Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 26, S. 191-196, Temmuz 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 26, pp. 191-196, July 2021

Research Article

Akıllı Telefonlar için Birleştirme Modeli Tabanlı Görüntü Altyazılama

Muharrem Baran*, Özge Taylan Moral, Volkan Kılıç

İzmir Katip Çelebi Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Elektrik-Elektronik Mühendisliği Bölümü, İzmir, Türkiye
(ORCID: 0000-0001-7394-3649, 0000-0003-0482-267X, 0000-0002-3164-1981), y190207002@ogr.ikcu.edu.tr, ozgetaylan.moral@ikcu.edu.tr, volkan.kilic@ikcu.edu.tr)

(3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications June 11-13, 2021)

(**DOI:** 10.31590/ejosat.950924)

ATIF/REFERENCE: Baran, M., Moral, Ö.T. & Kılıç, V. (2021). Akıllı Telefonlar için Birleştirme Modeli Tabanlı Görüntü Altyazılama. *Avrupa Bilim ve Teknoloji Dergisi*, (26), 191-196.

Öz

Görüntü altyazılama, bir görüntünün metinsel açıklamasını doğal dil işleme ve bilgisayarlı görü kullanılarak oluşturma işlemidir. Bir görüntünün görsel içeriğini makineye tanımlatmak, potansiyel uygulamaları nedeniyle son yıllarda artarak ilgi görmüştür. Bu çalışmada, akıllı telefonlarda uygulanabilir, kodlayıcı-kod çözücü yaklaşımına dayanan birleştirme modeli tabanlı bir görüntü altyazılama sistemi önerilmektedir. Önerilen birleştirme modelinde kodlayıcı olarak görüntü özniteliklerini çıkarmak için VGG16 evrişimsel sinir ağları ve kelime özelliklerini çıkarmak için uzun-kısa dönemli bellek yapısı kullanılmıştır. Bu iki işlem sonrası, görüntü özniteliklerinin ve oluşturulan kelime özelliklerinin kodlanmış biçimleri önerilen modelde birleştirilmiştir. Bu iki kodlanmış girdinin kombinasyonu daha sonra dizideki bir sonraki kelimeyi oluşturmak için çok basit bir kod çözücü modeli tarafından kullanılarak görüntülerin doğal dile uygun altyazıları başarıyla üretilmiştir. Önerilen sistem Flickr8k/30k veri kümeleri üzerinde BLEUn metriği kullanılarak test edilmiş ve literatürdeki çalışmalarla kıyaslanarak sağladığı üstünlük gösterilmiştir. Önerilen sistem, ayrıca, benzer çalışmalardan farklı olarak internet bağlantısı olmadan görüntü altyazısı üretebilecek şekilde geliştirdiğimiz *ImCap* adlı Android uygulamamız üzerinde de başarıyla çalıştırılmıştır. Bu uygulama ile görüntü altyazılamanın daha çok kullanıcıya ulaşması amaçlanmıştır.

Anahtar Kelimeler: Görüntü Altyazılama, Bilgisayarlı Görü, Doğal Dil işleme, Android.

Merge Model Based Image Captioning for Smartphones

Abstract

Image Captioning is the process of generating a textual description of an image by using both natural language processing and computer vision. Definition of the visual content of an image to the machine has attracted increasing attention in recent years due to its potential applications. In this study, an image captioning system based on an encoder-decoder merge model approach, applicable to smartphones, is proposed. In the proposed merge model, VGG16 convolutional neural networks are used to extract the image features and long-short term memory are used to extract the word features as encoder. After these two processes, the encoded forms of the images and the word features were merged in the proposed model. Image captioning was done successfully after the combination of these two encoded inputs had been used by a very simple decoder model to generate the next word in the sequence. The proposed system was tested using the BLEUn metric on the Flickr8k/30k dataset and its superiority was shown by comparing it with the studies in the literature. The proposed system was also integrated with our Android application called *ImCap*, which we have developed to generate captions without an internet connection, unlike other similar studies. With this application, image captioning is aimed to reach more users.

Keywords: Image Captioning, Computer Vision, Natural Language Processing, Android.

