Journal of the Faculty of Engineering and Architecture of Gazi University 39:4 (2024) 2177-2186

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University
Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

An efficient activity recognition model by integrating object recognition and image captioning with deep learning techniques for the visually impaired

Zeynep Hilal Kilimci[1]*, Ayhan Küçükmanisa[2]
[1]Department of Information Systems Engineering, Faculty of Technology, Kocaeli University, 41001, İzmit, Kocaeli, Türkiye
[2]Department of Electronics and Communication Engineering, Faculty of Engineering, Kocaeli University, 41001, İzmit, Kocaeli, Türkiye

Highlights:

- High performance activity recognition model
- Constructing image captions using deep learning architectures
- Converting the recognized activity from image texts to sound for visually impaired

Keywords:

- Activity recognition
- Deep learning models
- Image caption generator
- Feature injection techniques
- Long short term memory networks

Article Info:

Research Article
Received: 31.01.2023
Accepted: 17.09.2023

DOI:

10.17341/gazimmfd.1245400

Correspondence:

Author: Zeynep Hilal Kilimci
e-mail: zeynep.kilimci@kocaeli.edu.tr
phone: +90 262 303 2242

Graphical/Tabular Abstract

Automatically identifying the content of an image is a core task in artificial intelligence that connects computer vision and natural language processing. In this study, a generative model based on a deep and recurrent architecture is proposed, combining the latest developments in computer vision and machine translation, to create natural sentences describing an image. With this model, the texts obtained from the images can be converted into audio file format and the activity of the objects around the person can be defined for visually impaired people. For this purpose, first, object recognition is performed on images with the YOLO model, which identifies the presence, location and type of one or more objects in a particular image. Next, long-short-term memory networks (LSTM) are trained to maximize the probability of the target statement sentence given the training image. Thus, the activities in the related image have been converted to text format as annotations. The activities, which are converted to text format, are obtained by using the Google text-to-speech platform, and the audio file describing the activity is obtained. Flickr8K, Flickr30K and MSCOCO datasets are employed evaluating four different features injection architectures to demonstrate the effectiveness of the proposed model. The experimental results show that the proposed model is successful in expressing the activity description audibly for visually impaired individuals. Table A demonstrates the performance of the proposed model in terms of BLEU-1 score in Flickr8K, Flickr30K, and MSCOCO datasets when different features injection techniques are included.

Table A. Performance of proposed model with different features injection techniques in terms of BLEU-1 score in Flickr8K, Flickr30K, and MSCOCO datasets.

Dataset	par-inject	init-inject	pre-inject	merge
Flickr8K	0.6345	0.6239	0.5469	0.5371
Flickr30K	0.6546	0.6370	0.6036	0.5922
MSCOCO	0.6433	0.6807	0.6584	0.6330

Purpose:
The purpose of this study is to construct an efficient activity recognition model by integrating object recognition and image captioning with deep learning techniques for the visually impaired.

Theory and Methods:
For object recognition task, YOLO model is employed to identify the presence, location and type of one or more objects in a particular image. Next, long-short-term memory networks (LSTM) are trained to maximize the probability of the target statement sentence given the training image. Thus, the activities in the related image have been converted to text format as annotations. The activities, which are converted to text format, are obtained by using the Google text-to-speech platform, and the audio file describing the activity is obtained. Flickr8K, Flickr30K and MSCOCO datasets are employed by assessing par-inject, pre-inject, init-inject, and merge features injection architectures to show the effeciency of the proposed model.

Results:
In Flickr8K and Flickr30K datasets, par-inject technique outperforms other models with 0.6345 and 0.6546 BLEU-1 scores, respectively while init-inject method performs better performance in MSCOCO dataset by ensuring 0.6807 BLEU-1 result.

Conclusion:
As a result, the performance of the proposed model is significantly superior compared to the state-of-the-art studies in terms of BLEU-1 score when different features injection techniques are considered.

Journal of the Faculty of Engineering and Architecture of Gazi University 39:4 (2024) 2177-2186

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Görme engelliler için nesne tanıma ve resim altyazısını derin öğrenme teknikleriyle entegre eden verimli bir aktivite tanıma modeli

Zeynep Hilal Kilimci[1]*, Ayhan Küçükmanisa[2]
[1]Kocaeli Üniversitesi, Teknoloji Fakültesi, Bilişim Sistemleri Mühendisliği Bölümü, 41001, İzmit, Kocaeli, Türkiye
[2]Kocaeli Üniversitesi, Mühendislik Fakültesi, Elektronik Haberleşme Mühendisliği Bölümü, 41001, İzmit, Kocaeli, Türkiye

ÖNEÇIKANLAR

- Yüksek performanslı aktivite tanıma modeli
- Derin öğrenme mimarileri kullanarak görüntü altyazısı oluşturma
- Görme engelliler için görüntü metinlerinden tanınan aktivitenin sese dönüştürülmesi

Makale Bilgileri
Araştırma Makalesi
Geliş: 31.01.2023
Kabul: 17.09.2023

DOI:
10.17341/gazimmfd.1245400

Anahtar Kelimeler:
Aktivite tanıma, derin öğrenme modelleri, görüntü altyazısı üretici, özellik enjeksiyon teknikleri, uzun kısa dönem hafıza ağları

ÖZ

Bir görüntünün içeriğini otomatik olarak tanımlamak, bilgisayarla görmeyi ve doğal dil işlemeyi birbirine bağlayan yapay zekadaki temel bir görevdir. Bu çalışmada, bilgisayarla görü ve makine çevirisindeki son gelişmeleri birleştiren ve bir görüntüyü tanımlayan doğal cümleler oluşturmak için derin ve tekrarlayan bir mimariye dayalı üretken bir model sunulmuştur. Oluşturulan bu model ile görüntülerden elde edilen metinler, ses dosyası formatına dönüştürülebilmekte ve görme engelli insanlar için kişinin etrafında bulunan nesnelerin aktivitesi tanımlanabilmektedir. Bu amaçla, ilk olarak, belirli bir görüntüdeki bir veya daha fazla nesnenin varlığını, konumunu ve türünü tanımlayan YOLO modeliyle görüntüler üzerinde nesne tanıma işlemi gerçekleştirilmiştir. Sonrasında, uzun kısa dönem hafıza ağları (LSTM) eğitim görüntüsü verilen hedef açıklama cümlesinin olasılığını en üst düzeye çıkarmak için eğitilmiştir. Böylece, ilgili görüntü içerisinde yer alan aktiviteler, açıklama olarak metin biçimine dönüştürülmüştür. Metin biçimine dönüştürülen aktiviteler, Google metin okuma platformundan faydalanılarak aktiviteyi tanımlayan ses dosyaları elde edilmiştir. Önerilen modelin etkinliğini göstermek amacıyla dört farklı özellik enjeksiyon mimarisi değerlendirilerek Flickr8K, Flickr30K ve MSCOCO veri kümeleri kullanılmıştır. Deney sonuçları, önerdiğimiz modelin görme engelli bireyler için aktivite tanımlamayı sesli olarak ifade etmede başarılı olduğunu göstermiştir.

An efficient activity recognition model by integrating object recognition and image captioning with deep learning techniques for the visually impaired

HIGHLIGHTS

- High-performance activity recognition model
- Constructing image captions using deep-learning architectures
- Converting the recognized activity from image texts to sound for visually impaired

Article Info
Research Article
Received: 31.01.2023
Accepted: 17.09.2023

DOI:
10.17341/gazimmfd.1245400

Keywords:
Activity recognition, deep learning models, image caption generator, feature injection techniques, long short-term memory networks

ABSTRACT

Automatically identifying the content of an image is a core task in artificial intelligence that connects computer vision and natural language processing. This study presents a generative model based on a deep and recurrent architecture, combining the latest developments in computer vision and machine translation, to create natural sentences describing an image. With this model, the texts obtained from the images can be converted into audio file format, and the activity of the objects around the person can be defined for visually impaired people. For this purpose, first, object recognition is performed on images with the YOLO model, which identifies the presence, location and type of one or more objects in a particular image. Next, long-short-term memory networks (LSTM) are trained to maximize the probability of the target statement sentence given the training image. Thus, the activities in the related image have been converted to text format as annotations. The activities, which are converted to text format, are obtained using the Google text-to-speech platform, and the audio file describing the activity is obtained. Flickr8K, Flickr30K and MSCOCO datasets are employed to evaluate four different features injection architectures to demonstrate the effectiveness of the proposed model. The experimental results show that our proposed model successfully expresses the activity description audibly for visually impaired individuals.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *zeynep.kilimci@kocaeli.edu.tr, ayhan.kucukmanisa@kocaeli.edu.tr / Tel: +90 262 303 2242

1. Giriş (Introduction)

Görme engelli bireyler, görme duyusu eksikliği nedeniyle yaşantıları boyunca birçok zorlukla karşı karşıyadır. Günlük yaşantı içerisinde karşılaştıkları sorunları çözebilmek için başka insanlara, rehber köpeklere veya onlara destek olabilecek alternatif çözümlere ihtiyaç duymaktadırlar. Fakat bu çözümler, kimi zaman sürdürülebilir ve etkin çözümler olmayabilmektedir. Bu çalışmada, görme engelli insanlar için son teknoloji yöntemleri kullanarak görme engellerini en aza indirebilecek bir çözüm üretilmesi hedeflenmiştir. Bir başka deyişle, çalışma kapsamında görme engelli bireylerin karşılaştığı sorunların çevrelerindeki aktiviteleri sesli bir asistan aracılığıyla duyma duyusu vasıtasıyla giderilmesi amaçlanmıştır.

