan uca sonuç elde etmek için her bir resme atanmış bir sözlük oluşturulmuştur ve her sözlük resimde sorulan soruların cevap setini içermektedir. Zayıf bağlamsal görevde ise tüm veri seti için benzersiz 30 000 kelimelik bir sözlük oluşturulmuştur. Bunun 22 000 tanesi doğruluğu kesin olan ve 8 000 tanesi de dikkat dağıtıcı özelliğe sahip kelimelerdir. Son olarak açık sözlükte ise hiçbir bilgi sağlanmamıştır. Bu çalışma GSC alanında, görsel nesneler içindeki metinlerin bilgisini kullanarak bir veri seti geliştirilmesi ile GSC çalışmasının farklı bir yönden ele alınmasını sağlamıştır. Geleneksel doğrulama metrikleri yeterli olmadığı ST-VQA için Laveshtein temelli yeni bir doğrulama metriğine dayandırılmıştır (Biten , Tito, Mafla, & Rusiol, 2019).

GSC için nesne düzeyinde temellendirme ile çok noktalı dikkat çalışmasında genelleşmiş bir problem olan, görüntü ve cümleler arasındaki kaba bir ilişkiden sonuç ulaşma yöntemindeki eksiklikler ele almışlardır. Bu sorunu ele almak için çalışmada çok boyutlu bir dikkat yöntemi önerilmiştir. Yöntem, ilk olarak aşağıdan yukarı ve yukarıdan aşağı görüntü özellik çıkarımı yaklaşımı ile başlamaktadır. Daha sonra bu model, detaylı bilgileri keşfetmek için iki tür nesne düzeyinde bilgi çıkarımı ve soruların vektörel temsili için karmaşık bir dil modeli ile bütünleştirilmiştir. Giriş görüntüsünü temsil etmek için nesne algılama tabanlı yaklaşımı benimsemişlerdir. Belirli kategorilere ait sınırlayıcı kutulara sahip, nesnelerin örneklerini tanımlamak için eğitilmiş model uygulanır. Bu

algılama modelinin hedef kategorileri 1 600 nesne ve 400 öznitelik içerir. Her girdi görüntüsü için görüntüyü temsil etmek üzere en yüksek güven puanlarına sahip K nesneleri seçilir. Her nesne için, 2 048 boyutlu bir vektör çıktısı alınmaktadır. Sunulan modelde, soru cümlesi GRU (Chung, Gulcehre, Cho, & Bengio, 2014) ağına beslenir. GRU ağının son gizli durumu cümle özelliği olarak alınır. GSC problemi için, nesne kategori etiketleri ile sorulardaki kelimeler arasındaki anlamsal benzerliğin, atıfta bulunulan nesnelerin yerini belirlemeye yardımcı olduğunu gözlemlenmiştir. Bu nedenle nesnelerin kategori etiketleri ile sorudaki kelimeler arasındaki anlamsal benzerlik açısından, görüntüdeki K nesnelerinin her birine ne kadar ağırlık verilmesi gerektiğini gösteren dikkat vektörünün hesaplanması mümkün kılınmıştır. Deneyler VQA-2 veri seti üzerinde yapılmıştır. Sorular 14 kelime uzunluğuna kadar alınmıştır. Maksimum algılama kutu sayısı bir görüntü için 36 olarak belirlenmiştir. Dil Modellerinden Gömmeler (Embedding from Language Models, ELMO) (Peters , et al., 2018) kelime gömmeleri için boyutu 1 024 olarak hesaplanmıştır. Birleştirilmiş temsil ise 2 048 olarak ele alınmıştır. Geliştirilen model, test setinde yukarı-aşağı modelinde %67,41 ve test standardında %67,73 sonuçlarını elde etmiştir (Huang , Huang , Guo , Qiao , & Zhu, 2019).

GSC içinde tekrarlayan öz ilgi modeli, temel olarak üç modül içerir. Soru temsillerinin çıkarılması, görüntü özelliklerinin çıkarılması ve özellik birleştirme ile sınıflandırma. Sundukları sistem, hem aşağıdan yukarı hem de yukarıdan aşağıya doğru elde edilen görüntü özelliklerini birleştirerek ortak bir temsil oluşturur. Görüntü özelliklerinin çıkarılması temel olarak üç bölüme ayrılmıştır. İlk olarak örüntülerin yoğun ilgisini öğrenmek için gelişmiş hedef algılama modeli Faster-RCNN (Ren , He , Girshick , & Sun, 2015) kullanılır. İkincil olarak öz-dikkat modeli ile elde edilen özellikler geliştirilmiştir. Son olarak soru vektörlerine yönlendirilen çift başlı yumuşak (nokta çarpımı) dikkat yaklaşımı ile çıkarılan özelliklerin bağlamlılığı arttırılmıştır. Tekrarlayan öz ilgi, Google tarafından 2017 yılında önerilen tamamen dikkat odaklı bir model olup, makine çevirisi alanında en iyi sonuçları elde etmiştir. O zamandan beri giderek daha fazla araştırmacı öz dikkat ağına önem vermeye başlamıştır. Metinsel tarafta en yaygın sözcük gömme modeli Word2Vec'tir, ancak Word2Vec (Mikolov , Chen , Corrado , &

Dean, 2013) tüm belgedeki sözcüklerin genel bilgilerini dikkate almadan yalnızca yerel bağlam özelliklerini kullanır. Gizli Anlamsal Analizi, genel istatistiksel özellikleri kullanmasına rağmen metinsel akıl yürütmede düşük performans gösterir. Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) ile karşılaştırıldığında, eş oluşma matrisine dayalı GloVe (Pennington , Socher , & Manning, 2014) hem yerel bağlamı hem de küresel istatistiksel özellikleri dikkate alabilir. Bu nedenle GloVe (Pennington , Socher , & Manning, 2014) modeli kullanılmıştır. Yapılan testlerde soru cümlesi temsillerini öğrenmek için 512 nöron boyutlu GRU (Chung, Gulcehre, Cho, & Bengio, 2014) kullanılmış ve parametreleri GSC modeli ile güncellenmiştir. Görüntü özellikleri Faster-RCNN (Ren , He , Girshick , & Sun, 2015) ile elde edilip 2 048 boyutunda temsil edilmiştir. Her iki farklı öznitelikler entegrasyon için 512 nöron boyutuna indirgenmektedir. Deneysel sonuçlar, çift başlı yumuşak dikkat kullanımının, doğrulama setindeki tekli dikkat modelinden %1,7 daha yüksek öğrenme yeteneğini etkili bir şekilde geliştirebileceğini göstermektedir. Bu çalışma ile GSC görevi için aşağıdan yukarıya, yukarıdan aşağıya dikkat ve öz-dikkat yaklaşımlarını içeren yeni bir model önerilmiştir. Önerilen model VQA-2 veri setinde %67,26'lık genel doğruluğa ulaşmıştır (Zeng , Zhou , & Wang, 2019).

Doğal dil işleme tabanlı GSC çalışması görüntü tanıma ve doğal dil işleme tekniklerinin bir kombinasyonu şeklinde probleme yaklaşım geliştirmiştir. Önerilen yaklaşım, görüntü işleme için EfficientDet (Tan , Pang , & Le, 2020) ve soru işleme için Çift Yönlü LSTM (Bi-LSTM) (Schuster & Paliwal, 1997) kullanır. Model, girdi olarak bir görüntü ve bir soru alır, bunları bağımsız olarak işler, ardından işlenmiş versiyonları birleştirir ve sorunun cevabını tahmin eder. EfficientDet (Tan , Pang , & Le, 2020), daha iyi verimlilikle en son teknoloji doğruluğu sağlayan hafif bir Evrişimsel Sinir Ağları (Covolutional Neural Networks) mimarisidir. Temel mimari, görüntüden daha fazla bilgi elde etmek için bileşik ölçekleme yöntemini kullanarak ölçeklendirme gerçekleştirir. Model tek tip olarak ölçeklendirme gerçekleştirmek için bileşik katsayı kullanan bir yönteme sahiptir. İki ana bileşeni vardır. İlk bileşen EfficientNet (Tan , Pang , & Le, 2020) tabanlı omurga, ikincil bileşen görüntü öznitelik çıkarımı için çift yönlü özellikli piramit ağı Evrişimsel Sinir Ağları mimarisidir. EfficientNet omurgası aslında her bloğun

sonunda aktivasyonu olan bir dizi evrişim (MBConv) bloğudur. Nihai model EfficientDet (Tan , Pang , & Le, 2020) modelin gelen görsel özellikler ile Çift Yönlü LSTM mimarisi ile işlenmiş metinsel çıktıların tek bir katman verecek şekilde birleştirilmesi ile devam eder. Model VQA-2 veri kümesi üzerinde test edilmiştir. Deneysel sonuçlar, modelin çağdaş GSC modelleri ile benzer performans gösterdiğini belirtmiştir (Gupta , Hooda , & Chikkara, 2020).

Liang Peng ve ekibi GSC çalışmalarının hem görüntü hem de doğal dilin sofistike bir şekilde anlaşılmasını gerektirdiğini, aynı zamanda çözümün derin bilgi ve deneyim çeşitliliğine bağlı olduğunu öne sürmüşlerdir. Bu çalışmada özgün olarak iki aşamalı modele genişleyen bir yaklaşım geliştirilmiştir. Kademeli yanıtlama modeli (Cascade Answering Model) yönteminde iki yaklaşım vardır; aday cevap oluşturma (Candidate Answer Generation, CAG) ve son cevap tahmini (Final Answer Prediction, FAP). Bu sayede tüm üretilen aday cevaplar ele alınmış olmaktadır. CAG kısmında ortak ilgi ile üretilmiş olan bütün cevaplardan en alakalı cevaplar seçilir. FAP kısmında ise soru ve resim bilgisini entegre ederek aday cevaplardan anlamsal keşif ile son cevap bulunur. Sundukları metot üç aşamadan oluşmaktadır: Özellik çıkarımı ve CAG & FAP aşamaları. Özellik çıkarımı aşamasında görsel ve metinsel özellikler, hazır modeller ile çıkarılmaktadır. Resim ve soruların güncellenmiş temsilleri elde edildikten sonra bu özellikler birleştirilerek cevap tahminine geçilir. İlk olarak görsel temsiller soru temsilleri ile birleştirilir, burada eleman tabanlı bir çarpım gerçekleşir. Sonuç olarak ortak temsil edilen çoklu modalite elde edilmektedir. Ardından çoklu modalite temsilleri bir softmax katmanına gönderilir ve cevap adayları elde edilmektedir. En yüksek kalitede cevabı elde etmek için ön tanımlı sayıda frekansı en yüksek cevaplar seçilir. Seçilen bu cevaplar soru ve resim ile en alakalı cevaplardır. Bu çalışma VQA-1, VQA-2, VQA-CP, TDUIC, MS COCO üzerinde çalıştırılmış ve açık uçlu cevaplı tarafında ele alınmıştır. Her bir resim 100 adet bölge özelliklerine ayrılmış ve bunların boyutu 2 048 olarak ayarlanmıştır. Metin kelime vektörleri boyutu 620 olarak ve bir nöral katman boyutu 1 024 olarak ayarlanmıştır. Soru uzunluğu 15 olarak belirlenmiş ve CAG tarafından en yüksek frekanslı 5 cevap seçilmiştir. Sınıflandırma tarafında cevap sayısı 3 000 olarak

sınırlandırılmıştır. Elde ettikleri sonuçlar güncel GSC çalışmaları ile oldukça yakın sonuçlar göstermiştir (Liang , et al., 2020).

Belge üzerinde GSC çalışması olan DOC-VQA (Mathew , Karatzas , Manmatha , & Jawahar, 2021), belge görüntülerindeki bilgileri koşullu olarak yorumlamak için DOC-VQA (Mathew , Karatzas , Manmatha , & Jawahar, 2021) çalışmasını tanıtmakta ve yeni bir veri kümesi sunmaktadır. Genel GSC çalışmalarından yola çıkarak, belge görüntülerinin doğası yukarıdaki tüm görsel ipuçlarından yararlanmak, kullanılan örtük yazılı iletişim kuralları hakkında önceden bilgi sahibi olmaktır. Bu tür görüntülerde iletilen yüksek yoğunluklu anlamsal bilgilerle başa çıkmak için farklı bir yaklaşım gerektirir. Belgeler alanına görsel soru yanıtını getirme konusundaki önceki yaklaşımlar, veri görselleştirmeleri gibi belirli belge öğelerine veya kitap kapakları gibi belirli koleksiyonlara odaklanılmıştır. Bu tür yaklaşımların aksine, sorunu genel biçimine göre yeniden şekillendirmişlerdir ve geniş ölçekli çeşitli belgeler koleksiyonu (veri seti) ortaya koymuşlardır. Bu çalışmanın temel katkıları şu şekilde özetlenebilir. İlk olarak 50 000 soru ve cevap tanımı barındıran çeşitli tür ve içerikteki 12 767 belge görüntüsünden oluşan büyük ölçekli bir veri seti DOC-VQA (Mathew , Karatzas , Manmatha , & Jawahar, 2021) tanıtılmıştır. Veri kümesi; tablolar, formlar ve şekiller gibi öğelerin yanı sıra bir dizi farklı metinsel, grafiksel ve yapısal öğeyi içeren çok sayıda farklı belge türünü kapsar. Tanımlanan sorular, muhakeme gereksinimlerine göre kategorize edilerek DOC-VQA yöntemlerinin ayrıntılı analizinin yapılmasına olanak vermiştir. İkincil olarak DOC-VQA veri seti, farklı varsayımlar ile üst performans sınırlarını zorlayan ve son teknoloji sahne tabanlı GSC modellerine ve doğal dil işleme modellerine kadar çeşitli temel yöntemlerin bir arada kullanılmasını teşvik eden bir zorluk seviyesine sahiptir (Mathew , Karatzas , Manmatha , & Jawahar, 2021).