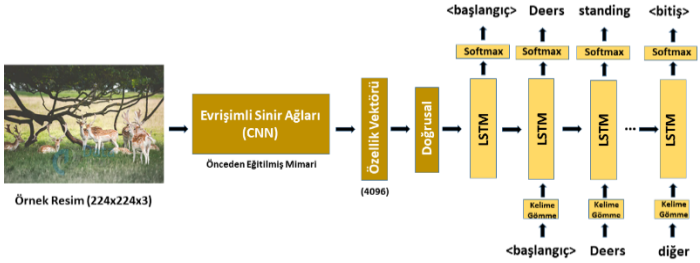
* Sorumlu Yazar: y190207002@ogr.ikcu.edu.tr

1. Giriş

Görüntülerden, bilgisayarlı görü ve doğal dil işleme yöntemleri kullanılarak, dil bilgisi açısından anlamlı ve okunabilir tanımlar üretilmesi, son yıllarda görsel arama, görüntü erişim ve dizinleme, görme engelli bireyler için sanal asistanlar gibi uygulamalar nedeniyle dikkat çekmektedir (Çaylı, Makav, Kılıç, & Onan, 2020; Makav & Kılıç, 2019a, 2019b).

Görüntü özellikleri ile görüntü altyazısı oluşturmak için geliştirilen önceki çalışmalar şablon tabanlı ve getirim tabanlı çalışmaları içermektedir (Elliott & Keller, 2013; Kuznetsova, Ordonez, Berg, & Choi, 2014; Mason & Charniak, 2014; Socher, Karpathy, Le, Manning, & Ng, 2014). Şablon tabanlı yöntemler ön-tanımlı cümle şablonlarından ve nesne algılayıcılardan yararlanarak altyazı üretirler. Getirim tabanlı yöntemler mevcut getirim kütüphanesindeki görüntülerden yararlanarak altyazı üretirler. Bu yöntemlerde altyazı kalitesi nesne algılayıcı ve kütüphane gibi etkenlere bağlı olduğundan, alternatif yöntemler önerilmiştir. Evrişimsel sinir ağları (Convolutional Neural Network – CNN) ve tekrarlayan sinir ağı (Recurrent Neural Network) yapılarını birleştiren kodlayıcı-kod çözücü yaklaşımı alternatif yöntemlerden birisidir (Mao et al., 2014). CNN, kodlayıcı olarak görüntü özniteliklerini çıkarmak için kullanılırken, RNN çıkarılan görüntü özniteliklerinin beslendiği kod çözücü olarak kullanılır. Son gelişmelerle birlikte, kodlayıcı tasarımında kullanılan VGGNet (Simonyan & Zisserman, 2014) ve Inception-v3 (Mathews, Xie, & He, 2018) gibi CNN mimarileri geliştirilmiştir. RNN tabanlı kod çözücüler, geleneksel RNN yapılarının neden olduğu kaybolan ve patlayan gradyan problemlerine çözüm olarak üretilen uzun-kısa dönemli bellek (Long Short Term Memory – LSTM) ve kapılı tekrarlayan hücre (Gated Recurrent Unit – GRU) yapılarını kullanarak, CNN ile çıkarılan görüntü özniteliklerini doğal dile uygun altyazıya dönüştürür (Hossain, Sohel, Shiratuddin, & Laga, 2019). LSTM yapısı, bilgi depolamak için uzun periyotlarda hafıza hücreleri kullanırken GRU yapısı ek hafıza hücresi kullanmadan bilgi akışı sağlar. Chen vd. çalışmalarında çift yönlü yinelenen sinir ağlarını (Bidirectional Recurrent Neural Network) kullanarak hem altyazılardan öznitelik çıkarmayı hem de özniteliklerden altyazı üreterek görüntü altyazılama yapmayı hedeflemiştir (Chen & Zitnick, 2014). Mao vd. çalışmalarında tek ve çok katmanlı tekrarlayan sinir ağlarını kullanarak altyazılama yapmıştır (Mao et al., 2014). Hodosh vd. çalışmalarında çekirdek kanonik korelasyon analizini (Kernel Canonical Correlation Analysis) uygulayarak altyazılamayı yapmıştır (Hodosh, Young, & Hockenmaier, 2013). Kiros vd. ise çalışmalarında uzun kısa dönemli hafıza (Long Short Term Memory) ağını kodlayıcı ve kod çözücü için kullanarak görüntü altyazılamayı hedeflemiştir (Kiros, Salakhutdinov, & Zemel, 2014). RNN tabanlı kod çözücülerde, kodlayıcıdan gelen görsel bilgi doğrudan RNN yapısına beslenebildiği gibi, RNN yapısından önce ek bir katmanla da beslenebilir (Tanti, Gatt, & Camilleri, 2018). Örneğin, kod çözücü mimarilerinden biri olan birleştirme (merge) mimari yaklaşımı, görüntüyü önek RNN kodlandıktan sonra modele verir ve her zaman adımında görüntü gösteriminden değişiklik yapılmaz (Hendricks et al., 2016; Tanti et al., 2018).