Bir görüntünün açıklamasını oluşturmaya görüntü yazısı veya görüntü altyazısı adı verilir. Görüntü altyazısı oluşturma, önemli nesnelerin, niteliklerinin ve bir resimdeki ilişkilerinin tanınmasını gerektirir. Ayrıca sözdizimsel ve anlamsal olarak doğru cümleler oluşturması gerekir. Çoğu çalışmada, bu yöntemin kullanılmasının sebebi görüntünün anlamlandırılıp metinlere dönüştürülmesini sağlamaktır [1-4]. Bunun yanı sıra, uygulanan nesne algılama yöntemi, dijital görüntülerde ve videolarda belirli bir sınıftaki (insanlar, binalar veya arabalar gibi) anlamsal nesnelerin örneklerini tespit etmekle ilgilenen bilgisayarla görme ve görüntü işleme ile ilgili bir bilgisayar teknolojisidir. Böylece, görüntüde bulunan nesnelerin algılanmasını sağlanmaktadır. Görüntü altyazısı oluşturma teknolojisi, görüntüyü anlamlandırmada oldukça başarılı olsa da görüntüde önemli olabilecek bir obje, cümle içerisinde kullanılmayabilir. Bu sorunun giderilmesi açısından nesne algılama teknolojisi de çalışmamızda kullanılmıştır.

Son yıllarda önemi artan makine öğrenmesi ve derin öğrenme metodolojilerinin etki ettiği farklı yöntemlerin günümüz ihtiyaçlarına çözümler sunması sayesinde farklı araştırmaların önü açılmıştır. Derin öğrenme, yapay sinir ağları ile beynin yapısını, işlevini ve öğrenme şeklini taklit eden bir mimari olmakla beraber bu çalışmada önerilen modelin temelini oluşturmaktadır. Derin öğrenme yöntemlerinin son dönemlerde çokça tercih edilmesinin ardında yatan nedenlerden bazıları, kolay ölçeklenebilirliği ve donanımsal olarak avantaj sağlamalarıdır. Bunlara ek olarak, derin öğrenme modellerinin sunmuş olduğu bir başka avantaj ise özellik çıkarımını ham verilerden herhangi bir dış müdahale olmadan otomatik olarak gerçekleştirebilmesidir. Evrişimsel sinir ağları (Convolutional Neural Network-CNN), tekrarlayan sinir ağları (Recurrent Neural Network-RNN), uzun kısa dönem hafıza ağları (Long Short Term Memory-LSTM), derin sinir ağları (Deep Neural Network-DNN) farklı alanlarda uygulanan ve sıklıkla tercih edilen derin öğrenme metodolojilerinden bazılarıdır [5-8]. Derin öğrenme modelleri, görüntü tanıma ve sınıflandırma [9], duygu analizi [10], aktivite tanıma [11], ses tanıma [12, 13], metin sınıflandırma [14] gibi birçok alanda kullanılmaktadır.

Bu çalışma, görüntü altyazısı oluşturma, nesne algılama, metni sese dönüştürme olmak üzere temelde üç aşamadan oluşmaktadır. İlk aşamada, bir görüntüdeki bir veya daha fazla nesnenin varlığını ve türünü tanımlayan YOLO modeliyle görüntüler üzerinde nesne tanıma işlemi gerçekleştirilmiştir. İkinci aşamada, uzun kısa dönem hafıza ağları (LSTM) kullanılarak eğitim görüntüsü, verilen hedef açıklama cümlesinin olasılığını en üst düzeye çıkarmak için eğitim işlemi gerçeklenmiştir. Bu sayede, ilgili görüntü içerisinde yer alan aktiviteler, açıklama olarak metin biçiminde ifade edilmiştir. Son aşamada, metin biçimine dönüştürülen aktiviteler, Google metin okuma platformundan faydalanılarak aktiviteyi tanımlayan ses dosyaları elde edilmiştir. Önerilen modelin etkinliğini kanıtlamak amacıyla dört farklı özellik enjeksiyon mimarisi Flickr8K, Flickr30K ve MSCOCO veri kümeleri üzerinde kullanılmıştır. Deney sonuçları, önerdiğimiz modelin görme engelli bireyler için aktivite tanımlamayı sesli olarak ifade etmede başarılı olduğunu göstermiştir. Önerilen yöntemin katkıları aşağıdaki gibidir:

- Aktivite tanıma işlevi için farklı özellik enjeksiyon yöntemleri değerlendirilmiştir.
- Yüksek performanslı derin öğrenme temelli aktivite tanıma modeli geliştirilmiştir.
- Görme engelliler için sadece tek bir görüntü girdi olarak alınarak tanınan aktivitenin sese dönüştürülmesi uçtan-uca bir şekilde gerçekleştirilmiştir.

Makalenin kalan kısmı şu şekilde düzenlenmiştir: Bölüm 2, resim altyazısı oluşturma ve nesne tanıma alanlarında yapılan çalışmaların özetini sunmaktadır. Bölüm 3, görüntü tanıma, altyazı oluşturma, özellik enjeksiyon teknikleri gibi sistemin inşası için kullanılan yöntemleri ve önerilen modelin mimarisini içermektedir. Deney sonuçları ise Bölüm 4 ve Bölüm 5' te sunulmaktadır.

2. Literatür Çalışmaları (Related Works)

Bu bölümde, resim altyazısı ve nesne algılama konularında yapılan literatür çalışmalarının bir özetini sunulmaktadır.

Yang vd. [15] nesne algılama ve yerelleştirme ile görüntü alt yazısı oluşturma alanında bir model önermişlerdir. Görüntülerin içeriğini otomatik olarak tanımlamayı öğrenen insan görsel sistemiyle yakından ilgili çok modelli bir sinir ağı yöntemi sunmuşlardır. Önerilen model, iki aşamadan oluşmaktadır. Bunlar sırasıyla, nesnelerin bilgilerini ve bunların uzaysal ilişkilerini görüntülerden çıkaran bir nesne algılama ve yerelleştirme modeli ile uzun kısa süreli bellek birimlerine dayanan ve cümle üretimi için dikkat mekanizmasına sahip derin bir tekrarlayan sinir ağıdır. Açıklamadaki her bir sözcük, üretildiğinde otomatik olarak girdi görüntüsünün farklı nesnelerine hizalanarak insan görsel sisteminin dikkat mekanizmasına benzer bir sistem oluşturulmuştur. Aneja vd. [16] evrişimli sinir ağlarını kullanarak görüntü altyazısı oluşturmayı amaçlamışlardır. Çalışmada, evrişimli bir resim altyazı ekleme modeli geliştirerek modelin etkinliğini MSCOCO veri kümesinde kanıtlamışlardır. Önerilen modelin performansı, uzun kısa dönem hafıza ağlarıyla (LSTM) karşılaştırılmış ve LSTM modeline çok yakın bir performans elde etmişlerdir. Ayrıca, evrişimli dil üretme yaklaşımları lehine ayrıntılı bir analiz gerçekleştirmişlerdir. Redmon vd. [17] nesne algılama görevi için YOLO modelinin etkinliğini gösteren bir yöntem önermişlerdir. Nesne algılama ile ilgili önceki çalışmalar, nesne tespiti için sınıflandırıcıları kullanırken bu çalışmada, uzamsal olarak ayrılmış sınırlayıcı kutulara ve ilişkili sınıf olasılıklarını bir regresyon problemi olarak çözüm üretmişlerdir. Kullanılan sinir ağı ile doğrudan görüntülerden sınırlayıcı kutuları ve sınıf olasılıklarını tahmin etmektedirler. Algılama ardışık düzeninin tamamı, tek bir ağ olduğundan doğrudan algılama performansına göre uçtan uca optimize edilebildiğini vurgulamışlardır. Kullanılan birleşik YOLO mimarisi ile görüntülerin gerçek zamanlı olarak saniyede 45 kare ile işlendiğini ve son derece hızlı olduğunu vurgulamışlardır. Ağın daha küçük bir versiyonu olan Fast YOLO ile gerçekleştirilen deneylerde diğer gerçek zamanlı dedektörlerin iki katı haritalama elde ederken saniyede 155 kare işlediklerini, YOLO ile gerçekleştirilen deneylerin sonuçlarını karşılaştırdıklarında R-CNN gibi modellere kıyasla daha üstün performans sergilediğini bildirmişlerdir.