Öncül çalışmalar ele alındığında ortaya çıkan sonuç, GSC çalışmalarında ağırlıklı olarak görsel ilgi yöntemlerine önem verdiğidir. Çalışmaların çoğu GloVe ön eğitimli vektörlerine dayandırılmaktadır. Yapılan tek merkezli kelime vektörleri yaklaşımı gerçekleştirilen tez için bir temel olmuştur. Literatürde birçok alanında kendi katkılarını yapmış kelime vektörleri vardır. Ayrıca, son yıllarda oldukça önem kazanmış olan çok başlı ilgi yöntemi ile geliştirilen bağlamsal kelime vektörleri çalışmaları gittikçe

popülerlik kazanmaya başlamıştır. Sunulan bu tez, GSC problemi için ön eğitimli ve bağlamsal kelime vektörlerinin ele alındığı, detaylı bir keşifsel analiz sunmaktadır. Bu sayede, sadece görsel ilgi yöntemlerinin değil, kelime ön eğitimli ve bağlamsal kelime vektörlerinin etkisi, GSC problemi için kapsamlı şekilde sunulmuştur.

3. MATERYAL ve YÖNTEM

Doğal dil işleme, bilgisayarlı görme, makine çevirisi, resimden yazı türetme gibi disiplinler arası problemlerin ortak yaşadığı sorun, eğitim sağlanacak veri seti eksikliğidir. GSC alanı oldukça karmaşık olduğundan, iyi bir veri kümesine sahip olunmalı hatta gerçek dünya senaryoların da bile sorular ve görüntü içeriği içindeki uzun olasılıkları yakalayacak kadar büyük olmalıdır.

Çalışmada en güncel GSC veri seti olan VQA-2 seti üzerinde çalışılmıştır. Çalışmada ön eğitimli kelime vektörleri olan Word2Vec, GloVe ve FastText, bağlamsal kelime vektörü olan BERT ile birlikte karşılaştırmalı olarak ele alınmıştır. Görsellerin vektör temsilleri için, nesne tanıma modellerinden alanında en iyi sonuçları veren Xception ve Inception-Resnet-V2 modelleri kullanılmıştır.

3.1 Öğrenme Tabanlı Yöntemler İçin Kelime Vektörleri Temsilleri

Öğrenmeye dayalı doğal dil işleme görevlerinin gerçekleştirilebilmesi için metinsel özellikler, sayısal özellikler olarak ifade edilmelidir. Matematiksel ifadelere kelime vektörleri/ gömmeleri denilmektedir. Yıllar içerisinde doğal dil işleme alanında Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013), GloVe (Pennington , Socher , & Manning, 2014) ve FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) gibi birçok kelime vektörü yöntemi geliştirilmiştir. Geleneksel ön eğitimli kelime vektörlerinin yanında bu çalışmada bağlamsal kelime vektörü olan BERT (Devlin , Chang , Lee , & Toutanova, 2019) yöntemide ele almıştır.

2013 yılında Thomas Mikolov skip-gram temelini benimseyen Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) kelime gömme yöntemi geliştirildi. Ardından, Word2Vec'in kısıtlamalarının aşılmasını amaç edinen GloVe (Pennington , Socher , & Manning, 2014) vektörleri ortaya çıkmıştır. Analojik doğrulama yöntemine göre, tanıtılan GloVe (Pennington , Socher , & Manning, 2014) vektörleri, Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013)'den daha iyi sonuç verdiği için yeni bir standart haline gelmiştir. 2017 yılında Mikolov ve ekibi tarafından, ön eğitimli vektörlerin performans kısıtlarını gideren bir yaklaşım olan FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) vektörleri tanıtılmıştır. Uzun bir süre ön eğitimli kelime vektörleri doğal dil işleme

alanında oldukça tercih edilmiştir. 2018 Yılında, çift yönlü kelime işleme ve gerçek bağlamdan anlamı koruyarak vektör çıkarma yöntemi olan BERT (Devlin , Chang , Lee , & Toutanova, 2019), Google beyin takımın tarafından geliştirilmiştir. BERT, kelime vektörlerini bağlı bulunduğu içerikten öğrenerek alt anlamı bütünüyle yansıtan bir yöntemdir. Tanıtılan yeni yöntem ile doğal dil işleme ve kelime vektörleri araştırma alanı yeni bir çağa girmiştir.

Kelime vektörlerinin yanında, paragraf ya da dokümanları sabit uzunlukta temsil eden doküman vektörleri de tanıtılmıştır. Cümlelerin ve Belgelerin Dağıtılmış Temsilleri (Distributed Representations of Sentences and Documents, Doc2Vec) (Le & Mikolov, 2014) genel olarak iyi performans vermesine karşın, kelimeler arası bağlamın, problemin çözümüne etkisinin yüksek olduğu yaklaşımlarda düşük performans göstermektedir.

3.1.1. Word2Vec

Kelimelerin dağıtık şekilde temsil edilmesinin, öğrenme temelli algoritmalarda büyük ölçüde performans artışı yapacağı varsayılmıştır. Mikolov ve ekibi bu yaklaşımdan yola çıkarak skip-gram modeline dayanan kelime vektörleri olan Word2Vec'i tanıtmıştır. Skip-gram yönteminin en büyük avantajı, yoğun matris çarpımları içermemesidir. Bu nedenle hesaplama karmaşıklığı kapsamında büyük bir kazanç sağlamıştır. Belirttiklerine göre, optimize edilmiş bir cihaz üzerinde tek bir makine bir günde yüz milyar kelime öbeğinden vektör temsili öğrenebilmektedir. Araştırmalarında ayrıca skip-gram yaklaşımını genişleterek frekansı yüksek kelimelerin alt örneklemini de uyarlamışlardır. Bu sayede eğitim esnasından 2 ile 10 kat arasında eğitim hızı kazancı sağlamışlardır. Alt örneklem ayrıca frekansı düşük olan kelimelerin bile doğru temsilini öğrenebilmektedir. Skip-gram modelinin temel görevi, seçilen kelimenin, etrafında bulunan kelimeler ile arasındaki ilişkiden vektör temsillerini öğrenmesidir. Formüle edilirse, verilen kelimeler $w_1, w_2, w_3 \dots \dots, w_t$ için amaç logaritmik olasılığı maksimize etmektedir.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} log_p(w_{t+j}|W_t) \tag{3.1}$$

Denklem 3.1'de gösterildiği üzere c eğitim içerik boyutunu temsil eder ve içerik arttıkça ortaya çıkacak olan modelin kalitesinin artacağı hipotezi kurulabilmektedir.

Her öğrenme yaklaşımının performansı test edilmelidir. Geleneksel test ve etiket tabanlı doğruluk ölçüm metriği kelime gömmelerinin kalitesini ölçmek için uygun değildir. Sunulan yeni yaklaşım için yeni bir doğruluk metriği gereksinimi ortaya çıkmıştır. Eğitilmiş modelin test edilmesi için analojik nedensellendirme yöntemi geliştirilmiştir. Örnek olarak, "*Montreal", "Montreal Canadiens", "Toronto", "Toronto Maple Leaf*" kelimeleri ele alınmış olsun. En yakın vektör temsillerini alınarak toplama işlemi uygulandığı varsayılır. *Vec("Montreal Canadiens") − Vec("Montreal") + Vec("Toronto")* işleminin sonucu *Vec("Toronto Maple Leaf")* kelimesini sonuç olarak elde edilmelidir. Ortaya çıkan cevap beklentiyi karşılıyorsa, eğitilen modelin doğru çalıştığı varsayılır.

Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) modelini eğitmek için haber metinlerinden oluşan geniş bir metin külliyatı kullanılmıştır. Kelimeler içinde frekansı beşten az olanlar önemsiz kabul edilip eğitim esnasında elenmiştir. Sonuç olarak 629 000 kelimelik bir sözlük boyutu elde edilmiştir. Ardından beraber görülme sıklığı yüksek olan kelimeler tanımlanmıştır. Bu esnada hesaplama karmaşıklığını azaltmak için çoğul yerine, ikili ya da üçlü kelime grupları ele alınmıştır. İlerleyen aşamada eğitilen model birkaç kez daha üzerinden geçilerek son çıktıya ulaşılmıştır. Doğrulama yöntemi daha önce bahsedilmiş olan analojik nedensellik ile gerçekleştirilmiştir. Eğitim süreci eğitim setinin birkaç katı daha büyük bir set ile gerçekleştirilerek keşifsel analiz yapılmıştır. Model 33 milyar kelime ile %72 doğruluğa kadar çıkabilmektedir. Skip-gram yaklaşımına dayanan Word2Vec yönteminin, kelimeleri vektör uzayında yüksek doğruluk ile temsil ettiği sonucuna ulaşılmıştır (Mikolov , Chen , Corrado , & Dean, 2013).

3.1.2. GloVe

Mikolov'un araştırmasında (Mikolov , Chen , Corrado , & Dean, 2013) bahsedildiği gibi, kelimelerin gerçek vektörler ile gösterilmesi anlamsal(semantik) vektör uzayı modellerinin temel amacıdır. Genel olarak kelime temsilleri, kelimeler arasındaki uzaklık

ya da birbiri arasındaki açıya bağımlıdır. Bu yaklaşım doğal dil işleme alanında oldukça kabul görmüştür.

Vektör uzay modelleri arasında iki ana metodoloji vardır. İlk yöntem, Gizli Anlamsal(Semantik) Analiz mantığına dayanan genel matris faktorizasyonudur. İkinci yöntem ise Mikolov'un temel aldığı skip-gram modelidir. Analojik doğrulama yöntemleri ile Gizli Anlamsal(Semantik) Analiz metotlarının istatistiksel veriler üzerinde yüksek performans verdiği ancak kelime analojik görevlerinde geri kaldığı ortaya çıkmıştır. Aynı şekilde skip-gram yöntemini temel alan modeller analojik nedensellik üzerinde yüksek performans verirken, istatistiksel yaklaşımlarda geride kalmaktadır. GloVe (Pennington , Socher , & Manning, 2014) yaklaşımı bu ilişkiden ortaya çıkarak, ağırlıklı en az kareler yöntemini önermektedir. Sunulan model, kelimelerin birlikte görülme frekanslarını hesaplarken etkili bir şekilde istatistik yöntemini de baz alabilmektedir.

Önerdikleri yönteme göre, denetimsiz kelime vektörü öğrenme yaklaşımının, başlangıç noktası kelimelerin olasılıkları yerine, beraber görülmelerinin oranları olmasıdır. Bu hipotezi, çalışmalarında kelimenin ham olasılığı ile görülme oranlarını karşılaştırarak kanıtlamışlardır. GloVe'a göre, ağırlıklı en az kareler yönteminin kullanılması, birlikte görülmenin eşit şekilde ağırlıklandırılması problemini ortadan kaldırmaktadır. Eşit şekilde ele alınan dağılımlar gürültülü bir veri oluşturmaktadır. Sıfır olarak atanmış girdiler birleştirildiğinde %75'e kadar veri düzensizliği yaratmaktadır. Sundukları ağırlıklı yaklaşım Denklem 3.2 de gösterilmiştir.

$$J = \sum_{i,j=1}^{v} f(X_{i,j}) \, (w_i^T \tilde{w}_j + b_i + \tilde{b}_j - log X_{ij})^2 \tag{3.2}$$

Denklem 3.2'de v sözlük boyutunu temsil etmektedir. Yeni ağırlıklı denklemin optimal sonuçları vermesi için aşağıdaki kriterleri sağlaması beklenmektedir.

- $F(0) = 0$ Eğer f bir devamlı fonksiyon olarak kabul edilmişse, $x \to 0$ durumunu nötralize etmesi beklenir.

- Nadir görülen kelimelerin gereğinden fazla ağırlıklandırılması sorununu önlemek için $f(x)$ azalmayan bir yapıda olmalıdır.
- Frekansı yüksek kelimelerin aşırı ağırlıklandırılmasını önlemek için $f(x)$ küçük olmalıdır.

GloVe (Pennington , Socher , & Manning, 2014) yöntemi $1M - 1{,}6M - 4{,}3M - 6M$ büyüklüğündeki kelime külliyatları ile eğitilmiştir ve üst limit olarak $42M$ kelime külliyatına kadar çıkılmıştır. Çalışmalarında, Mikolov'un (2013) öne sürdüğü model kalitesi ve külliyat büyüklüğü arasındaki ilişkinin her zaman doğru sonuç vermediğini iddia etmişlerdir. GloVe (Pennington , Socher , & Manning, 2014) yöntemi, Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) gibi analojik nedenselleme yöntemi ile test edilmiştir ve en iyi sonucu 300 boyutlu kelime vektörlerinde verdiği gözlenerek, Word2Vec'den daha yüksek bir temsil skoru elde etmiştir (Pennington , Socher , & Manning, 2014).

3.1.3. FastText

Her yeni araştırma, kendisinden önce sunulan yöntemlerin eksikliklerine odaklanmaktadır. Yıllar içerisinde ön eğitimli kelime vektörleri, devamlı kelime torbası yöntemi (bag-of-words) ya da skip-gram yöntemlerine dayanmaktaydı. Mikolov ve ekibi (2013) sundukları bu çalışmada, dikkatli bir şekilde ele alınmayan basit ön işlemlerin uygulanması sonucunda, kelime vektörlerinin kalitesinin arttırılabileceğini öne sürmüştür. Bu çalışma temel olarak kelime alt örneklemleri ve sözcük grubu temsillerine odaklanmaktadır.

Standart kelime torbası yöntemi, içinde bulunduğu bağlamdaki kelimeleri tahmin etmesi ile vektör temsili öğrenmektedir. Burada bağlam, kelime üzerinde kayan pencere olarak ifade edilmiştir. Formülize şekilde amaç $w_1, w_2, w_3, \dots\dots, w_t$ kelimelerinden oluşan T cümlesinde logaritmik benzerliği maksimize etmektir.

$$\sum_{t=1}^{T} \sum_{c \in C_t} logp(w_t | C_t) \tag{3.3}$$

Denklem 3.3, maksimize edilmesi hedeflenen logaritmik benzerliği temsil etmektedir. Burada w_t, sekans içindeki kelimeleri C_t, ise içinde bulunduğu bağlamı temsil eder. Yapılan çalışmada kelimeler arasındaki skorlama fonksiyonu $s(w, C)$ olarak gösterilmiştir. Bu bilgiler ışığında, koşullu olasılık, bir bağlam ve kelimeye uygulanan softmax fonksiyonu olarak ele alınır. İddia edildiği üzere bu yaklaşım zengin kelime külliyatları için uygun değildir. Var olan soruna dikkat çekmek için önerilen çözüm, kelimelerin ikili sınıflandırılmasının yerine, doğru kelimenin, örneklenmiş negatif adaylarla birlikte ele alınmasıdır.

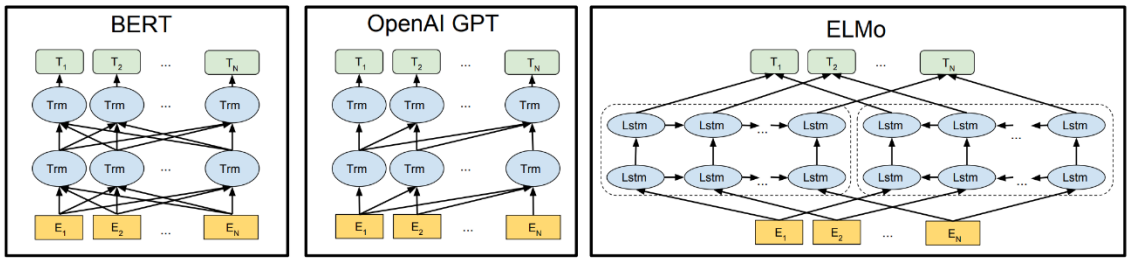
$$\sum_{t=1}^{T}\left[\log\left(1+e^{-s(w_t,C_t)}\right)+\sum_{n\in N_{C_t}}\log\left(1+e^{s(n,C_t)}\right)\right] \tag{3.4}$$

Denklem 3.4 negatif logaritmik benzerliğin, ikili lojistik kayıp fonksiyonu ile bağlam pozisyonu c için elde edilebileceğini gösterir. Sunulan çözüm basitçe kelime olasılığını ve negatif örneklemi temsil eden N_c'yi değiştirerek maksimize edilmiş hedef fonksiyonunu elde eder.

Bu alanda yapılan bir diğer geliştirme ise kelime alt örneklemidir. Belirteçlerin tüm oluşumlarını eşit olarak ele almak, en sık kullanılan kelimeler için fazla uyum sorununa ve daha az sıklıkta olan kelimeler için yetersiz uyum sorununa neden olur. ZipF dağılımı, kelimelerin çoğunun, külliyatın küçük alt kümesine ait olduğunu belirtir. N-gram yaklaşımının kullanılmasının, temsil problemini çözebileceği öne sürülmüştür. En büyük dezavantajı, n-gram kullanmanın eğitim karmaşıklığını artırabilmesidir. Bu çalışma, n-gramları yinelemeli olarak seçmeyi önerir, daha sonra bu n-gramlar ön işlem aşamasında tek bir kelimede birleştirilir. Mevcut kelime vektör modellerinin bir kelimenin iç yapısını göz ardı ettiği belirtilmektedir. Çoğu durumda iç yapı, yanlış yazılmış veya nadir bir kelime biçiminde yüksek değerli bilgiler taşır. Bu durum genellikle Türkçe veya Fince gibi morfolojik olarak zengin dillerde olur. Önerilen çözüm, karakter n-gramlarını kullanarak kelime ağırlıklarını zenginleştirmektir. Bu, her kelimeyi karakter n-gramlarına

ayrıştırarak ve her n-gramı vektörle temsil ederek elde edilir. Vektör, basitçe n-gramların toplamıdır. Bu, sözcük dağarcığı dışındaki kelimelerin, n-gram karakterlerine dayalı olarak, ortak oluşum matrisi kullanılarak temsil edilebileceğini varsayar. Bu araştırma; wiki dökümleri, haber verileri ve genel tarama gibi çeşitli eğitim verilerini kullanır. FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) üzerinde eğitilen en büyük derlem, 630 milyar kelime içeren yaygın tarama derlemidir. FastText, kelime analojisi, nadir kelimeler ve Stanford Soru Cevaplama Veri Seti (Stanford Question Answering Dataset) üzerinde değerlendirilmiştir. Sonuçlar, GloVe (Pennington , Socher , & Manning, 2014) kelime vektörlerinden daha iyi performans göstermektedir.

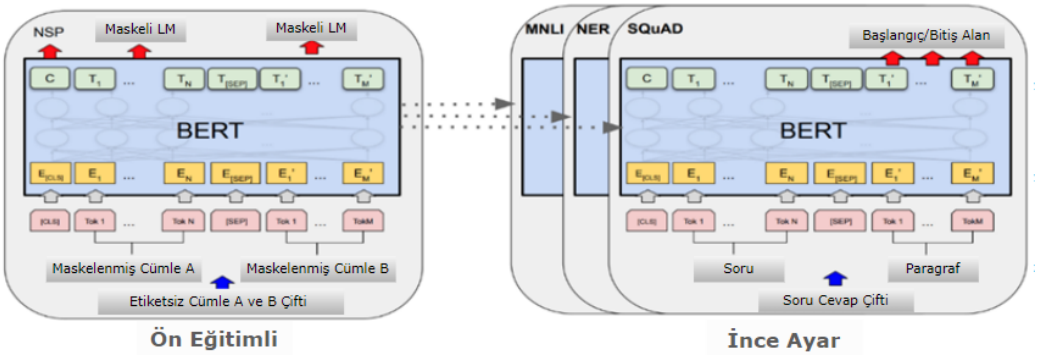
3.1.4 BERT



Şekil 3.1. Bert ve diğer transformatör temelli mimarilerin karşılaştırması (Devlin , Chang , Lee , & Toutanova, 2019)

Doğal dil işleme araştırma alanındaki en yeni atılım BERT'dir. BERT (Devlin , Chang , Lee , & Toutanova, 2019), 11 farklı dil işleme görevini tek başına çözebilmektedir. Bu bölüm, BERT altyapısına ve dil görevleri üzerindeki yeteneklerine odaklanmaktadır. BERT, etiketlenmemiş metin külliyatı kullanılarak önceden eğitilmiş derin çift yönlü temsil üzerine kuruludur. Ayrıca, tüm katmanlarda soldan sağa ve sağdan sola gösterimlerden öğrenebilmek için ortak koşul kullanır. Önceden eğitilmiş dil temsilleri için öznitelik tabanlı ve ince ayar (fine-tuning) olmak üzere iki ana yaklaşım vardır. Özellik tabanlı yaklaşıma harika bir örnek olan başka bir transformatör kodlayıcı modeli ELMO'dur. ELMO (Peters , et al., 2018), ek özellik olarak önceden eğitilmiş temsilleri içeren, göreve özel yaklaşım kullanan, başka bir temsilci ağıdır. Bir diğer bağlamsal temsilci olan Üretken Önceden Eğitilmiş Transformatör (Generative Pre-trained Transformer, GPT) (Brown , et al., 2020), göreve özel minimum parametre

kullanmaktadır. Bu iki görev farklı görünebilir, ancak her iki yöntem de tek yönlü eğitim yaklaşımından muzdariptir.

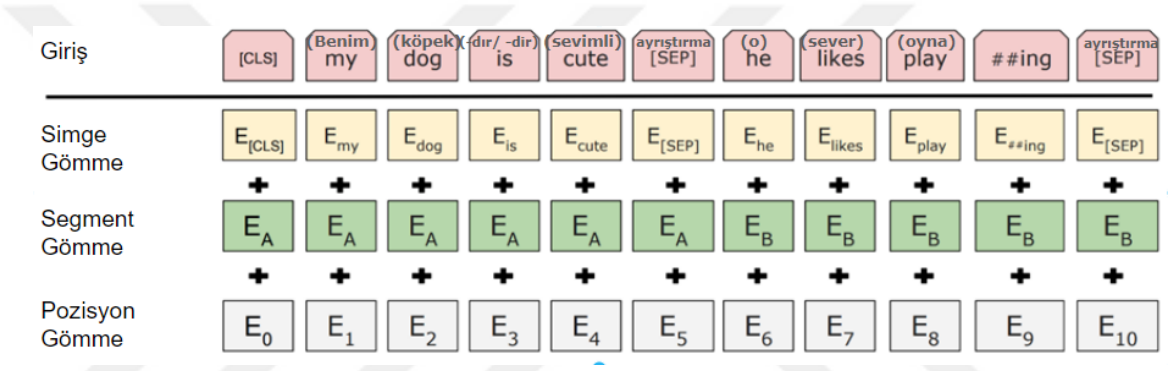


Şekil 3.2. Bert ön eğitim ve ince ayar mimarileri temsili (Devlin , Chang , Lee , & Toutanova, 2019)

Yeni tanıtılan BERT (Devlin , Chang , Lee , & Toutanova, 2019), maskelenmiş dil modeli ile bu sınırlamanın üstesinden gelebilmiştir. Maskelenmiş dil modeli kullanmak, BERT'in eğitim esnasında soldan sağa ve sağdan sola bilgileri işlemesine ve çift yönlülüğün gerçek gücünü ortaya çıkarmasına olanak sağlar. Transformatör ağlarında ön eğitim ve ince ayar olmak üzere iki yaklaşımın olduğundan bahsedilmektedir. İnce Ayar, etiketli eğitim setini kullanarak eldeki problemin büyük birçoğunu genelleştirebilir. İnce ayar yaklaşımı için BERT, önceden eğitilmiş olan tüm parametreler ile başlatılır ve böylece az miktarda çalışma ile istenen çözümlerin elde edilmesini sağlar. BERT'in model mimarisi çok katmanlı perspektife dayanmaktadır. Tüm katmanlar, maskelenmiş dil modeli kullanılarak çift yönlülükten yararlanacak şekilde ayarlanır. Bir terminoloji olarak BERT, katman boyutunu belirtmek için L'yi (Layer), gizli durum için H'yi (Hidden) ve dikkat kafaları için A'yı (Attention) kullanır. Model, $BERT_{Base}$ ve $BERT_{Large}$ olmak üzere iki ana modele sahiptir. Temel model 12 katmanlı, 768 gizli durum ve 12 dikkat başlığından oluşur. Büyük model 24 katman, 1 024 gizli durum ve 16 dikkat başlığı ile gelir. $BERT_{Base}$ ve GPT (Brown , et al., 2020) arasındaki en büyük fark, GPT kısıtlı öz ilgi kullanırken BERT'in ortak koşullu dikkat kullanmasıdır. Geleneksel yaklaşımlardan farklı olarak BERT (Devlin , Chang , Lee , & Toutanova, 2019), çeşitli aşağı akış görevlerini ele almak için farklı girdi ve çıktı temsilleri uygulayabilmektedir. BERT aynı zamanda WordPiece bölücüsü (tokenizer) adlı yeni bölücü sürecini de tanıtmaktadır. Karakter boşluklarını kullanarak basitçe bölme gerçekleştiren veya büyük kelime

dağarcığı gerektiren öncül çalışmaların aksine BERT, verilen cümleleri bölmek için kelime kök ve olası eklerden oluşan 30 522 boyutunda bir sözlük kullanmaktadır. Kelime parçası bölücüsü çeşitli seçeneklerle birlikte gelir. Ancak en iyi yaklaşım, kelimeleri ön ekine ve son ekine ayrıştıran ve bu bilgiyi kullanarak simgeleştiren Tam WordPiece belirtecidir. Bu yaklaşım, az miktarda bilgi kullanarak daha iyi sonuçlar elde eder.

Her dizi girişi (tek veya cümle çiftleri) '[CLS]' belirteci ile işaretlenir. Bu simge, dizinin başlangıcını temsil eder ve bir cümle temsili olarak kullanılabilir. Verilen girdi cümle çiftlerinde ise, her çift '[SEP]' belirteci ile ayrılır. Daha sonra her cümle dizisi şekilde gösterildiği gibi belirteç kimlikleri, segment kimlikleri ve konum yerleştirme ile temsil edilir.



Şekil 3.3. BERT mimarisi kelime girdi gösterimi (Devlin , Chang , Lee , & Toutanova, 2019)

Maskelenmiş dil modeli, giriş belirteçlerinden bazılarını rastgele maskeler ve bu belirteçleri tahmin etmeye çalışır. BERT (Devlin , Chang , Lee , & Toutanova, 2019) mimarisi, maskeli dil modeli için sözcük parçası belirteçlerinin %15'ini rastgele maskelemektedir. Daha sonra maskelenmiş belirteçleri temsil eden son gizli vektörler, sözlük üzerinden softmax çıktısına beslenir. Bu süreç, dil modellerinin çoğu için benzerdir.

Dizilerdeki gürültü sorununu önlemek için BERT (Devlin , Chang , Lee , & Toutanova, 2019), tüm diziyi yeniden oluşturmak yerine yalnızca maskelenen sözcüğü tahmin eder. Ön eğitim aşaması 800 milyon kelimeye sahip kitap külliyatı ve 2,5 milyar kelimeye sahip İngilizce Wikipedia kullanılarak gerçekleştirilmiştir. Wikipedia külliyatından yalnızca metin pasajları çıkarılır ve bilgilerin geri kalanı atılmaktadır. İnce ayar aşaması basit ve

yalındır, bu nedenle araştırmacıların BERT'i hem cümle çiftleriyle hem de tekli cümlelerle birçok alt görev için kullanmasını sağlar (Devlin , Chang , Lee , & Toutanova, 2019).