Chun vd. [18] köprü hasarının kapsamlı açıklamalarını otomatik olarak oluşturmak için derin öğrenmeye dayalı bir resim alt yazısı yöntemi önermişlerdir. Çalışmada, derin öğrenme modeline bir dikkat mekanizması getirerek oldukça tanımlayıcı cümleler üretilebileceği

gösterilmiştir. Ek olarak, köprülerin görüntülerinde çoğu zaman birden fazla hasar biçimi gözlemlenebildiğinden önerilen yöntem, karmaşık görüntülerin kapsamlı bir yorumunu sağlamak için birden fazla cümle çıkaracak şekilde uyarlanmıştır. Veri kümesinde, iki dilli değerlendirme eğitimi puanları (BLEU-1'den BLEU-4'e) sırasıyla 0,782, 0,749, 0,711, 0,693 olarak sunulmuştur ve açıklayıcı cümlelerin doğru şekilde çıkma yüzdesi % 69,3 olarak bildirilmiştir. Sonuçlar değerlendirildiğinde dikkat mekanizması olmayan bir modele kıyasla daha iyi bir performans sergilediğini belirtmişlerdir. Geliştirilen yöntem ile görüntülerde köprü hasarlarının kullanıcı dostu, metin tabanlı açıklamalarını sunmayı mümkün kılarak, nispeten az deneyime sahip mühendislerin ve hatta kapsamlı teknik uzmanlığa sahip olmayan idari personelin köprü hasar görüntülerini anlamasına olanak tanındığı belirtilmiştir. Wang vd. [19] derin öğrenme ile nesne algılama ve görüntü altyazısını entegre ederek yapılarda semantik bilgi çıkarımı için vizyon tabanlı bir yöntem önermişlerdir. Bu çalışmada, inşaat görüntülerinden veya videolarından göze çarpan bilgileri keşfetmeyi amaçlayan derin öğrenme nesne algılama ve görüntü altyazısını entegre ederek yeni bir anlamsal bilgi çıkarma yöntemi önerilmektedir. Önerilen yöntemde, inşaat nesne bölgelerinin özellik haritalarını ve bütünsel görüntüyü çıkarmak için bir kodlayıcı olarak nesne algılama kullanılmıştır. Anlamsal bilgileri çıkarmak için kod çözücü olarak görüntü alt yazısı seçilmiştir. Daha iyi erişilebilirlik ve görselleştirme için anlamsal bilgileri bir grafik formatına ayrıştırmak amacıyla son bir işleme yöntemi önerilmiştir. Deneylerde, önerilen yöntem, 1,84'lük Mutabakata Dayalı Görüntü Açıklama Değerlendirmesi (CIDEr) elde edildiği bildirilmiştir.

Al-Malla vd. [20] insan görüntüsünün anlaşılmasını taklit etmek için dikkat ve nesne özelliklerini kullanan görüntü altyazı modeli sunmuşlardır. Bu çalışmada, MSCOCO veri kümesi üzerinde önceden eğitilmiş YOLOv4 modelinden çıkarılan nesne özellikleriyle birlikte, ImageNet (Xception) üzerinde önceden eğitilmiş bir CNN modelinden çıkarılan evrişimli özellikleri kullanan, dikkat tabanlı, Kodlayıcı-Kod Çözücü derin bir mimari sunulmuştur. Aynı zamanda, nesne özellikleri için yeni bir konumsal kodlama şeması olan önem faktörü tanıtılmaktadır. Önerilen model, MS COCO ve Flickr30k veri kümelerinde test edilmiş ve benzer çalışmalardaki performansları ile karşılaştırılmıştır. Deney sonuçları, önerilen yöntemin CIDEr puanını %15,04 arttırdığı gözlenmiştir. Bhalekar ve Bedekar [21] görme engelli bireyler için derin öğrenmeyi kullanarak metin çıkarma ile resim altyazıları oluşturmak için bir model önermişlerdir. Bu çalışmada, ayrıntılı altyazılar oluşturan ve varsa bir görüntüden metin çıkaran ve görüntünün daha kesin bir tanımını sağlamak için bunu altyazının bir parçası olarak kullanan bir görüntü altyazı sistemi önerilmektedir. Görüntü özelliklerini çıkarmak için, önerilen model Evrişimli Sinir Ağlarını (CNN'ler) ve ardından öğrenilen görüntü özelliklerine dayalı olarak karşılık gelen cümleler üreten Uzun Kısa Süreli Bellek (LSTM) kullanılır. Ayrıca, metin çıkarma modülü kullanılarak, çıkarılan metin (varsa) görüntü açıklamasına dahil edilmekte ve altyazılar sesli biçimde sunulmaktadır. MS COCO, Flickr-8k, Flickr-30k veri kümeleri görüntü alt yazısı oluşturmak için kullanılmıştır. Yazarlar, deney sonuçlarının önerilen modelin görüntü alt yazı modellerinde mevcut modeller kadar etkili olduğunu ve metin çıkarımı gerçekleştirerek görüntü hakkında daha fazla içgörü sağladığını bildirmişlerdir.

Bizim çalışmamız, diğer çalışmalardan farklı olarak 2 farklı yöntemin bütünleştirilerek kullanılmasına dayanır. Literatürdeki yöntemlerin birçoğu hem görüntü altyazısı oluşturma hem de nesne algılama görevleri için çoğunlukla derin öğrenme tabanlı metodolojiler kullanırken çalışmamızda nesne algılama görevi için YOLO platformu kullanılmış ve görüntü altyazısı oluşturmak için bu görüntüler uzun kısa dönem hafıza ağlarına beslenmiştir. Ayrıca, önerilen modelin etkinliğini göstermek amacıyla farklı enjeksiyon teknikleri kullanılmıştır. Çalışmanın kapsamı dahilinde görme engelli bireylere yardımcı olması amacıyla için elde edilen alt yazılar sesli bir aktivite tanıma sistemine dönüştürülmüştür.

3. Önerilen Çerçeve (Proposed Framework)

Çalışmamızın bu bölümünde, görüntü altyazısı ve nesne algılama görevlerinde kullanılan modellerden, özellik enjeksiyon yöntemlerinden ve önerilen modelin mimarisinden bahsedilmiştir.

3.1. Görüntü Altyazısı (Image Captioning)

Altyazı oluşturma, belirli bir görüntü için metinsel bir açıklamanın oluşturulması gereken zorlu bir yapay zekâ görevidir. Görüntünün içeriğini anlamak için bilgisayar görüsünden her iki yöntemi ve görüntünün anlaşılmasını doğru sırayla kelimelere dönüştürmek için doğal dil işleme alanından bir dil modeli gerekir. Son zamanlarda, derin öğrenme yöntemleri, bu problemin örnekleri üzerinde son teknoloji sonuçlara ulaşmıştır. Bu yöntemlerle, karmaşık veri hazırlığı veya özel olarak tasarlanmış modellerden oluşan bir ardışık düzen gerektirmek yerine, tek bir uçtan uca model ile verilen görüntünün altyazısının tahminlemesi yapılabilmektedir [22-25].