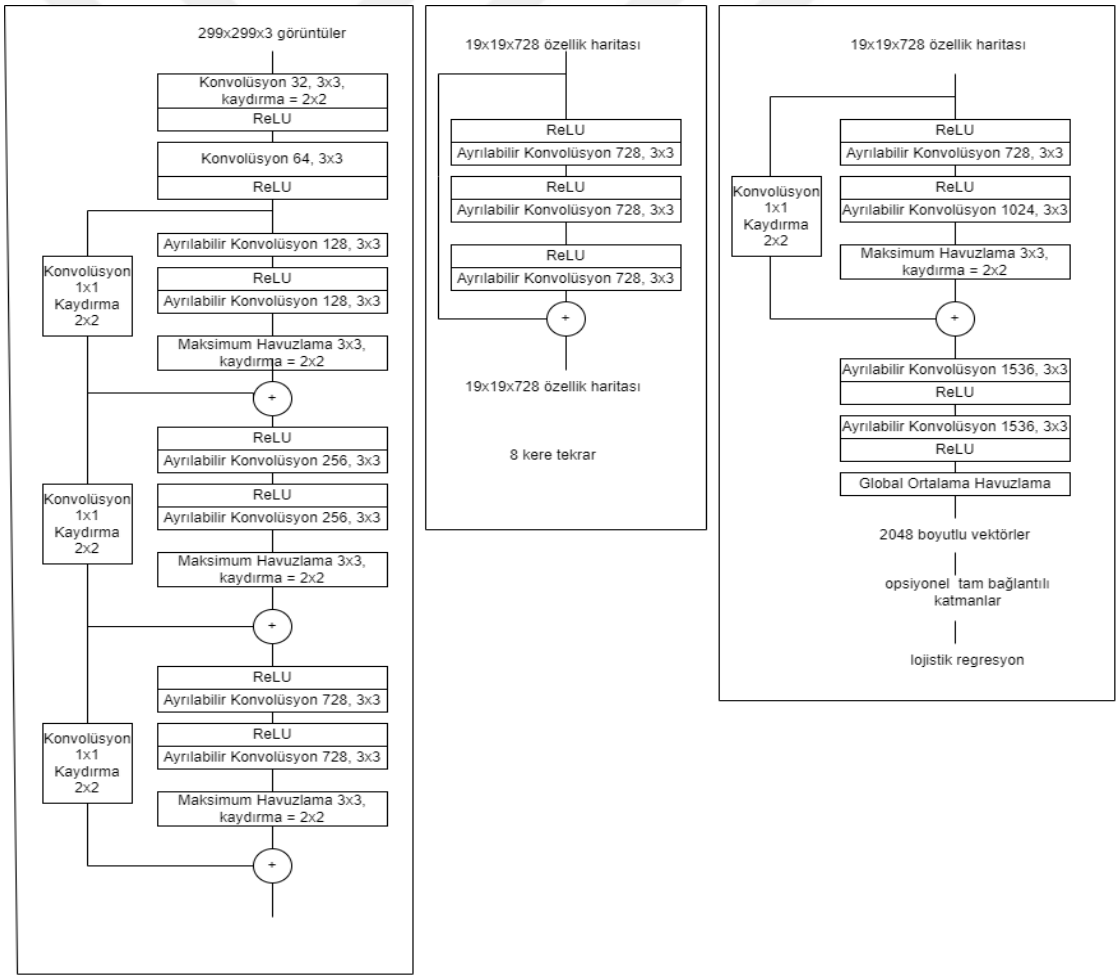
3.2 Görsel Vektörlerin Elde Edinimi İçin Konvolüsyonel Ağlar

GSC çalışması, bir resim ve resme sorulan soru üzerinden cevap edinmeyi amaçlar. Bu amaç, doğal dil işleme ve bilgisayarda görü görevlerinin kesiştiği yerde çözüme ulaşmaktadır. Görsellerden bilgi çıkarımı yapabilmek ve geniş analizler gerçekleştirebilmek için, kelime vektörleri gibi matematiksel düzlemde temsil edilmeleri gerekmektedir. Bu bağlamda çeşitli konvolüsyonel ağlar geliştirilmiştir. Konvolüsyonel ağlar oluşturduğu bir kernel üzerinden resmin üzerinden kayan pencereler gibi ilerleyerek öncül vektör çıkarımı gerçekleştirir. Ardından, elde edilen görüntü vektörleri derin konvolüsyonel ağlar içerisinden bilgi çıkarımı işlemlerine tabii tutularak ağırlıklı son durum vektörleri elde edilmektedir. İlk olarak LeNet5 (LeCun , Bottou , Bengio , & Haffner, 1998) ile literatüre giren konvolüsyonel yöntemler, işlem gücünün artması ile AlexNet'in (Krizhevsky , Sutskever , & Hinton, 2012) ortaya çıkmasına ve konvolüsyonel ağların popülerlik kazanmasına olanak tanımıştır. Artan işlem gücü ile daha derin ağlar ve ilgi yöntemleri geliştirilmiştir. Günümüzde en yüksek doğruluk veren ağlardan ikisi Xception (Chollet, 2017) ve Inception-Resnet-V2'dir (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017).

3.2.1 Xception

Xception (Chollet, 2017), Inception ağlarının aşırı (extreme) versiyonu olarak ortaya çıkmıştır. Yapısındaki değiştirilebilir derinlemesine ayrılabilen evrişim ağları ile Inception-v3'ten daha iyi performans göstermektedir. Orijinal derinlemesine ayrışabilir konvolüsyonel ağ mantığı, derinlemesine ayrışabilir ağların noktasal bir ağ ile birleştiği durumdur. Basitçe derinlemesine konvolüsyonel ağ, kanal bazında $N\ x\ N$ uzaysal evrişimdir. Eğer beş adet kanal bilgimiz var ise, 5 farklı $N\ x\ N$ uzaysal konvolüsyon elde ederiz. Noktasal konvolüsyon ise bu mimarinin $1\ x\ 1$ konvolüsyon ile tamamlandığı durumdur. Geleneksel konvolüsyon ile karşılaştırıldığında, tüm kanallarda konvolüsyon gerçekleştirmemiz gerekmez. Bu, bağlantı sayısının daha az olduğu ve modelin daha hafif

olduğu anlamına gelir. Değiştirilmiş derinlikte ayrılabilir konvolüsyon, noktasal konvolüsyon ve ardından derinlik yönünde konvolüsyon ile tamamlanan ağlardır. Bu değişiklik, Inception-v3'teki başlangıç modülü tarafından motive edilir. Mimariler arasında iki temel farklılık vardır. İlk farklılık, noktasal konvolüsyonların uyarlama farklılığıdır. İkincil olarak orijinal başlangıç modülünde, ilk işlemden sonra doğrusal olmayan (non-linear) bir yaklaşım vardır. Değiştirilmiş derinlemesine ayrılabilir evrişim olan Xception'da, hiçbir ara katmanda ReLU doğrusalsızlığı yoktur. Farklı aktivasyon birimleri ile değiştirilmiş derinlikte ayrılabilir konvolüsyon test edilmiştir. Elde edilen sonuçlarda herhangi bir ara aktivasyon olmadan Xception, ReLU kullananlara kıyasla en yüksek doğruluğa sahiptir.

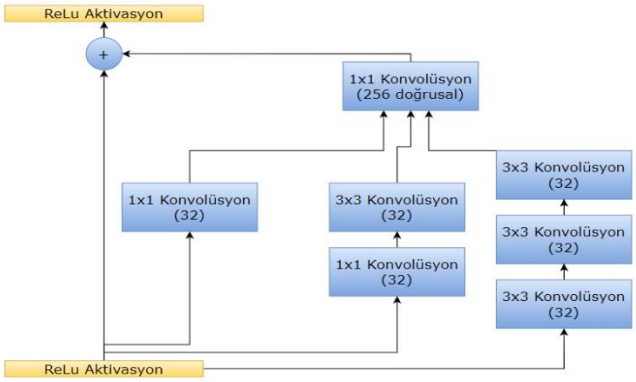


Şekil 3.4. Detaylı Xception mimarisi (Chollet, 2017)

Şekil 3.4'de gösterilen detaylı Xception (Chollet, 2017) mimarisinden anlaşıldığı gibi ayrıştırılabilir konvolüsyon modülleri birer Inception (Szegedy , et al., 2015) modülü gibi ele alınmıştır. Mimaride görüldüğü gibi, artık bağlantılar vardır ve bu bağlantılar modelin performansını derinden etkileyen etmendir. Xception mimarisinin, ImageNet (Russakovsky , et al., 2015) veri seti ile gerçekleştirilen eğitim çıktısı diğer derin konvolüsyonel ağlar ile karşılaştırıldığında birçoğunu geride bırakarak en yüksek 5 doğruluk metriğinde %94,5'lik bir başarı sergilemiştir (Chollet, 2017).

3.2.2 Inception-ResNet v2

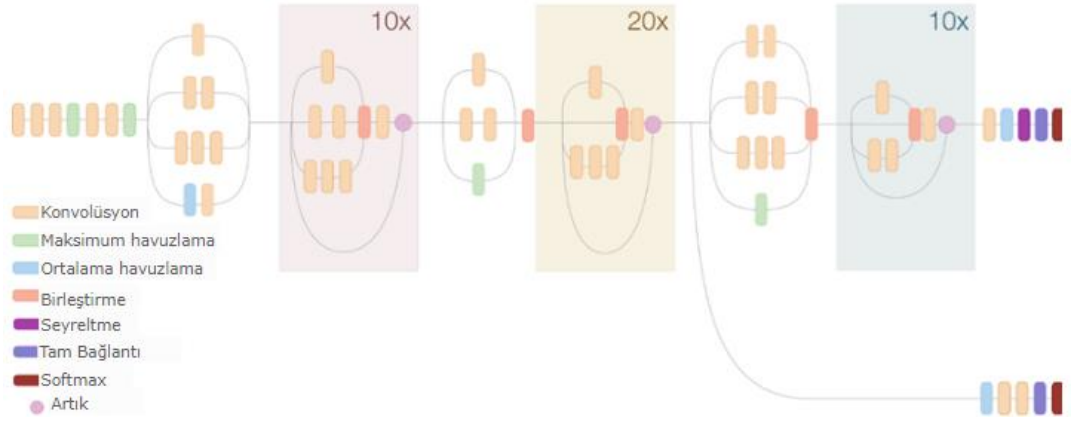
ResNet ve Inception, düşük bir hesaplama karmaşıklığı ve çok iyi bir performans sunarak, son yıllarda görüntü tanıma problemlerindeki en büyük ilerlemelerden biri olmuştur. Inception-ResNet, Inception mimarisini artık bağlantılarla bütünleştirmektedir.



Şekil 3.5. Residual Inception blok mimarisi (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017)

Şekil 3.5'de gösterildiği gibi her bir Inception bloğu bir filtre genişletme katmanını izlemektedir (1 x 1 aktivasyon olmadan evrişim). Bu yapı ile birleştirme öncesinde derinliklerin eşitlenmesi sağlanır. Öncüllerinden farklı olarak Inception-Resnet mimarisinde blok normalizasyonu sadece geleneksel katmanın tepesinde kullanılmıştır. Yapılan testlerde, filtre sayısının 1 000'i aşması durumunda, kalan değişkenler kararsızlıklar sergilemeye başlar ve ağ eğitimin henüz başlarında iken bozulma eğilimi gösterir. Sonuç olarak ortalama havuzlamadan önceki son katmanda yalnızca sıfır vektörleri üretmeye başladığı anlamına gelir. Araştırmaya göre artıkları, bir önceki katmandaki aktivasyonlara eklemeden önce küçültmek, eğitimi stabilize ederek, ölen

ağırlıklar sorununu çözmekteydi. Kalıntı bloklarını ölçeklendirmek için 0,1 ile 0,3 arasındaki faktörleri seçilerek, kurulan yapının optimal sonuç vermesi sağlanmıştır.



Şekil 3.6. Inception-Resnet-V2 mimarisi sıkıştırılmış gösterimi (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017)

Şekil 3.6'da gösterilen Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) mimarisi 164 katmandan oluşan derin konvolüsyonel bir ağdır. Inception yapısı ve kalıntı bağlantısının bir kombinasyonuna dayalı olarak formüle edilmiştir. Inception-Resnet bloğunda, çok boyutlu konvolüsyonel filtreleri artık bağlantılarla birleştirilir. Artık bağlantıların kullanılması, derin yapıların neden olduğu bozulma problemini ortadan kaldırmakla kalmaz, aynı zamanda eğitim süresini de azaltır. Oluşturulan modelin performansının ölçülmesi için ImageNet (Krizhevsky , Sutskever , & Hinton, 2012) veri seti ile eğitimi gerçekleştirilmiştir. Sonuç olarak ağ, çok çeşitli görüntüler için zengin özellik temsillerini öğrenebilmektedir. Elde edilen en yüksek beş doğruluk skorunda %95,3'e kadar çıkabilmektedir (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017).

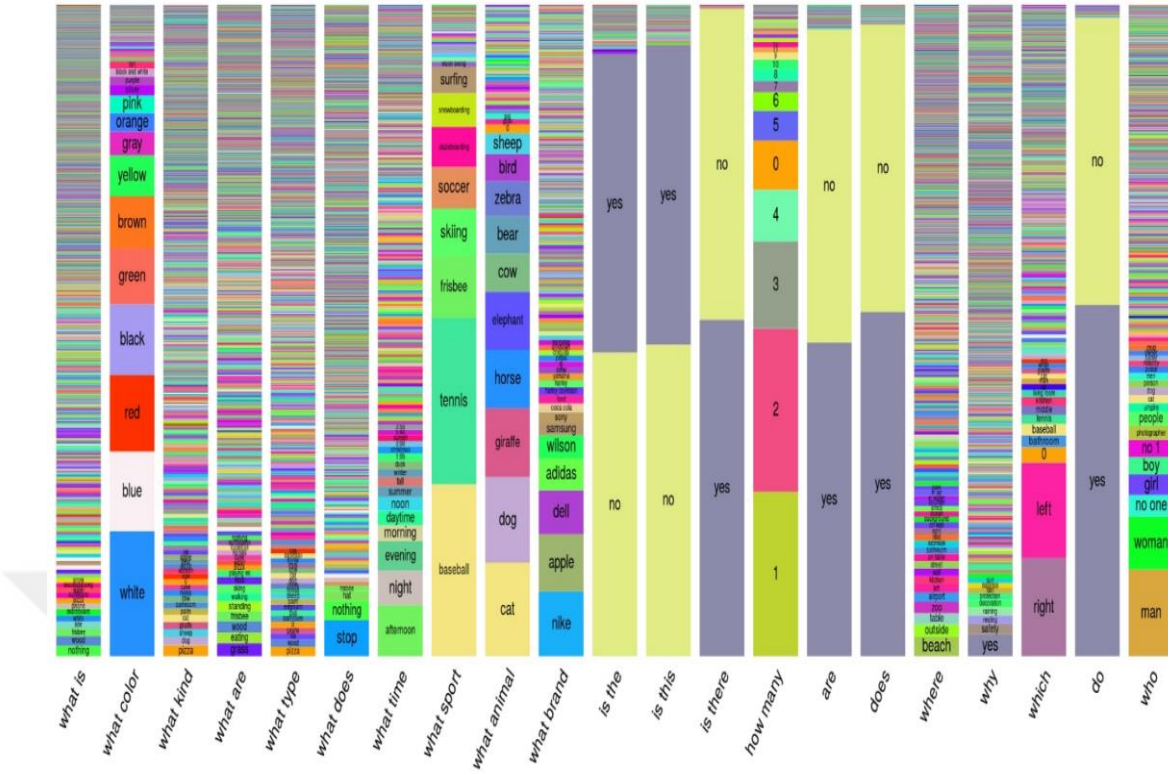
3.3. Görsel Soru Cevaplama Veri Seti VQA

İlk GSC veri seti, MS COCO (Lin , et al., 2014) veri setinden türetilmiştir. MS COCO veri seti 328 000 resim, 91 nesne tipiyle ve 2,5 milyon etiketli örnekler ile çok kolay bir şekilde tanımlanabilir. MS COCO veri seti, GSC veri setinin oluşturulma aşamalarını hızlandırıp, kolaylaştırabilmektedir ancak kolay bir süreç değildir. Örnek olarak kullanışlı ve belirsiz olmayan soruları toplamak oldukça zordur. Çeşitlilik ve kesinlik sorununun yanı sıra iyi bir veri seti taraf (bias) sorununda arındırılmış olmalıdır. Örnek olarak

yanıtların %90'lık kısmının sadece evet olduğu bir evet – hayır yanıtlarına sahip veri kümesinde, en basit yöntemler bile %90'lık bir başarı elde edebilecektir.

VQA-1 veri seti bu bilgiler ışığında oluşturulmuştur. Veri seti içindeki 204 721 adet resim MS COCO (Lin , et al., 2014)(içerisinde insan, kedi, köpek, masa gibi birçok nesneden oluşan görüntüler) veri setinden alınarak her bir resim için üç soru ve on adet cevap ilişkilendirmesi yapılmıştır. İlişkilendirme sonucunda toplam olarak 760 000'den fazla soru ve yaklaşık olarak on milyon adet cevap kümesi bulunmaktadır. Bu sonuca ulaşmak için Amazon Mekanik Türk ekibi ikiye bölünerek bir ekip resimleri oluşturmuş ve diğer ekip soruları ilişkilendirmiştir. Çalışmanın sonucunda açık uçlu ve çoktan seçmeli olarak iki sınıf altında toplanan veri seti elde edilmiştir.

VQA-2 veri seti, VQA-1 veri setinde yapılan çalışmalar sonucunda ortaya çıkan sorunlara ışık tutan bir çalışma sayesinde elde edilmiştir. Yash Goyal ve ekibi, VQA-1 veri setindeki taraf sorunu ve dengesizliğin güçlü derin öğrenme ağlarında tutarsızlığa neden olduğunu ortaya sürmüştür. Bu soruna sundukları çözüm, sorulan sorular için birbirini tamamlayan görsel imgelerin desteklenmesi ve soruların cevaplarının daha dengeli bir hale getirilmesi olmuştur. VQA-2 veri setinde 204 721 COCO görüntüsü, 1,1 milyon soru ve 11 milyon cevap bulunmaktadır. Dengelenmiş veri seti Şekil 3.7'de gösterilmiştir.



Şekil 3.7. VQA-2 veri seti dengelenmiş soru ve cevap analizi (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017)

3.4. Keşifsel Veri Analizi

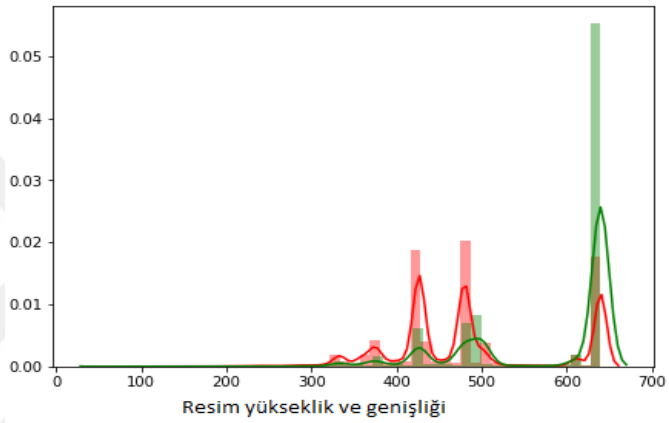


Soru: Yakınlarda orman var mı? (Is there a forest nearby?)
Cevap: evet (yes)

Şekil 3.8. VQA-2 veri setinden alınan örnek görsel ve soru cevap ikilisi (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017)

Keşifsel veri analizi çalışmasında VQA-2 veri seti karakteristikleri ele alınmıştır. İçinde barındırmış olduğu veriden rastgele olarak seçilmiş bir resim, soru ve cevap örneği Şekil 3.8'de verilmiştir.

Daha önce de bahsedildiği gibi VQA-2 veri seti yüksek miktarda görsel içerik barındırmaktadır. Bu yoğunluktaki bir görsel veri setinde her resim farklı boyutlara sahip olabilmektedir. Boyut verisi, özellikle görsel veriden özellik çıkarımı yapacak olan model seçiminde büyük rol oynamaktadır. Şekil 3.9 veri seti içindeki resim boyutlarının dağılımını göstermektedir.



Şekil 3.9. VQA-2 veri içindeki resimlerin ortalama boyut bilgisi

Yapılan analiz sonucunda yüksek ve genişlik boyutunda en fazla 640 x 640 sonuçları elde edilirken, en düşük sonuçlarda 51 x 59 elde edilmiştir. Ortalama olarak ise yükseklik boyutunda 483 piksel genişlik boyutunda ise 578 piksel elde edilmiştir.

VQA-2 veri seti sadece görsel veriden oluşmamaktadır. Metinsel veriler üzerinde gerçekleştirilen keşifsel analiz sonucu Şekil 3.10 'de gösterilmiştir. En uzun soru cümlesi 22 kelimeye sahiptir. En kısa soru cümlesi 2 kelimeden oluşmaktadır. Ortalama olarak bir soru cümlesinde 6 kelime vardır. Sorular, kelime olarak ayırma(tokenizer) işleminden geçirilip benzersiz kelime sayısı hesaplanmıştır. Çıkan sonuçta toplamda 22 226 benzerlik kelime öbeği bulunmaktadır. Şekil 3.10'da elde edilen sonuçlar grafiksel düzlemde gösterilmiştir.



Şekil 3.10. VQA-2 veri setindeki soruların kelime bazında dağılım grafiği

Veri seti içindeki sorular detaylı bir şekilde incelenmiştir. Detaylı incelemenin amacı her bir soru tümcesinin veri içindeki oranlarını elde etmektir. Çizelge 3.1 soru tümceleri ve oranlarını göstermektedir.

Çizelge 3.1. VQA-2 veri seti soru tümcelerinin sayı ve dağılımları

Soru Tipi	**Soru Tipi Sayısı**	**Kapsam**
Ne (What)	182842	%42,9
Dır -Dir (Is)	113774	%26,7
Nasıl (How)	53426	%12,54
Dır -Dir (Are)	33136	%7,78
Nerede (Where)	12409	%2,91
Yap (Does)	12021	%2,82
Hangi (Which)	5382	%1,26
Yap (Do)	4983	%1,17
Neden (Why)	4891	%1,15
Kim (Who)	3322	%0,78

Aynı metodoloji takip edilerek soruların ilişkili olduğu cevaplar üzerinden analiz gerçekleştirilmiştir. Bu analizde cevapların sayıları ve oranları en yüksek frekansa göre Çizelge 3.2 de sunulmuştur.

Çizelge 3.2. VQA-2 veri seti cevapların sayı ve dağılımı

Cevap Tipi	**Cevap Sayısı**	**Kapsam**
Evet (Yes)	84 978	%21,89
Hayır (No)	82 516	%21,26
1	12 540	%3,23
2	12 215	%3,15
Beyaz (White)	8 916	%2,3
3	6 536	%1,68
Mavi (Blue)	5 455	%1,41
Kırmızı (Red)	5 201	%1,34
Siyah (Black)	5 066	%1,31
0	4 977	%1,28

3.5 Yöntem Tasarımı

GSC çalışması klasik makine öğrenmesi yöntemleri ile çözülebilecek bir problem değildir. Yapısı gereği görsel ve metinsel temsillerin bir noktada birleşmesi ve beraber olacak şekilde sonuca ulaştırılması gerekmektedir. Problem dolayısı ile ancak derin öğrenme algoritmaları ile sonuca ulaşabilecek bir seviyeye gelmektedir. Problemin ilk aşamasında VQA-2 veri setinde bulunan görseller eğitim(train) ve doğrulama(validation) olarak ayrı veri yollarında saklanmaktadır. Ardından her biri Xception (Chollet, 2017) ve Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) algoritmaları ile ayrı vektör temsilleri elde edilir. Bu aşamada görsel modellerin son katmanı silinerek, en son katmanındaki, ortalama havuz bilgisi alınır. Bu sayede Xception (Chollet, 2017) için tek bir resim 2 048 boyutunda temsil edilirken. Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) için bir resim 1 536 boyutlu vektör ile temsil edilmektedir.

İkinci aşamada, bahsedildiği gibi metinsel vektörlerin ele alınması vardır. Problemin bu aşamasına Spacy (Matthew, Ines, Landeghem, Adriane, & Adriane, 2020) doğal dil işleme mimarisi kullanılmıştır. Burada her bir kelime vektörü Word2Vec (Mikolov ,

Chen , Corrado , & Dean, 2013), GloVe (Pennington , Socher , & Manning, 2014) ve FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) mimari altyapısı kullanılarak içeri alınmıştır. Spacy, kendi yapısında bulunan özümseme yaklaşımı ile birbirine belli bir eşiğin üzerinde yakınlık gösteren vektörleri aynı düzleme haritalandırma imkânı sunmaktadır. Bu sayede, örnek olarak üç milyon sözlük girdisine sahip olan Word2Vec vektörleri, çalışmayı gerçekleştiren kişinin belirlediği miktara kadar indirilerek haritalandırması gerçekleştirilir. Yapılan keşifsel analiz sonuçlarında her bir sözlük miktarının yarısı kadar özümseme yapıldığında doğruluk metriklerinin, virgülden sonraki ondalık düzlemde değiştiği ancak derin öğrenme modellerinin eğitim ve yorumlama aşamalarında iki kata kadar performans artışı gözlenmiştir.

Sunulan ön işlemler sayesinde elde edilen vektör temsilleri Keras altyapısı kullanılarak derin öğrenme ağında eğitilmiştir. Derin öğrenme ağı, EfficientDet (Gupta , Hooda , & Chikkara, 2020) çalışmasında kullanılan sade ancak güçlü algoritma baz alınarak tasarlanmıştır. Bu tasarım Bi-LSTM blokları ile devam edip, (Gupta , Hooda , & Chikkara, 2020) çalışmasında önerildiği gibi element bazında çarpım işlemleri ile vektörleri birleştirerek, ileri yönlü besleme ağına göndermektedir. Ağın en son çıktısı olarak, 1 000 adet cevap sınıfından en yüksek güven değerine sahip olan değişkenlerin ikili olarak temsil edilmesi (one-hot-encoding) sonucu döndürülür.

LSTM (Hochreiter & Schmidhuber, 1997) katmanı, zaman içerisinde sıralı olan bilgiyi, baştan sonra doğru işlerken, her adımda bir önceki bilgiyi bir sonraki gizli katmana taşımaktadır (Hochreiter & Schmidhuber, 1997). LSTM yapısından girdi olarak bir cümle $S = [w_1, w_2, \dots, w_T]$ olarak temsil edilmektedir. Burada T bir cümlenin toplam uzunluğunu temsil eder. Cümle temsili gömme katmanından geçerek her bir kelime için 300 boyutlu vektörel temsiller elde edilir. Bu temsiller $X' = (X'_1, X'_2, \dots, X'_T)$ olarak ifade edilmektedir. Çalışma LSTM (Hochreiter & Schmidhuber, 1997) mimarisini genişleterek zamanlar bilginin çift yönlü kullanıldığı Bi-LSTM (Schuster & Paliwal, 1997) yapısını kullanmıştır. Bi-LSTM mimarisi, girdi olarak alınan vektörel temsilleri zaman içerisinde hem ileri hem de geri yönde işleyerek birleşik bir temsil oluşturur.