3.2. Nesne Tespiti (Object Detection)

Nesne algılama, belirli bir görüntüdeki bir veya daha fazla nesnenin varlığını, konumunu ve türünü tanımlamayı içeren bir görevdir. Son yıllarda, derin öğrenme teknikleri, standart karşılaştırma veri kümelerinde ve bilgisayarla görme yarışmalarında olduğu gibi nesne tespiti için kayda değer sonuçlar sunmaktadır. "Sadece Bir Kez Bakarsınız" veya diğer adıyla You Only Look Once (YOLO), gerçek zamanlı olarak nesne algılamayı gerçekleştirebilen tek bir uçtan-uca modelle kayda değer sonuçlar sunan evrişimli sinir ağları ailesinden bir modeldir. "Sadece Bir Kez Bakarsınız" veya YOLO modeli, Redmon vd. [17] tarafından geliştirilen, hızlı nesne tespiti için tasarlanmış bir dizi uçtan uca derin öğrenme modelidir. Orijinal olarak GoogLeNet' in bir sürümüyle daha sonra güncellenen ve Visual Geometry Group (VGG) yöntemine dayalı DarkNet olarak adlandırılmaktadır. Yaklaşım, girişi bir hücre ızgarasına bölen ve her hücreyi doğrudan bir sınırlayıcı kutu ve nesne sınıflandırmasını tahmin eden tek bir derin evrişimli sinir ağını içerir. Sonuç, işlem sonrası bir adımla nihai bir tahminde birleştirilen çok sayıda aday sınırlama kutusuyla elde edilir. YOLO mimarisinin ilk sürümü genel mimariyi önerirken, ikinci sürümünde tasarımı iyileştirilmiş ve sınırlayıcı kutu önerisini iyileştirmek için önceden tanımlanmış bağlantı kutuları kullanılmıştır. Sonraki sürümlerinde ise model mimarisi ve eğitim süreci iyileştirilmiştir.

3.3. Özellik Enjeksiyon Teknikleri (Feature Injection Techniques)

Görüntü alt yazısı görevi için literatürde sıkça kullanılan init-inject, pre-inject, par-inject ve merge olmak üzere dört farklı enjeksiyon yöntemi bulunmaktadır [2, 26-29]. İnit-inject için, görüntü özelliklerini içeren vektör, önerilen dekoderin başlangıç durumuna beslenirken pre-inject, kelimelerden önce dekodere girdi olarak bu vektörü kullanmaktadır. Par-inject yönteminde ise vektör, bir kelime yerleştirme katmanından elde edilen çıktıyla birleştirilip giriş olarak dekodere verilmektedir. Merge yöntemi ise, bu vektöre dekoderin çıktı katmanından önce dekoderden gelen dile ait özellikler eklenerek gerçeklenir.

3.4. Uzun Kısa Dönem Hafıza Ağları (Long Short Term Memory Networks)

Derin öğrenme yaklaşımı insan beyninin sinir yapısını taklit etmesi sebebiyle daha çok yapay sinir ağlarının gelişimi sonrası popüler

olmuştur. Derin öğrenme modelleri ile makine öğrenmesi yöntemlerinin aksine özellik çıkarım aşaması, insan müdahalesi gerektirmeden gerçekleşmektedir. Bu nedenle, derin öğrenme modelleri büyük miktarda veriyi işleyebilen, kendi kendine öğrenebilen modeller olarak bilinir. Uzun kısa dönem hafıza ağları (LSTM) [30], metinlerdeki uzun süreli bağımlılıkları yorumlayabilen yinelemeli sinir ağı modelinin eksikliklerini gidermek üzerine geliştirilmiştir. Geleneksel ileri beslemeli sinir ağlarından farklı olarak geri bildirim bağlantıları da bulunmaktadır. LSTM, yinelemeli sinir ağları modellerine benzer bir mantıkla yürütülüyor olsa da bellek adı verilen yapılar, kapı olarak isimlendirilen birimlerle bilgi akışı üzerinde müdahale yeteneğine sahiptir. Böylece, sinir ağının hangi durumda olduğu bilgisini hafızasında saklayabilen hücrelerle yinelenen sinir ağlarının performansını iyileştirebilmektedir. Hafıza birimleri bir ya da birden fazla hafıza bloklarından oluşan ve bu blokların paylaştığı çarpımsal ve toplamsal kapılardan oluşmaktadır. Her LSTM yapısı, bilginin hangi kısımlarının unutulacağını veya hatırlanacağını, bir sonraki aşamaya geçilip geçilmeyeceği bilgisini içeren unutma, giriş ve çıkış kapılarından meydana gelmektedir. Uzun kısa dönem hafıza ağlarının metin sınıflandırma [31-32], hastalık tespiti [33], zaman serisi analizi [34, 35], görüntü tanıma [36-37], finansal ürünlerin fiyat tahmini [38], satış tahmini [39] gibi alanlarda uygulamaları bulunmaktadır. Bu çalışmada, uzun kısa dönem hafıza ağları (LSTM) modeli görüntü altyazısı oluşturma amacıyla kullanılmıştır.

3.5. Metni Sese Dönüştürme (Text to Speech)

Metin okuma teknolojisi (TTS), dijital metni bilgisayarlarda, akıllı telefonlarda ve tabletlerde bulunan kelimelerin yüksek sesle okunmasına dayanır. Metin okuma, genellikle okuma problemiyle mücadele eden çocuklara yardımcı olmasıyla bilinse de çalışmamızın da odak noktasını oluşturan görme engelli bireylerin hayatını kolaylaştıran çalışmalar için de tercih edilen bir yöntemdir. Günümüzde, neredeyse her dijital cihaz için kullanılabilen ve "yüksek sesle okuma" teknolojisi olarak bilinen metin okuma araçları bulunmaktadır. Bu çalışmada, görüntü altyazısı, Google'ın sağladığı Google Metin Okuma (gTTS) Python kütüphanesi vasıtasıyla ses dosyalarına dönüştürülmektedir.

3.6. Önerilen Çerçeve (Proposed Framework)

Önerilen çerçevenin inşasında iki farklı modelin eğitim işlemi için farklı veri kümeleri kullanılmaktadır. Görüntü altyazısı oluşturmak için ilk aşamada, gerçekçi ve nispeten küçük olması sebebiyle Flickr8K [40] veri kümesi tercih edilmiştir. Veri kümesi, her biri göze çarpan varlıkların ve olayların net açıklamalarını sağlayan beş farklı başlık ile eşleştirilmiş 8,000 görüntü içermektedir. Veri kümesindeki görüntüler, altı farklı Flickr grubundan seçilmiştir ve herhangi bir tanınmış kişi veya nesnenin konumu içermeme eğilimindedir. Çeşitli sahneleri ve durumları tasvir etmek için manuel olarak seçilmiştir. Veri kümesinde önceden tanımlanmış bir eğitim veri kümesi 6.000 görüntü, geliştirme veri kümesi 1.000 görüntü ve test veri kümesi 1.000 görüntü içermektedir. Özet olarak, Flickr8k ve Flickr30k [41], belirli olay ve faaliyetlerde yer alan insanlara ve nesnelere odaklanan beş referans başlığıyla 8000 ve 31,783 resim içerir. MSCOCO [42], her biri en az beş referans başlığıyla açıklamalı 118,287 eğitim ve 5,000 doğrulama görüntüsü içeren nispeten büyük bir veri kümesidir.

Görüntülerin içeriğini yorumlamak için önceden eğitilmiş bir modelden faydalanılmaktadır. Bu çalışmada, 2014 yılında ImageNet yarışmasını kazanan Oxford Visual Geometry Group (VGG) modeli kullanılmıştır. Önceden eğitilmiş modeli kullanarak görüntü özellikleri hesaplanarak veri kümesindeki belirli bir görüntünün yorumu olarak modele aktarılmıştır. Veri kümesi, her bir görüntü için birden çok açıklama içerdiğinden bu açıklamalar ön işleme aşamasına tabi tutulmuştur. Veri ön işleme adımında, tüm kelimeler küçük harfe dönüştürülmüş, tüm noktalama işaretlerini kaldırılmış, bir karakter veya daha kısa olan tüm kelimeler kaldırılmış, içinde sayı bulunan tüm kelimeler kaldırılmıştır. Ön işleme aşaması tamamlandıktan sonra, ideal olarak hem anlamlı hem de olabildiğince küçük bir kelime dağarcığı elde etmek adına kelime dağarcığının boyutu küçültülmüştür. Daha küçük bir kelime dağarcığı, daha hızlı çalışacak daha küçük bir modelle sonuçlanacaktır. Görüntü tanımlayıcıları ve açıklamaları sözlüğünü, her satırda bir görüntü tanımlayıcı ve açıklama olacak şekilde yeni bir dosyaya kaydedilmiş ve Tablo 1'de de görüntü tanımlayıcıları ve açıklamaları örnek olarak gösterilmiştir.