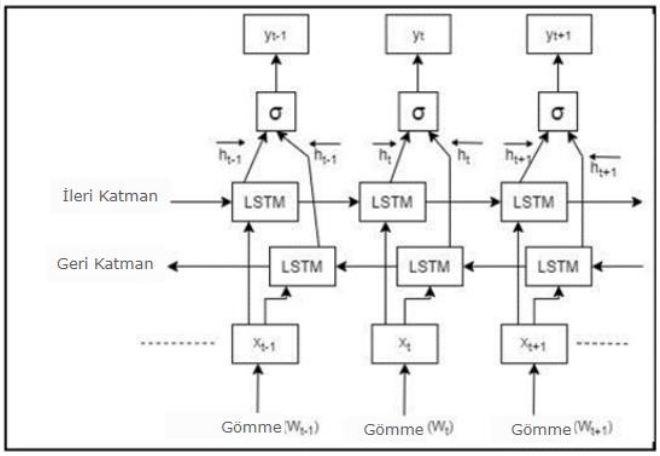
$$\vec{H} = [\overrightarrow{h_o}, \overrightarrow{h_1}, \dots, \overrightarrow{h_T}] \qquad \overleftarrow{H} = [\overleftarrow{h_T}, \overleftarrow{h_{T-1}}, \dots, \overleftarrow{h_0}] \tag{3.5}$$

Denklem 3.5'de Bi-LSTM (Schuster & Paliwal, 1997) mimarisinin veriyi işleyiş gösterimi yapılmıştır. Çift yönlü işlenen veri daha sonra tanh aktivasyon fonksiyonuna gönderilir.

$$h_i = \overrightarrow{h_l} \oplus \overleftarrow{h_l} \tag{3.6}$$

$$y_i = \tanh(h_i + b) \tag{3.7}$$

Denklem 3.6 ve 3.7, birleştirilmiş olan Bi-LSTM (Schuster & Paliwal, 1997) vektörlerinin tanh fonksiyonu ile aktive edilmesini göstermektedir. Elde edilen vektörler görsel veriler ile çarpım gerçekleştirildikten sonra ileri beslemeli ağlar ile son katmana gönderilir. Son katmanda geleneksel softmax çıktısı yerine sigmoid çıktısı kullanılarak, çıktı etiketlerinin olasılık değerlerin daha iyi bir temsili alınmıştır.



Şekil 3.11. Çift yönlü LSTM mimarisi yapısı gösterimi (Schuster & Paliwal, 1997)

Şekil 3.11 de Bi-LSTM (Schuster & Paliwal, 1997) mimarisinin veriyi zaman içinde nasıl ileri ve geri yönlü işlediğinin temsili verilmiştir.

3.6. Word2Vec, GloVe, FastText modelleri Hiper parametreler ve Keras Tuner

Bir derin öğrenme ağında sonucu en çok etkileyen seçimler, parametre seçim ile yapılmıştır. Çalışmamızdaki birçok parametre Goyal'ın (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017) keşifsel çalışmasında önerdiği uzaydan seçilmiştir. Seçilen örneklem uzayı ardından mimari üzerinde Keras-Tuner çerçevesi ile Hiper Bant (Hyperband) istatistiksel yaklaşımları kullanılarak 19 grafik işlemci saati eşdeğerinde araştırma ile son haline getirilmiştir. En yüksek sonuçları veren parametreler Çizelge 3.3'teki gibi elde edilmiştir:

Çizelge 3.3. Kelime Vektörleri modelleri derin öğrenme modeli hiper parametreler

Parametreler	**Değerler**
Yığın Boyutu	512
Döngü	128
Öğrenme Oranı	0,01
LSTM Gizli Nöron Sayısı	512 / 1 024
Cümle Kelime Sayısı	14
Seyreltme	0,5

3.7. BERT modeli Hiper parametreler ve Keras Tuner

Bir derin öğrenme ağında sonucu en çok etkileyen seçimler, parametre seçim ile yapılmıştır. Çalışmamızdaki birçok parametre Goyal'ın (Goyal , Khot , Summers-Stay, Batra , & Parikh, 2017) keşifsel çalışmasında önerdiği uzaydan seçilmiştir. Seçilen örneklem uzayı ardından mimari üzerinde Keras-Tuner çerçevesi ile Hiper Bant (Hyperband) istatistiksel yaklaşımları kullanılarak BERT-Xception modelini 160 grafik işlemci saati eşdeğerinde ve BERT- InceptionRV2 modelini 172 grafik işlemci saati eşdeğerinde toplam 332 grafik işlemci saati eşdeğerinde araştırma ile son haline getirilmiştir. En yüksek sonuçları veren parametreler Çizelge 3.4'teki gibi elde edilmiştir.

Çizelge 3.4. BERT derin öğrenme modeli hiper parametreler

Parametreler	**Değerler**
Yığın Boyutu	512
Döngü	256
Öğrenme Oranı	0,0001
LSTM Gizli Nöron Sayısı	960
Cümle Kelime Sayısı	14
Seyreltme	0,2

Son aşamada VQA-2 veri seti öncül çalışmalarda önerildiği gibi tamamıyla birleştirilmiştir. Ardından akıllı karıştırma (smart-shuffle) yöntemi ile resim – kelime – cevap bilgisi bütünlüğünü bozmadan karıştırılmıştır. Çıktı olarak literatürde sıkça kullanılan oranlar olan 80 – 10 – 10, Eğitim – Doğrulama – Test oranların bölünerek eğitim aşamasında geçilmiştir.

3.8. Doğruluk Metriği

GSC problemi, klasik etiketli verilerde olduğu gibi, doğrulama esnasında etiketlerin eşleşmesi üzerinden hesaplanması gerçek sonuç vermemektedir. Her bir resim – soru ikilisi için on adet aday cevap vardır ve modelin döndürdüğü cevap, aday cevaplar ile korele edilerek elde doğruluk metriği elde edilir.

$$Doğruluk(cevap) = \min\{\frac{ck}{3}, 1\} \tag{3.8}$$

Denklem 3.8'de gösterildiği gibi modelden gelen her bir cevap, insanlardan gelen cevap gibi ele alınır. Her bir soru için potansiyel cevaplarla karşılaştırılarak korelasyon elde edilir. Karşılaştırma esnasında eşleşme ikiden fazla ise doğruluk metriği bir olarak alınır. Eğer ikiden az ise gelen cevap formüldeki gibi ele alınmaktadır. Bir cevabın doğru sayılabilmesi için o küme içerisinde en az üç defa geçmelidir.

4. BULGULAR ve TARTIŞMA

Model eğitimleri, vektörlerin disk üzerinden girdi–çıktı performans kaybı olmadan okunması için 1 TB NVM-e SSD disk, okunan vektörlerin RAM üzerinde tutulabilmesi için 64GB RAM içeren, AMD Ryzen 3700x (8-16) işlemci ve derin öğrenme matris çarpımları için Nvidia GTX 1080 Ti konfigürasyonu üzerinde yapılmıştır. Yazılım mimarisi, Anaconda dağıtımı üzerinde kurulan Python dili ile, Tensorflow-Keras 2.4.1 derin öğrenme çerçevesiyle geliştirilmiştir. Sonuçlar tablo ve grafik halinde sunuldu.

Çizelge 4.1. Xception-Word2Vec döngü başına model performans değerleri

Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,399389	0,335274	1,815769	0,392211
2	1,863038	0,392597	1,683967	0,422423
3	1,754617	0,411620	1,613728	0,426824
4	1,695392	0,424310	1,575155	0,451036
5	1,650404	0,433720	1,544590	0,462707
6	1,615663	0,442877	1,526402	0,458011
7	1,585152	0,450366	1,506743	0,469787
8	1,560290	0,456531	1,504969	0,466847
9	1,539299	0,461629	1,485089	0,476205
10	1,519398	0,466913	1,479511	0,480640
11	1,501761	0,471043	1,461905	0,485615
12	1,483894	0,475674	1,455355	0,485545
13	1,466559	0,481287	1,455015	0,484676
14	1,453194	0,485150	1,455129	0,490729
15	1,434464	0,490857	1,451474	0,490990
16	1,419812	0,494927	1,437148	0,497999
17	1,408075	0,499110	1,448842	0,497721
18	1,392125	0,502702	1,444473	0,497060
19	1,377771	0,506494	1,438027	0,498521
20	1,364566	0,510646	**1,433630**	0,494625

Çizelge 4.2. InceptionRV2-Word2Vec döngü başına model performans değerleri

Döngü	**Kayıp**	**Doğruluk**	**Doğrulama Kayıp**	**Doğrulama Doğruluk**
1	2,375975	0,334230	1,820657	0,393028
2	1,882187	0,384914	1,692342	0,420441
3	1,781301	0,400683	1,634493	0,432390
4	1,726632	0,412111	1,595567	0,444670
5	1,690356	0,420344	1,569765	0,449714
6	1,659298	0,427265	1,553664	0,452254
7	1,635232	0,433300	1,539703	0,454758
8	1,617948	0,436851	1,524413	0,463438
9	1,598498	0,442160	1,519518	0,460272
10	1,581869	0,446272	1,503169	0,465038
11	1,568194	0,449284	1,494575	0,469630
12	1,555252	0,452148	1,489409	0,468499
13	1,543302	0,455337	1,483239	0,470552
14	1,529689	0,459751	1,482461	0,471700
15	1,518552	0,460630	1,479514	0,478014
16	1,506067	0,464739	1,470852	0,478188
17	1,495466	0,468405	1,462407	0,480762
18	1,484426	0,470709	1,466689	0,478449
19	1,474045	0,474180	1,464895	0,483145
20	1,462686	0,478749	1,461841	0,485667
21	1,450966	0,481331	1,456385	0,484450
22	1,440502	0,483592	**1,454687**	0,486085

Çizelge 4.3. Xception-GloVe döngü başına model performans değerleri

Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,411288	0,424101	1,824137	0,393515
2	1,873138	0,433271	1,672688	0,421554
3	1,763481	0,442007	1,616885	0,440130
4	1,699910	0,448998	1,577462	0,453924
5	1,657664	0,455022	1,548101	0,461959
6	1,621563	0,460186	1,531981	0,464273
7	1,591436	0,465879	1,509137	0,468117
8	1,568707	0,471428	1,496526	0,470569
9	1,546328	0,475473	1,485673	0,473300
10	1,525008	0,480525	1,480725	0,478223
11	1,505878	0,485475	1,486276	0,480484
12	1,489529	0,490219	1,466952	0,480031
13	1,471205	0,494608	1,464055	0,478884
14	1,455325	0,497753	1,448035	0,490485
15	1,440596	0,502295	1,465717	0,491216
16	1,425395	0,506962	1,440436	0,496173
17	1,413748	0,510951	1,445962	0,493390
18	1,396666	0,515048	1,446272	0,496399
19	1,381480	0,517593	**1,437312**	0,502313

Çizelge 4.4. InceptionRV2-GloVe döngü başına model performans değerleri

Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,381756	0,331954	1,831774	0,385897
2	1,888936	0,383701	1,708330	0,403221
3	1,787882	0,400273	1,637156	0,431799
4	1,731942	0,411647	1,608160	0,436860
5	1,694413	0,418085	1,581909	0,445522
6	1,662767	0,427243	1,560010	0,438600
7	1,639237	0,433097	1,534082	0,459037
8	1,620167	0,438304	1,521368	0,457733
9	1,602625	0,442185	1,509401	0,459733
10	1,585968	0,447291	1,502180	0,467369
11	1,574461	0,448424	1,494478	0,470221
12	1,558283	0,453182	1,489457	0,471822
13	1,545110	0,456740	1,484586	0,475179
14	1,533943	0,459590	1,482358	0,476483
15	1,523181	0,461585	1,474870	0,478675
16	1,511154	0,464884	1,474475	0,472152
17	1,501142	0,466884	1,465260	0,482832
18	1,489895	0,470274	1,460353	0,483945
19	1,477802	0,475085	1,464809	0,483232
20	1,469322	0,477338	1,460617	0,481301
21	1,459220	0,479460	1,473237	0,487667
22	1,448373	0,483101	**1,459219**	0,486606

Çizelge 4.5. Xception-FastText döngü başına model performans değerleri

Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,369926	0,338627	1,797475	1,797475
2	1,849449	0,395360	1,656615	1,656615
3	1,744347	0,415224	1,603121	1,603121
4	1,685130	0,427470	1,567718	1,567718
5	1,643252	0,437674	1,534150	1,534150
6	1,606638	0,445957	1,536048	1,536048
7	1,580429	0,452239	1,510167	1,510167
8	1,556009	0,458711	1,494769	1,494769
9	1,533831	0,464167	1,479375	1,479375
10	1,514630	0,470268	1,475357	1,475357
11	1,495287	0,474462	1,466839	1,466839
12	1,479120	0,479508	1,452016	1,452016
13	1,463725	0,485523	1,452684	1,452684
14	1,446957	0,489388	1,442792	1,442792
15	1,429986	0,494738	1,443867	1,443867
16	1,414730	0,498738	1,436875	1,436875
17	1,400952	0,502256	**1,433344**	1,433344

Çizelge 4.6. InceptionRV2-FastText döngü başına model performans değerleri

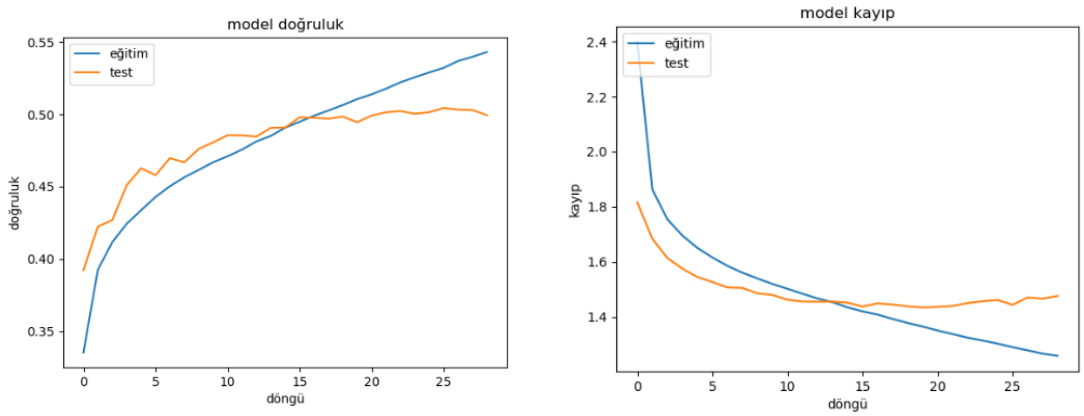
Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,387053	0,333496	1,830490	0,390628
2	1,877163	0,383755	1,685769	0,413257
3	1,774017	0,401361	1,628476	0,430894
4	1,721056	0,413477	1,593945	0,428529
5	1,682920	0,422087	1,561939	0,452323
6	1,651922	0,429097	1,558791	0,445470
7	1,629198	0,434369	1,535170	0,458846
8	1,609663	0,439902	1,512623	0,459194
9	1,592472	0,443839	1,509878	0,463925
10	1,575223	0,447722	1,493411	0,468186
11	1,559905	0,450880	1,496435	0,460985
12	1,547706	0,454908	1,481586	0,474709
13	1,531865	0,457843	1,480698	0,469195
14	1,520858	0,461063	1,474118	0,474796
15	1,508828	0,463023	1,472845	0,476344
16	1,498514	0,467116	1,466676	0,478083
17	1,484608	0,470800	1,462143	0,480432
18	1,476426	0,472377	1,460382	0,480971
19	1,464181	0,475755	1,466838	0,484241
20	1,452770	0,479593	1,460645	0,486137
21	1,442472	0,482782	1,456336	0,490224
22	1,432491	0,485573	**1,451003**	0,489650

Çizelge 4.7. Xception-BERT döngü başına model performans değerleri

Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,584854	0,330184	2,028162	0,388742
2	1,919671	0,404337	1,796862	0,424585
3	1,746602	0,431482	1,695049	0,442304
4	1,657173	0,448397	1,621552	0,456196
5	1,593791	0,461401	1,593718	0,459141
6	1,543327	0,473505	1,569112	0,468708
7	1,498786	0,485294	1,545104	0,476848
8	1,456998	0,496246	1,537535	0,479538
9	1,419513	0,507491	1,536670	0,484559
10	1,382460	0,517271	1,523899	0,488050
11	1,346893	0,527971	1,525202	0,488201
12	1,315134	0,537415	**1,520310**	0,489557

Çizelge 4.8. InceptionRV2-BERT döngü başına model performans değerleri

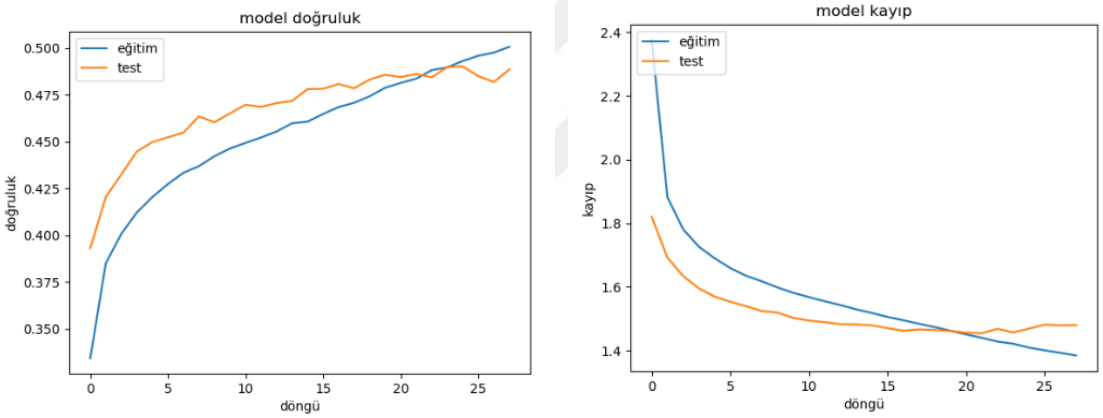
Döngü	Kayıp	Doğruluk	Doğrulama Kayıp	Doğrulama Doğruluk
1	2,562070	0,335177	2,023637	0,399399
2	1,921897	0,402614	1,783033	0,430882
3	1,759791	0,428517	1,684057	0,439428
4	1,673364	0,443960	1,639457	0,451094
5	1,612102	0,455874	1,604109	0,456752
6	1,562217	0,466341	1,573422	0,462063
7	1,523294	0,476489	1,561468	0,470749
8	1,487578	0,485894	1,544046	0,476570
9	1,455220	0,495021	1,542082	0,478460
10	1,425430	0,503020	1,532683	0,483504
11	1,395994	0,512660	**1,531332**	0,485963



(A) (B)

Şekil 4.1. Xception-Word2Vec modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

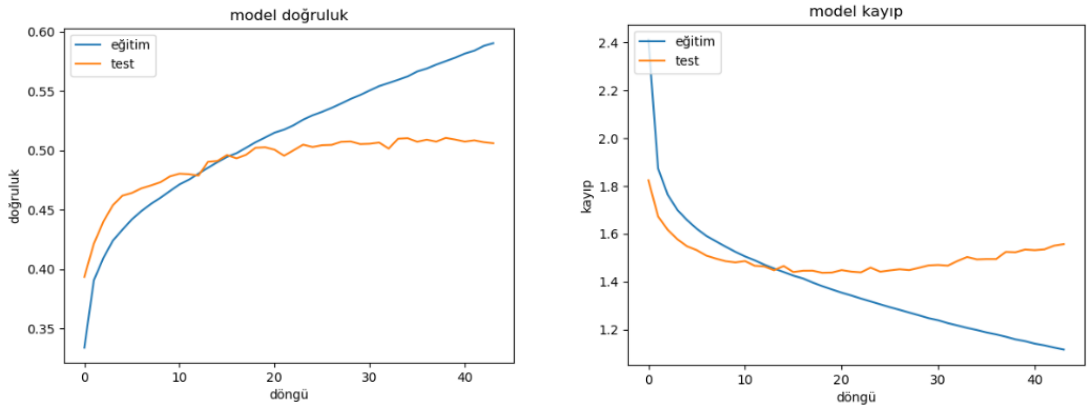
Şekil 4.1'de Xception-Word2Vec modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.2. InceptionRV2-Word2Vec modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

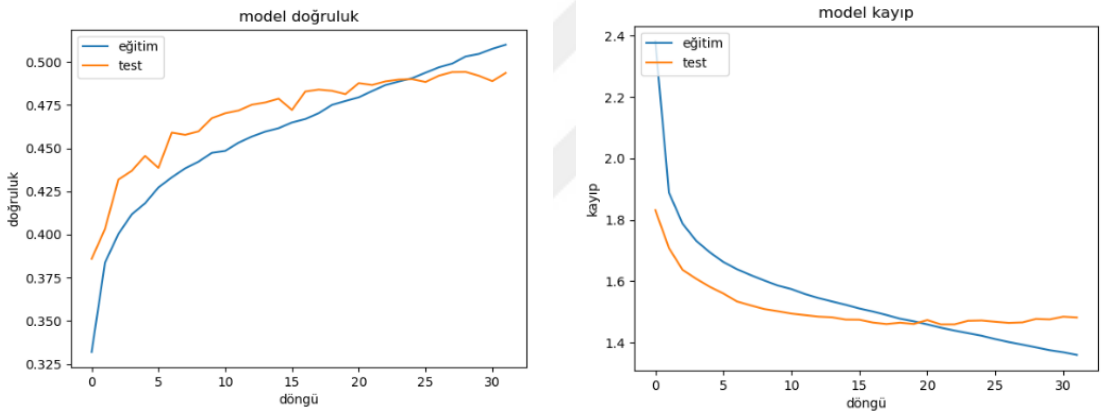
Şekil 4.2'de InceptionRV2-Word2Vec modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.3. Xception-GloVe modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

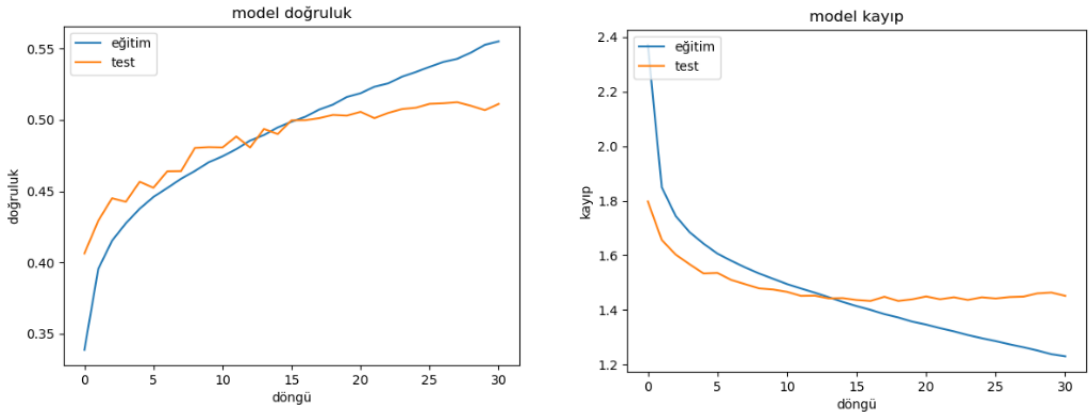
Şekil 4.3'de Xception-GloVe modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.4. InceptionRV2-GloVe modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

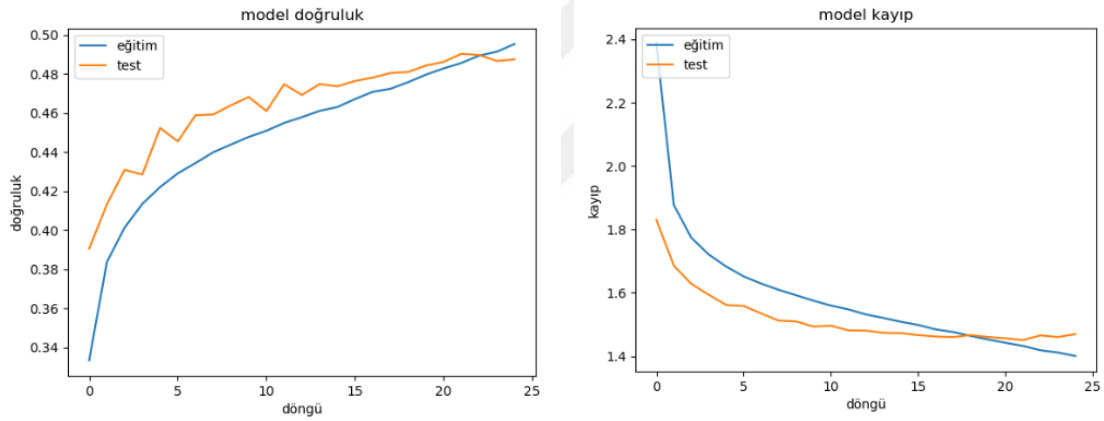
Şekil 4.4'de InceptionRV2-GloVe modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.5 Xception-FastText modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

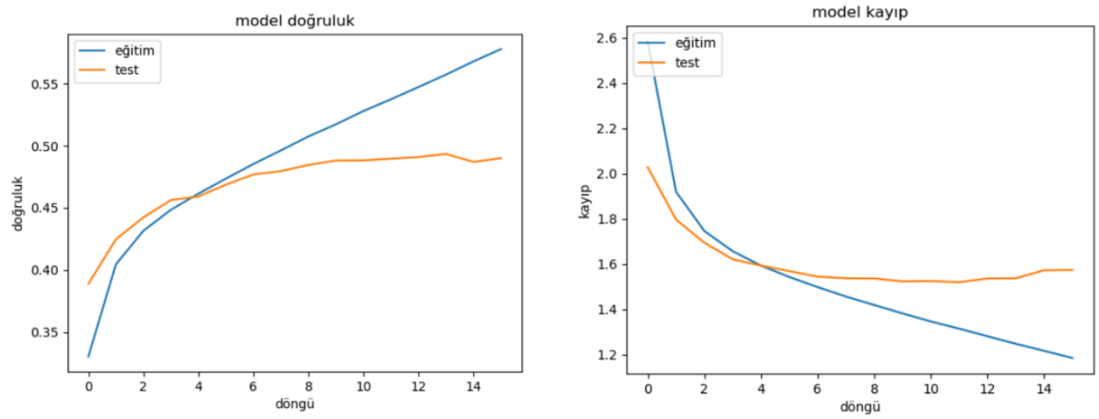
Şekil 4.5'de Xception-FastText modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.6. InceptionRV2-FastText modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

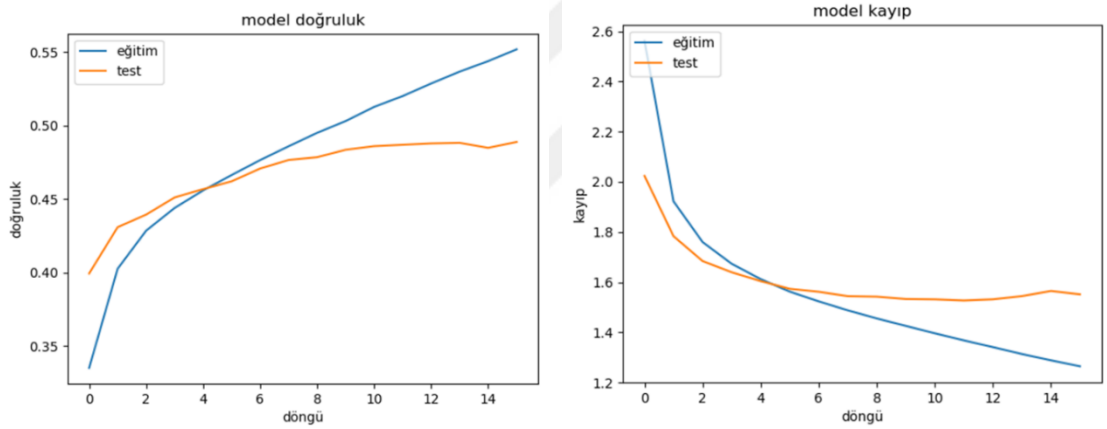
Şekil 4.6'da InceptionRV2-FastText modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.7. Xception-BERT modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

Şekil 4.7'de Xception-BERT modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.



(A) (B)

Şekil 4.8. InceptionRV2-BERT modeli eğitim A) Doğruluk B**)** Kayıp grafikleri

Şekil 4.8'de InceptionRV2-BERT modeline ait eğitim esnasında elde edilen doğruluk ve kayıp değerleri verilmiştir.