Belirli bir tanımlayıcı kümesi için temizlenmiş metin açıklamalarını yükleyen ve tanımlayıcılardan oluşan bir sözlüğü metin açıklamaları listelerine döndüren işlev uygulanmıştır. Geliştirilen model, bir görüntüye verilen bir altyazı oluşturmakta ve altyazı içeriğinde her seferinde bir kelime oluşturmaktır. Önceden üretilen kelimelerin sırası girdi olarak sağlanmaktadır. Bu nedenle, üretim sürecini başlatmak için ilk kelimeye ve altyazının sonunu belirtmek için son kelimeye ihtiyaç vardır. Açıklama metninin, modele girdi olarak sunulabilmesi veya modelin tahminleriyle karşılaştırılabilmesi için önce sayısal olarak kodlanması gerekmektedir. Verileri kodlamanın ilk adımı, sözcüklerden benzersiz tamsayı değerlerine tutarlı bir eşleme oluşturmaktır. Keras kütüphanesi, bu eşlemeyi yüklenen açıklama verilerinden öğrenebilen Tokenizer sınıfı vasıtasıyla sunabilmektedir.

Model eğitimi şu şekilde gerçekleştirilir: Her açıklama, kelimelere bölünmüştür. Modele bir kelime ve görüntü verilerek bir sonraki kelime oluşturulmuştur. Daha sonra, açıklamanın ilk iki kelimesi, sonraki kelimeyi oluşturmak için görüntü ile önerilen modele girdi olarak sunulmuştur. Sonra, model açıklamalarını oluşturmak için kullanıldığında, oluşturulan sözcükler birleştirilmekte ve bir görüntü için bir açıklama oluşturmak üzere girdi olarak tekrar tekrar sağlanmaktadır. Görüntü özellikleri ve sayısal olarak ifade edilen metin özellikleri olmak üzere önerilen modelin iki giriş dizisi bulunmaktadır. Metin dizisinde, sonraki kelimeyi kodlayan model için ayrı bir çıktı bulunmaktadır. Girdi, metni tamsayılar olarak kodlayarak bir kelime yerleştirme katmanına beslemektedir. Model, sözcük dağarcığındaki tüm sözcükler üzerinde bir olasılık dağılımı olacak bir tahmin çıkarmaktadır. Bu nedenle, çıktılar, 1 değerine sahip olan gerçek kelime konumu hariç tüm kelime pozisyonlarında "0" değerleri ile idealleştirilmiş bir olasılık dağılımını temsil eden her kelimenin bir kodlanmış versiyonu olmaktadır. Görüntü özellikleri ise doğrudan modelin başka bir bölümüne beslenmektedir. Çalışma kapsamında önerilen bütünleşik modelin ilk aşamasında görüntü özellikleri çıkarılmaktadır. Bu, ImageNet veri kümesinde önceden

Tablo 1. Görüntü tanımlayıcı ve açıklamaları (Image identifier and descriptions)

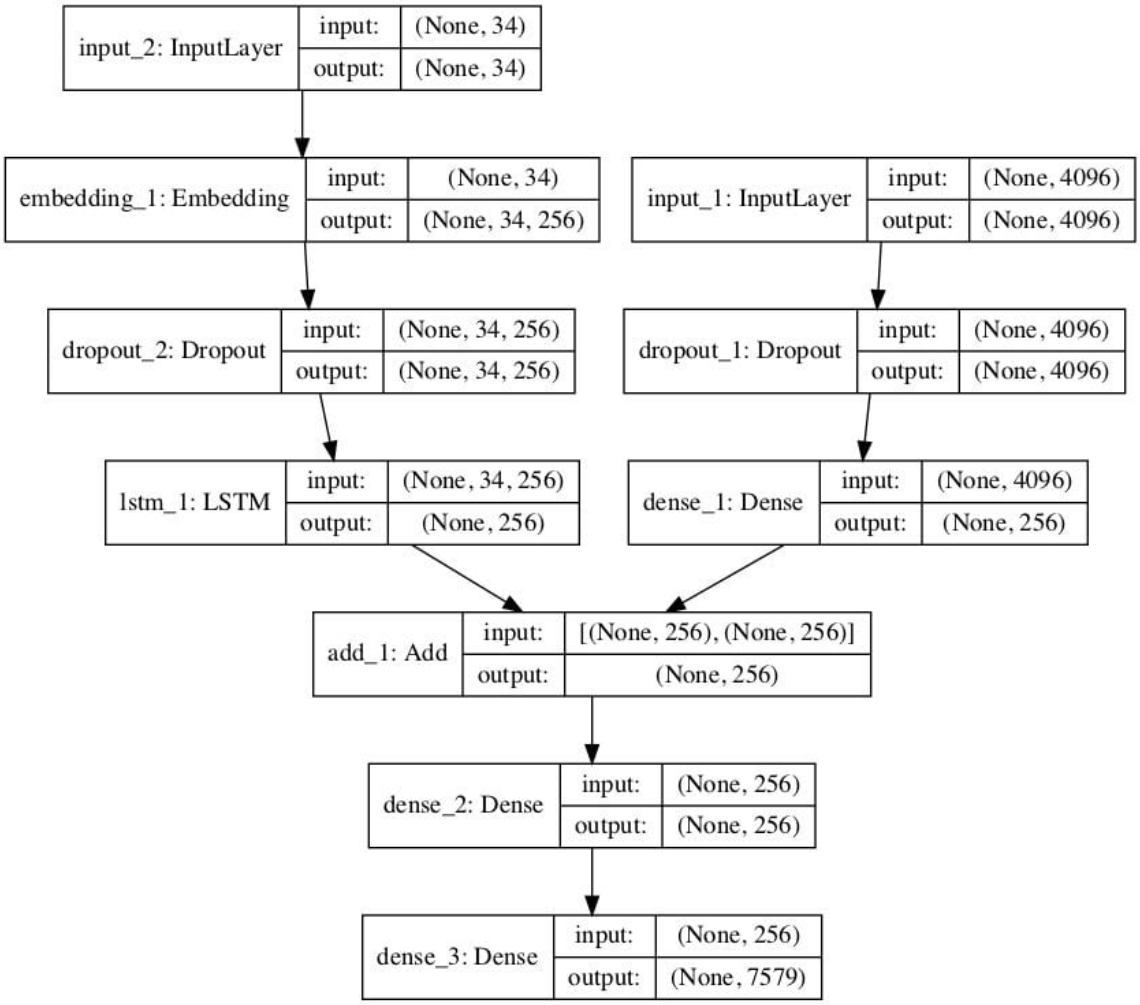
Görüntü Tanımlayıcısı	Görüntü Açıklaması
2252123185_487f21e336	bunch on people are seated in stadium
2252123185_487f21e336	crowded stadium is full of people watching an event
2252123185_487f21e336	crowd of people fill up packed stadium
2252123185_487f21e336	crowd sitting in an indoor stadium
2252123185_487f21e336	stadium full of people watch game

eğitilmiş 16 katmanlı bir VGG modelidir. Görüntüler, VGG modeliyle (çıktı katmanı olmadan) işlenerek çıkarılan özellik kümesi oluşturulmuştur. Sonrasında, metin girişini işlemek için bir kelime gömme katmanı ve ardından bir Uzun Kısa Süreli Bellek (LSTM) tekrarlayan sinir ağı katmanı kullanılmaktadır. Son olarak hem özellik çıkarıcı hem de sıra işlemcisi kod çözücü vasıtasıyla sabit uzunlukta bir vektör sunmaktadır. Elde edilen çıktılar birleştirilerek son bir tahminde bulunmak için Dense katmanı tarafından işlenmektedir. Görüntü özelliği çıkarma aşamasında, giriş görüntü özelliklerinin 4,096 öğeden oluşan bir vektör olması gerekmektedir. Bu öğeler, görüntünün 256 öğeli bir temsilini oluşturmak için yoğun bir katman tarafından işlenir. Sıralı işlemci modeliyle, elde edilen değerleri yok saymak için bir maske kullanan kelime yerleştirme katmanına beslenen, önceden tanımlanmış bir uzunluğa (34 kelime) sahip girdi dizileri verilmektedir. Bunu, 256 bellek birimine sahip bir LSTM katmanı izler. Her iki girdi modeli de 256 elemanlı bir vektör üretir. Ayrıca, her iki girdi modeli de %50 bırakma şeklinde düzenlileştirmeyi kullanır. Bu, model yapılandırması çok hızlı öğrendiğinden, eğitim veri kümesinin aşırı uyumunu azaltmak içindir. Dekoder modeli, bir toplama işlemi kullanarak her iki giriş modelindeki vektörleri birleştirir. Buradan elde edilen vektör ise, yoğun 256 nöron katmanına ve ardından dizideki bir sonraki kelime için tüm çıktı kelime dağarcığı üzerinde bir softmax tahmini yapan son çıktı yoğun katmanına beslenir. Şekil 1'de yukarıda detayları verilen altyazı oluşturma derin öğrenme modelinin yapısı sunulmuştur.

4. Deney Sonuçları (Experiment Results)

Çalışmamızın bu bölümünde önerilen modelin, görme engelli bireyler için nesne tanıma ve altyazı oluşturma görevlerindeki performansı sunulmaktadır. İlk aşamada, gerçek ve tahmin edilen açıklamalar, oluşturulan metin içeriğinin beklenen metin içeriğine ne kadar yakın olduğunu değerlendiren iki dilli değerlendirme (Bilingual Evaluation Understudy-BLEU) puanı kullanılarak belirlenmektedir. BLEU puanları, çevrilmiş metni bir veya daha fazla referans çevirisine karşı değerlendirmek için metin çevirisinde kullanılır. Bu çalışmada, oluşturulan her bir açıklama görüntünün tüm referans tanımlarıyla karşılaştırılmaktadır. Daha sonra 1, 2, 3 ve 4 kümülatif n-gram için BLEU puanları hesaplanmaktadır. 1'e yakın hesaplanan BLEU puan değeri modelin performansında başarılı olarak değerlendirilirken sıfıra yakın değer, model başarısının yeterli olmadığı anlamına gelmektedir. Önerilen modelin Flickr8K, Flickr30K ve MSCOCO veri kümelerindeki performansı, BLEU-1, BLEU-2, BLEU-3, BLEU-4 değerlendirme metrikleriyle sırasıyla Tablo 2, Tablo 3 ve Tablo 4' de sunulmuştur. Koyu fontta gösterilen BLEU değerleri ilgili veri kümeleri için en iyi sonuçların elde edildiği anlamına gelmektedir.

Tablo 2'de Flickr8K veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri sunulmaktadır. Tablo 2'den de açıkça görüldüğü üzere par-inject tekniği, Flickr8K veri kümesinde 0,6345 BLEU-1 değeri ile en yüksek başarıma sahiptir. Par-inject sonuçları incelendiğinde BLEU-1 sonucunu sırasıyla 0,4633 ile



Şekil 1. Önerilen modelin mimarisi (Architecture of the proposed model)

BLEU-2, 0,3128 ile BLEU-3 ve 0,1945 ile BLEU-4 takip etmektedir. Özellik enjeksiyon teknikleri BLEU-1 skorlarına göre kıyaslandığında en iyi performans, 0,6345 ile par-inject tekniği vasıtasıyla elde edilirken onu, 0,6239 ile init-inject, 0,5469 ile pre-inject, 0,5371 ile merge izlemektedir.

Tablo 3'de Flickr30K veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri sunulmaktadır. Tablo 3'den de açıkça görüldüğü üzere par-inject tekniği, Flickr30K veri kümesinde 0,6546 BLEU-1 skoru ile en yüksek performansı sergilemiştir. Par-inject sonuçları incelendiğinde Flickr8K veri kümesinde elde edilen sonuçlara benzer şekilde performans göstermektedir. BLEU değerleri sırasıyla 0,5125 ile BLEU-2, 0,3614 ile BLEU-3 ve 0,2417 ile BLEU-4 şeklindedir. Özellik enjeksiyon teknikleri BLEU-1 skorlarına göre kıyaslandığında en iyi performans 0,6546 ile par-inject tekniği vasıtasıyla elde edilirken onu, 0,6370 ile init-inject, 0,6036 ile pre-inject, 0,5922 ile merge yöntemi takip etmektedir.

Tablo 4' de MSCOCO veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri gösterilmiştir. Tablo 4' den de görüldüğü üzere init-inject tekniği, MSCOCO veri kümesinde 0,6807 BLEU-1 skoru ile en yüksek başarıma sahiptir. İnit-inject sonuçları incelendiğinde BLEU skorları Flickr8K ve Flickr30K veri kümelerinden elde edilen sonuçlara benzer şekilde performans gözlenmektedir. İnit-inject teknikleri BLEU skorlarına göre sırasıyla 0,5103 ile BLEU-2, 0,3766 ile BLEU-3 ve 0,2446 ile BLEU-4 şeklinde elde edilmektedir. Özellik enjeksiyon teknikleri BLEU-1 skorlarına göre kıyaslandığında en iyi performans 0,6807 ile init-inject tekniği ile elde edilirken onu, 0,6584 ile pre-inject, 0,6433 ile par-inject, 0,6330 ile merge yöntemi izlemektedir. Tablo 2, Tablo 3 ve Tablo 4 sonuçları genel olarak değerlendirildiğinde merge enjeksiyon tekniği tüm veri kümelerinde başarımı en zayıf yöntemdir. Öte yandan, par-inject modeli Flickr8K ve Flickr30K veri kümelerinde en iyi performansı sergilerken MSCOCO veri kümesinde init-inject yöntemi başarımı en iyi modeldir.

Tablo 5, Tablo 6 ve Tablo 7'de, sırasıyla Flickr8K, Flickr30K ve MSCOCO veri kümeleri üzerinde önerdiğimiz model ve literatür çalışmalarının farklı özellik enjeksiyon teknikleri kullanılarak BLEU-1 değerleri açısından karşılaştırması sunulmuştur. Çalışmamızda önerilen model, Tablo 5' ten de görüldüğü üzere 0,6345 BLEU-1

Tablo 2. Flickr8K veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri
(BLEU values obtained with different feature injection results on Flickr8K dataset)

Enjeksiyon Teknikleri	BLEU-4	BLEU-3	BLEU-2	BLEU-1
par (eşit)	0,1945	0,3128	0,4633	0,6345
init (başta)	0,1912	0,3012	0,4774	0,6239
pre (önceden)	0,1715	0,2905	0,4628	0,5469
merge (birleştirme)	0,1880	0,2911	0,4556	0,5371

Tablo 3. Flickr30K veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri
(BLEU values obtained with different feature injection results on Flickr30K dataset)

Enjeksiyon Teknikleri	BLEU-4	BLEU-3	BLEU-2	BLEU-1
par (eşit)	0,2417	0,3614	0,5125	0,6546
init (başta)	0,2355	0,3502	0,5182	0,6370
pre (önceden)	0,2201	0,3377	0,4916	0,6036
merge (birleştirme)	0,2231	0,3356	0,4831	0,5922

Tablo 4. MSCOCO veri kümesi üzerinde farklı özellik enjeksiyon teknikleriyle elde edilen BLEU değerleri
(BLEU values obtained with different feature injection results on MSCOCO dataset)

Enjeksiyon Teknikleri	BLEU-4	BLEU-3	BLEU-2	BLEU-1
par (eşit)	0,2362	0,3648	0,5190	0,6433
init (başta)	0,2446	0,3766	0,5103	0,6807
pre (önceden)	0,2375	0,3578	0,5279	0,6584
merge (birleştirme)	0,2208	0,3427	0,4918	0,6330

Tablo 5. Flickr8K veri kümesi üzerinde önerilen model ve literatür çalışmalarının farklı özellik enjeksiyon teknikleri kullanılarak BLEU-1 değerleri açısından kıyaslanması
(Comparison of the proposed model and literature studies on the Flickr8K dataset in terms of BLEU-1 values using different feature injection techniques)