Çizelge 4.9. Xception ve kelime vektörlerine dayalı doğruluk çıktıları.

	Eğitim	**Doğrulama**	**Test**
Xception-Word2Vec	51,06	49,46	54,97
Xception-GloVe	51,75	**50,23**	53,51
Xception-FastText	50,22	43,33	55,26
Xception-BERT	**53,74**	48,95	**59.00**

Çizelge 4.9'da göstertildiği gibi eğitilmiş olan modellerin başarım değerleri gösterilmiştir. Elde edilen sonuçlar içinde Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) diğer üç modele yakın sonuç çıkmıştır. GloVe (Pennington , Socher , & Manning, 2014) modeli, FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) modeline rekabetçi bir başarım sağlamasına rağmen açık bir şekilde aşırı öğrenme (overfit) sorunu yaşamaktadır. Xception (Chollet, 2017) vektörleri ile elde edilen sonuçlarda kelime gömme vektörlerinden en iyi performansı FastText vektörleri göstermektedir. Bunun en açık sebebi Mikolov'un araştırmasında (Bojanowski , Grave , Joulin , & Mikolov, 2017) bahsettiği gibi, vektörler oluşturulurken alt-kelime anlamlarının çıkarılabilmesi için ön işlemlerin titizlikle ele alınmış olmasıdır. Bu çalışmada ele alınan BERT (Devlin , Chang , Lee , & Toutanova, 2019) yönteminin kelime vektörleri yöntemlerinden daha başarılı olduğu görülmektedir. Bunun temelinde BERT'in çok başlı ilgi yapısı ile kelimeler arasındaki alt bağlamı öğrenmekte ve temsil etmektedir.

Çizelge 4.10. Inception-Resnet-V2 ve kelime vektörlerine dayalı doğruluk çıktıları.

	Eğitim	**Doğrulama**	**Test**
InceptionRV2-Word2Vec	48,35	48,60	55,50
InceptionRV2-GloVe	48,31	48,66	55,06
InceptionRV2-FastText	48,55	**48,96**	55,42
InceptionRV2-BERT	**51,26**	48,59	**58,68**

Çizelge 4.10'da Inception-Resnet-V2 (Szegedy , Ioffe , Vanhoucke , & Alemi, 2017) görsel modeli ile eğitilmiş olan GSC modelleri çıktıları açık bir şekilde düşük

performansa işaret etmektedir. Bunun iki sebebi olduğu varsayılmaktadır. İlk olarak Xception (Chollet, 2017) modeli, ortalama havuzlama çıktısı olarak her bir görsel için 2 048 boyutlu vektör üretirken, Inception-Resnet-V2 modeli 1 536 boyutlu vektör üretmektedir. Verilen değerlerden görüldüğü gibi Xception mimarisi görsel başına daha çok bilgi üretmekte ve temsil etmektedir. Bir diğer sorun, kelime vektörlerinin görsel model çıktılarını genelleştirememesidir. Word2Vec (Mikolov , Chen , Corrado , & Dean, 2013) vektörleri test setinde az bir fark ile yukarıya çıkmıştır. Benzer durum FastText (Bojanowski , Grave , Joulin , & Mikolov, 2017) vektörlerinde gözlemlenmektedir. Ancak görsel vektör değişiminde kelime gömme yöntemlerinde performansını en iyi şekilde Word2Vec (Pennington , Socher , & Manning, 2014) vektörleri korumaktadır. Bu durumu, kelime gömmelerini oluşturulurken, ağırlıklı kareler ile istatistiksel yaklaşımı beraber kullanabilmesiyle açıklayabilmekteyiz. Bu sayede, çoklu girdi modellerde (görsel – metinsel ya da görsel – metinsel – işitsel) oluşan değişikliklere karşı daha az duyarlı olduğunu söyleyebiliriz. BERT (Devlin , Chang , Lee , & Toutanova, 2019) yönteminin Inception-Resnet-V2 modeli ile diğer yöntemlerden daha yüksek başarıya uğradığı görülmektedir. Eğitim ve test sonucunda en yüksek doğruluk elde edilmiştir. Çalışmada göze çarpan bir diğer detay test setinin büyüklüğü ya da derişiminin geliştirilmeye müsait olduğudur. Bu şekilde daha rekabetçi model test durumlarının elde edilebileceği düşünülmektedir.

Çalışma kapsamında veri seti dışında rastgele bir görsel ve görsel ile ilgili soru sorularak modelin cevap üretmesi sağlanılmıştır. Aşağıdaki iki örnek verilmiştir. Şekil 4.9'da modelin görsel ile ilgili sorusuna doğru cevap verdiği görülürken, Şekil 4.10'da model yanlış cevap vermektedir.



Soru: Köpeğin ağzında ne var? (What is in the dog's mouth?)
Cevap: frizbi(frisbee)

Şekil 4.9. Modelin doğru cevap verdiği örnek

Şekil 4.9'da görüldüğü üzere ağzında frizbi olan bir köpek görseli ve görsel ile ilgili köpeğin ağzında ne olduğu sorulmaktadır. Bu soru için oluşturulan aday cevap kümesi "frizbi(frisbee), frizbi(frisbee), top(ball), frizbi(frisbee), frizbi(frisbee), frizbi(frisbee), frizbi(frisbee), top(ball), frizbi(frisbee), frizbi(frisbee)" şeklindedir. Modelden gelen cevap "frizbi(frisbee)" şeklindedir ve doğrulama metriğine göre cevap, verilen küme içerisinde sekiz defa geçtiğinden verilen cevap doğru kabul edilmektedir.



Soru: Kedinin düşme tehlikesi var mı? (Is the cat in danger of falling?)
Cevap: Hayır(No)

Şekil 4.10. Modelin yanlış cevap verdiği örnek

Şekil 4.10'da balkon demirinin üzerinde duran kedi görseli ve görsel ile ilgili kedinin düşme tehlikesi var mı sorusu sorulmaktadır. Bu soru için oluşturulan aday cevap kümesi

“evet(yes), evet(yes), hayır(no), evet(yes), evet(yes), evet(yes), evet(yes), hayır(no), evet(yes), evet(yes)” olarak tanımlanmıştır. Modelden gelen cevap “hayır(no)” şeklindedir ve doğrulama metriğine göre cevap, verilen küme içerisinde iki defa geçtiğinden verilen cevap yanlış kabul edilmektedir. Bu görselde karanlık olduğu için kedinin balkon demiri yerine düz zeminde duruyormuş gibi algılamış olabileceği tahmin edilmektedir. Modelde kullanılan veri setinde her görsel için bir soru ve on aday cevap oluşturulduğu bilinmektedir ve model tarafından üretilen cevap bu cevaplar ile karşılaştırılmaktadır. Tekil resim ve soru sorulduğunda (örneklerdeki gibi) sorulan sorunun cevabı bilindiğinden cevap kümesinin oluşturulup karşılaştırılmasına gerek olmadan, verilen cevabın doğru olup olmadığı görülebilmektedir.

5. SONUÇ

Sunulan çalışma, GSC problemi için Xception, Inception-Resnet-V2 görsel vektörel modelleri, ile alanında kendini kanıtlamış ön eğitimli kelime vektörleri olan Word2Vec, GloVe, FastText'i Bi-LSTM ağlarını kullanarak eğitmiştir. Modeller ortalama 30-35 döngü değerinde modeller yerel maksimum değerlerine ulaşmıştır ve her bir model bir saat yirmi dakika ile bir buçuk saat arasında değişen sürelerde eğitilmektedir.

Elde edilen bulgular, kullanılan mimarinin temel aldığı *Natural Language Processing based Visual Question Answering Efficient: an Efficient Approach* (Gupta , Hooda , & Chikkara, 2020) çalışmasına yaklaşan sonuçlar elde etmiştir. En iyi çıktı olarak kabul edilen Xception-GloVe mimarisi ile alınan sonuçlar ile Gupta'nın çalışmasından %5 geride doğruluk elde ederek literatür seviyesinde çıktı elde edilmiştir.

Elde edilen sonuçlarda, Xception modelinin yüksek boyutlu vektör çıktısı etkisini oldukça göstermektedir. Kelime vektörleri tarafında ise FastText ve Word2Vec en yüksek sonuçları elde etmiştir. GloVe vektörleri değişen görsel mimarilerin farklı boyutlardaki çıktılarına daha dayanıklı vektör çıktıları sunmakta iken, FastText kelime vektörleri en yüksek sonucu elde etmektedir.

Çalışma kapsamında öne atılan fikir doğrultusunda GSC problemi için Xception, Inception-Resnet-V2 görsel vektörel modelleri ile BERT'i Bi-LSTM ağları kullanılarak eğitilmiştir. Modeller ortalama 10-15 döngü değerinde modeller yerel maksimum değerlerine ulaştığı görüldü ve her bir model beş saat ile sekiz saat arasında değişen sürelerde eğitilmektedir.

Görsel vektörel modellerinden bağımsız olarak BERT modelinde daha iyi sonuç verdiği görüldü. BERT modelinin birçok problemde başarı gösterdiği bilinmektedir ve bu etkisi GSC probleminde de görülmüştür. GSC probleminde BERT modelinin etkisini daha iyi analiz edebilmek için sade fakat güçlü bir model kullanılarak BERT'in performansı ve diğer literatürde yaygın olarak kullanılan modeller karşılaştırıldığında BERT yaklaşık olarak %3 daha fazla doğruluk elde edilmiştir. Bu çalışma ile GSC probleminde ön eğitimli kelime gömme vektörleri yerine BERT'in kullanılmasının doğruluğu arttırdığı gösterilmiştir.

Sunulan tez çalışmasının geliştirilmesi için gelecek çalışmalarda bağlamsal vektörler ile ön eğitimli vektörlerin birleştirilerek yüksek boyutlu bileşik kelime vektörleri ile analiz yapılması. Görsel ilgi ve metinsel ilgi üzerine odaklanılabilmesi. Son olarak, görsel – metinsel vektörlerin birleştirilmesi aşamasında çok modalite füzyon modeli eklenebilmesi örnek olarak verilebilir. Ancak bu yaklaşımlar yüksek boyutlu hesaplamalar gerektirdiği için, çoklu ekran kartı temelinde bir altyapı gerektirmektedir.

KAYNAKLAR

Anderson, P., He, X., Buehler, C., Teney , D., Johnson, M., Gould, S., & Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6077-6086.

Biten, A. F., Tito, R. P., Mafla, A., & Rusiol, M. 2019. Scene Text Visual Question Answering. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 4290-4300.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 135-146.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., & Dhariwal, P. 2020. Language Models are Few-Shot Learners. ArXiv.

Chollet, F. 2017. Xception: Deep Learning with Depthwise Separable Convolutions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1800-1807.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *ArXiv*.

Devlin, J., Chang, M.-W., Lee , K., & Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv.

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. EMNLP.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. 2017. Making the V in VQA Matter Elevating the Role of Image Understanding in Visual Question Answering. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6325-6334.

Gupta, R., Hooda, P., & Chikkara, N. K. 2020. Natural Language Processing based Visual Question Answering Efficient: an EfficientDet Approach. *4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 900-904.

Hochreiter, S., & Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Computation*, 1735-1780.

Huang, P., Huang, J., Guo, Y., Qiao, M., & Zhu, Y. 2019. Multi-grained Attention with Object-level Grounding for Visual Question Answering. The Association for Computational Linguistics.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 84 - 90.

L. P., Yang, Y., Zhang, X., Ji , Y., Lu, H., & Shen, H. T. 2020. Answer Again: Improving VQA with Cascaded-Answering Model. *IEEE Transactions on Knowledge and Data Engineering*, 1-1.

Le, Q. V., & Mikolov, T. 2014. Distributed Representations of Sentences and Documents. *ArXiv*.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998. Gradient-based learning applied to document recognition.

Liang, J., Jiang, L., Cao, L., Li, L.-J., & Hauptmann, A. 2018. Focal Visual-Text Attention for Visual Question Answering. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6135-6143.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Zitnick, L. 2014. *Microsoft COCO Common Objects in Context.*

Malinowski, M., Rohrbach, M., & Fritz, M. 2015. Ask Your Neurons a Neural-Based Approach to Answering Questions about Images. *2015 IEEE International Conference on Computer Vision*, 1-9.

Malinowski, M., & Fritz, M. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. NIPS.

Mathew, M., Karatzas, D., Manmatha, R., & Jawahar, C. 2021. DocVQA A Dataset for VQA on Document Images. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2199-2208.

Matthew, H., Ines, M., Landeghem, V., Adriane, S., & Adriane, B. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* Retrieved from zenodo: https://doi.org/10.5281/zenodo.1212303

Mikolov, T., Chen , K., Corrado, G. S., & Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. International Conference on Learning Representations (ICLR).

Pennington, J., Socher, R., & Manning, C. D. 2014. GloVe Global Vectors for Word Representation. EMNLP.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. 2018. Deep Contextualized Word Representations. North American Chapter of the Association for Computational Linguistics.

Ren, S., He, K., Girshick , R. B., & Sun, J. 2015. Faster R-CNN Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1137-1149.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 211-252.

Schuster, M., & Paliwal, K. 1997. Bidirectional recurrent neural networks. *IEEE*.

Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. 2017. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. Association for the Advancement of Artificial Intelligence.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Rabinovich, A. 2015. Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.

Tan, M., Pang, R., & Le, Q. V. 2020. EfficientDet Scalable and Efficient Object Detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10778-10787.

Teney, D., Anderson, P., He , X., & Hengel, A. v. 2018. Tips and Tricks for Visual Question Answering Learnings from the 2017 Challenge. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4223-4232.

Vaswani, A., Shazeer, N. M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. 2017. Attention is All you Need. ArXiv.

Zeng, D., Zhou, G., & Wang, J. 2019. Residual Self-Attention for Visual Question Answering. *1st International Conference on Electrical, Control and Instrumentation Engineering (ICECIE)*, 1-7.

Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., & Parikh, D. 2016. Yin and Yang Balancing and Answering Binary Visual Questions. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5014-5022.