Kod Çözücü	par (eşit)	init (başta)	pre (önceden)	merge (birleştirme)
RNN [43]	0,611	0,611	0,609	0,600
LSTM [44]	-	-	-	0,500
CNN+LSTM [45]	0,6671	0,6579	0,6406	-
LSTM [46]	-	-	-	0,4537
Önerilen Model	0,6345	0,6239	0,5469	0,5371

skoruyla par-inject yöntemi kullanan çalışmalarla kıyaslandığında oldukça rekabetçidir. Yine aynı veri kümesi üzerinde merge tekniği ile 0,5371 BLEU-1 skoru elde edilmiş olup genellikle literatür çalışmalarından üstün performans sergilemektedir. Tablo 6'da gözlendiği üzere, Flickr30K veri kümesinde BLEU-1 skorlarında dikkate değer bir artış gözlenmekte olup 0,6546 ile en iyi başarım literatür çalışmalarına benzer şekilde par-inject tekniğinde sağlanmıştır. BLEU-1 skorları literatür çalışmalarıyla kıyaslandığında çoğunlukla daha iyi performans sergilediği gözlenmektedir. Tablo 5'te Suresh vd. [45] yaptığı çalışmada elde edilen BLEU-1 sonuçlarının yaklaşık %3 daha yüksek performans göstermesinin sebebi, CNN ve LSTM modellerinin kod çözücü olarak birleştirilmesinden kaynaklanmaktadır. Tablo 6' da Nugraha vd. [47] yapmış olduğu çalışmada par- inject modelinin yaklaşık %3 daha yüksek performans göstermesinin sebebi ise farklı konuşma dili kullanımından kaynaklanmaktadır. Tablo 7'den görüldüğü üzere, MSCOCO veri kümesinde init-inject tekniğiyle 0,6807 BLEU-1 skoru elde edilmiştir. Tablo 7'de, literatür çalışmalarıyla kıyaslandığında önerilen modelin init-inject tekniği kullanan çalışmalarda en yüksek performansı sergilediği gözlenmektedir. Yine aynı veri kümesinde merge tekniği ile elde edilen BLEU-1 skorlarına bakıldığında 0,6330 değeriyle ile önerilen model oldukça rekabetçidir. Şekil 2'de veri kümelerinden seçilen görüntülerin önerilen model ile nesne tanıma başarıları görüntüler üzerinde

Tablo 6. Flickr30K veri kümesi üzerinde önerilen model ve literatür çalışmalarının farklı özellik enjeksiyon teknikleri kullanılarak BLEU-1 değerleri açısından kıyaslanması
(Comparison of the proposed model and literature studies on the Flickr30K dataset in terms of BLEU-1 values using different feature injection techniques)

Kod Çözücü	par (eşit)	init (başta)	pre (önceden)	merge (birleştirme)
RNN [43]	0,613	0,613	0,614	0,605
LSTM [45]	0,686	-	-	-
GRU [47]	-	-	-	0,3670
Önerilen Model	0,6546	0,6370	0,6036	0,5922

Tablo 7. MSCOCO veri kümesi üzerinde önerilen model ve literatür çalışmalarının farklı özellik enjeksiyon teknikleri kullanılarak BLEU-1 değerleri açısından kıyaslanması (Comparison of the proposed model and literature studies on the Flickr30K dataset in terms of BLEU-1 values using different feature injection techniques)

Kod Çözücü	par (eşit)	init (başta)	pre (önceden)	merge (birleştirme)
RNN [43]	0,6670	0,6790	0,6770	0,6770
3-layer GRU [48]	0,6200	0,6379	0,6169	0,5898
LSTM [49]	-	-	0,5220	-
Önerilen Model	0,6433	0,6807	0,6584	0,6330

(a) (b)

(c) (d)

Şekil 2. Veri kümelerinden seçilen örnek görüntüler a) Görüntü-1, b) Görüntü-2, c) Görüntü-3 d) Görüntü-4
(Sample images selected from datasets a) Image-1, b) Image-2, c) Image-3 d) Image-4)

Tablo 8. Veri kümelerinden seçilen görüntülerinden oluşturulmuş altyazı örnekleri
(Examples of captions created from images selected from datasets)

Görüntü	Görüntünün Yer Tanımı Referansları	Görüntünün Tespit Edilen Tanımı	Oluşturulan Ses Dosyası
Görüntü-1	• A man at the top of a mountain with a beautiful view in the background • A man is sitting on a snowbank • A man on a snowy peak • A person in red snow gear is kneeling on a snowy ridge under a blue sky • Mountain climber in a red suit poses on snowy peak with mountains in the background	Man is standing on the top of mountain	başarılı-örnek1.mp3
Görüntü-2	• A climber wearing a blue helmet and headlamp is attached to a rope on the rock face • A man climbs a rocky wall • A rock climber climbs a large rock • A woman in purple snakeskin pants climbs a rock Person with blue helmet and purple pants is rock climbing	Man is climbing up mountain	başarılı-örnek2.mp3
Görüntü-3	A dog with mud stuck on his underside is running on grass • A golden retriever has on a collar and a harness and has a muddy lower body • A muddy dog prances through the grass • A muddy golden dog running in grass • A white dog that is muddy walks across the green grass	Dog is running through the grass	başarılı-örnek3.mp3
Görüntü-4	• A brown dog is running over snow near leafless trees • A brown dog runs across a snowy field • A dog is running through the snow • A dog runs through the snow • A medium sized brown dog is running through an open white wooded area	Dog is running through the snow	başarılı-örnek4.mp3

gösterilmiştir. Tablo 8'de ise, Şekil 2'deki görüntülerin görüntü referans tanımları, tespit edilen tanımları ve ses dosyaları sunulmuştur.

5. Sonuçlar (Conclusions)

Çalışma kapsamında, bilgisayarla görü ve makine çevirisindeki son gelişmeleri harmanlayan ve bir görüntüyü ifade cümleler oluşturmak için derin ve tekrarlayan mimariye dayalı üretken bir model sunulmuştur. Önerilen bu model ile görüntülerden elde edilen metinler, ses dosyası formatına dönüştürülebilmekte ve görme engelli bireyler için kişinin etrafında bulunan nesnelerin aktivitesi tanımlanabilmektedir. Bu amaçla, ilk olarak, belirli bir görüntüdeki bir veya daha fazla nesnenin varlığını, konumunu ve türünü tanımlayan YOLO modeliyle görüntüler üzerinde nesne tanıma işlemi gerçekleştirilmiştir. Sonrasında, uzun kısa dönem hafıza ağları (LSTM) eğitim görüntüsü verilen hedef açıklama cümlesinin olasılığını en üst düzeye çıkarmak için eğitilmiştir. Böylece, ilgili görüntü içerisinde yer alan aktiviteler, açıklama olarak metin biçimine dönüştürülmüştür. Metin biçimine dönüştürülen aktiviteler, Google metin okuma platformundan faydalanılarak aktiviteyi tanımlayan ses dosyaları elde edilmiştir. Önerilen modelin etkinliğini göstermek amacıyla dört farklı özellik enjeksiyon mimarisi değerlendirilerek Flickr8K, Flickr30K ve MSCOCO veri kümeleri kullanılmıştır. Deney sonuçları, Flickr8K ve Flickr30K veri kümelerinde par-inject özellik enjeksiyon yönteminin kullanımının literatür çalışmalarıyla kıyaslandığında oldukça rekabetçi olduğunu göstermiştir. MSCOCO veri kümesinde init-inject tekniğinin önerilen LSTM dekoderi ile harmanlandığında literatür çalışmalarına kıyasla üstün bir performans sergilediği gözlenmiştir.

Kaynaklar (References)

1. Hossain M.Z., Sohel F., Shiratuddin M.F., Laga, H., A comprehensive survey of deep learning for image captioning, ACM Computing Surveys 51 (6), 1-36, 2019.
2. Yao T., Pan Y., Li Y., Qiu Z., Mei T., Boosting image captioning with attributes, IEEE International Conference on Computer Vision, Venice, Italia, 4894-4902, 22-29 Ekim, 2017.
3. You Q., Jin H., Wang Z., Fang C., Luo J., Image captioning with semantic attention, IEEE Conference on Computer Vision And Pattern Recognition, Las Vegas, USA, 4651-4659, 26 Haziran-1 Temmuz, 2016.
4. Pan J.Y., Yang H.J., Duygulu P., Faloutsos C., Automatic image captioning, IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 1987-1990, 27-30 Haziran, 2004.
5. O'Shea K. ve Nash R. An introduction to convolutional neural networks. https://arxiv.org/abs/1511.08458. Aralık 2, 2015. Temmuz 30, 2019.
6. Medsker L.R. ve Jain L.C., Recurrent neural networks, Design and Applications, 5, 64-67, 2001.
7. Hochreiter S. ve Schmidhuber J., Long short-term memory, Neural Computation, 9 (8), 1735-1780, 1997.
8. Montavon G., Samek W., Müller K.R., Methods for interpreting and understanding deep neural networks, Digital Signal Processing, 73, 1-15, 2018.
9. Guo T., Dong J., Li H., Gao Y., Simple convolutional neural network on image classification. IEEE International Conference on Big Data Analysis, Beijing, China, 721-724, 10-12 Mart, 2017.
10. Ouyang X., Zhou P., Li C.H., Liu L., Sentiment analysis using convolutional neural network, IEEE International Conference on Computer and Information Technology, Dhaka, Bangladesh, 2359-2364, 21-23 Aralık, 2015.
11. Yang J., Nguyen M.N., San P.P., Li X.L., Krishnaswamy S., Deep convolutional neural networks on multichannel time series for human activity recognition. International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina, 3995-4001, 25-31 Temmuz, 2015.
12. Salamon J. ve Bello J.P., Deep convolutional neural networks and data augmentation for environmental sound classification, IEEE Signal Processing Letters, 24 (3), 279-283, 2017.
13. Eyben F., Petridis S., Schuller B., Tzimiropoulos G., Zafeiriou S., Pantic M., Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks, IEEE International Conference on Acoustics, Speech and Signal Processing, Prague, Czech Republic, 5844-5847, 22-27 Mayıs, 2011.

14. Khataei Maragheh H., Gharehchopogh F.S., Majidzadeh K., Sangar A.B., A new hybrid based on long short-term memory network with spotted hyena optimization algorithm for multi-label text classification. Mathematics 10 (3), 1-24, 2022.

15. Yang Z., Zhang Y., Rehman S., Huang Y., Image captioning with object detection and localization, International Conference on Image and Graphics, Shanghai, China, 109-118, 13-15 Eylül, 2017.

16. Aneja J., Deshpande A., Schwing A.G., Convolutional image captioning, IEEE Conference on Computer Vision and Pattern Recognition, Utah, USA, 5561-5570, 18-22 Haziran, 2018.

17. Redmon J., Divvala S., Girshick R., Farhadi A., You only look once: Unified, real-time object detection, IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 779-788, 26 Haziran-1 Temmuz, 2016.

18. Chun P.J., Yamane T., Maemura Y., A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage. Computer-Aided Civil and Infrastructure Engineering, 37 (11), 1387-1401, 2022.

19. Wang Y., Xiao B., Bouferguene A., Al-Hussein M., Li H., Vision-based method for semantic information extraction in construction by integrating deep learning object detection and image captioning, Advanced Engineering Informatics, 53, 1-13, 2022.

20. Al-Malla M.A., Jafar A., Ghneim N., Image captioning model using attention and object features to mimic human image understanding, Journal of Big Data, 9 (1), 1-16, 2022.

21. Bhalekar M. ve Bedekar M., D-CNN: A New model for generating image captions with text extraction using deep learning for visually challenged individuals, Engineering, Technology & Applied Science Research 12 (2), 8366-8373, 2022.

22. Herdade S., Kappeler A., Boakye K., Soares J., Image captioning: Transforming objects into words. Neural International Conference on Neural Information Processing Systems, Vancouver, Canada, 11137-11147, 8-14 Aralık, 2019.

23. Feng Y., Ma L., Liu W., Luo J., Unsupervised image captioning. IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, USA, 4125-4134, 15-20 Haziran, 2019.

24. Huang L., Wang W., Chen J., Wei X.Y., Attention on attention for image captioning. IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 4634-4643, 27 Ekim-2 Kasım, 2019.

25. Staniūtė R. ve Šešok D., A systematic literature review on image captioning, Applied Sciences 9(10), 1-20, 2019.

26. Devlin J., Cheng H., Fang H., Gupta S., Deng L., He X., Mitchell M., Language models for image captioning: The quirks and what works. https://arxiv.org/abs/1505.01809. Mayıs 7, 2015.

27. Nina O. ve Rodriguez A., Simplified LSTM unit and search space probability exploration for image description, IEEE International Conference on Information, Communications and Signal Processing, Singapore, 1-5, 2-4 Aralık, 2015.

28. Liu S., Zhu Z., Ye N., Guadarrama S., Murphy K., Improved image captioning via policy gradient optimization of spider, IEEE International Conference on Computer Vision, Venice, Italia, 873-881, 27-29 Ekim, 2017.

29. Mao J., Wei X., Yang Y., Wang J., Huang Z., Yuille, A.L., Learning like a child: Fast novel visual concept learning from sentence descriptions of images, IEEE International Conference on Computer Vision, Las Condes, Şili, 2533-2541, 11-18 Aralık, 2015.

30. Sak H., Senior A., Beaufays F., Long short-term memory recurrent neural network architectures for large scale acoustic modeling, Annual Conference of the International Speech Communication Association, Singapore, 338-342, 14-18 Eylül, 2014.

31. Gültekin I., Artuner H., Turkish dialect recognition in terms of prosodic by long short-term memory neural networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (1), 213-224, 2020.

32. Kilimci Z.H., Financial sentiment analysis with Deep Ensemble Models (DEMs) for stock market prediction, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (2), 635-650, 2020.

33. Altun S. ve Alkan A., LSTM-based deep learning application in brain tumor detection using MR spectroscopy, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (2), 1193-1202, 2022.

34. Gökdemir A., ve Çalhan A., Deep learning and machine learning based anomaly detection in internet of things environments, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (4), 1945-1956, 2022.

35. Utku A., Using network traffic analysis deep learning based Android malware detection, Journal of the Faculty of Engineering and Architecture of Gazi University, 37 (4), 1823-1838, 2022.

36. Akalın F., Yumuşak N., Classification of ALL, AML and MLL leukaemia types on microarray dataset using LSTM neural network approach, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (3), 1299-1306, 2023.

37. Dölek İ., Kurt A., Ottoman Optical Character Recognition with deep neural networks, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (4), 2579-2594, 2023.

38. Kantar O., Kilimci Z.H., Deep learning based hybrid gold index (XAU/USD) direction forecast model, Journal of the Faculty of Engineering and Architecture of Gazi University, 38 (2), 1117-1128, 2023.

39. Erol B., İnkaya, T., Long short-term memory network based deep transfer learning approach for sales forecasting, Journal of the Faculty of Engineering and Architecture of Gazi University, 39 (1), 191-202, 2024.

40. Hodosh M., Young P., Hockenmaier J., Framing image description as a ranking task: Data, models and evaluation metrics, Journal of Artificial Intelligence Research, 47, 853-899, 2013.

41. Plummer B.A., Wang L., Cervantes C.M., Caicedo J.C., Hockenmaier J., Lazebnik S., Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, IEEE International Conference on Computer Vision, Las Condes, Şili, 2641-2649, 2015.

42. Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollar P., Zitnick C.L., (2014). Microsoft coco: Common objects in context, European Conference on Computer Vision, Zurich, Switzerland, 740-755, 6-12 September, 2014.

43. Tanti M., Gatt A., Camilleri K.P., Where to put the image in an image caption generator, Natural Language Engineering, 24 (3), 467-489, 2018.

44. Mulyanto E., Setiawan E.I., Yuniarno E.M., Purnomo M.H., Automatic ındonesian ımage caption generation using CNN-LSTM model and FEEH-ID dataset, IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications, Tianjin, China, 1-5, 14-16 Haziran, 2019.

45. Suresh K.R., Jarapala A., Sudeep P.V., Image captioning encoder–decoder models using CNN-RNN architectures: A comparative study, Circuits, Systems, and Signal Processing, 41 (10), 5719-5742, 2022.

46. Martin A.D., Ahmadzade E., Moon I., Privacy-preserving image captioning with deep learning and double random phase encoding, Mathematics 10 (16), 1-14, 2022.

47. Nugraha A.A. ve Arifianto A., Generating image description on Indonesian language using convolutional neural network and gated recurrent unit, International Conference on Information and Communication Technology, Kuala Lumpur, Malaysia, 1-6, 24-26 Temmuz, 2019.

48. Keskin R., Çaylı Ö., Moral Ö.T., Kılıç V., Aytuğ O., A benchmark for feature-injection architectures in image captioning, Avrupa Bilim ve Teknoloji Dergisi, 31, 461-468, 2021.

49. You Q., Jin H., Luo J. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. https://arxiv.org/abs/1801.10121. Ocak 30, 2018.