Bu çalışmada ise, kodlayıcı-kod çözücü yaklaşımına dayanan birleştirme modeli tabanlı görüntü altyazılama sistemi önerilmektedir. Görüntülerin görsel özniteliklerini çıkarmak için VGG16 CNN mimarisi, metin açıklamasının kodlanması için de LSTM RNN mimarisi kodlayıcı olarak kullanılmıştır.



Şekil 1. Görüntü Altyazılamaya Genel Bakış (Flair, 2019)

Kodlanan verilerin basit bir kod çözücü modeline girdi olarak verilmesiyle de görüntü altyazılama yapılmıştır. Önerilen yöntem, Flickr8k/30k (Hodosh et al., 2013) veri kümelerinin mevcut İngilizce görüntü altyazıları ile eğitilmiş ve birleştirme modelinin üretilen görüntü altyazı başarısı üzerindeki etkisi ve literatürdeki çalışmalarla kıyaslanması için BLEUn (n = 1,..., 4) (Bilingual Evaluation Understudy) (Papineni, Roukos, Ward, & Zhu, 2002) performans metriği ile değerlendirilmiştir. Bu çalışmada diğer çalışmalardan farklı olarak; görüntü altyazılama modeli herkesin kolaylıkla erişebileceği, internetsiz ve hızlı bir sonuç alabileceği Android tabanlı akıllı telefon uygulamasına gömülmüştür.

Makalenin geri kalanı ise şu şekilde düzenlenmiştir: Bölüm 2'de, önerilen kodlayıcı-kod çözücü yaklaşımı ve Android tabanlı uygulama tanıtılmıştır. Bölüm 3'te veri kümesi, performans metriği ve karşılaştırmalı sonuçlar verilmiştir. Bölüm 4'te ise sonuçlara değinilmiştir.

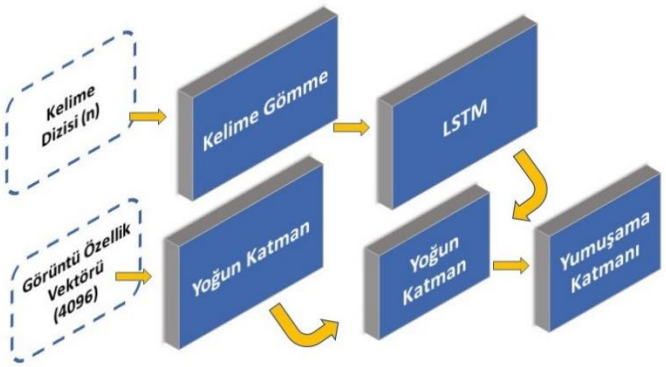
2. Materyal ve Metot

Bu bölümde önerilen görüntü altyazılama yöntemi ve *ImCap* adlı geliştirdiğimiz Android tabanlı uygulaması anlatılacaktır.

2.1. Önerilen Kodlayıcı-Kod Çözücü Yaklaşımı

Kodlayıcı-kod çözücü yaklaşımına dayalı görüntü altyazılama sisteminde, kodlayıcı olarak görüntülerin görsel özelliklerinin kayda değer içerikle çıkarılması için CNN mimarisi ve altyazı metinlerinin kodlanması - için sıralı verileri işleyebilen RNN mimarisi kulanılmıştır. CNN mimarileri evrişim işlemi yapan filtreler kullanan evrişim katmanı, bu katmanda çıkarılan öznitelik haritasının uzamsal boyutunu düşüren bir havuzlama katmanı ve bir önceki katmandan gelen tüm girdilere bağlı son çıktıyı üreten tam bağlı katmandan oluşur (Zhang, Wang, & Liu, 2018). Kodlayıcı olarak büyük veri kümelerinde önceden eğitilmiş CNN mimarilerinden yararlanılmaktadır.

Kod çözücü olarak kullanılan RNN, dil bilgisine uygun ve anlamlı bir cümle üretmek için kodlayıcıdan gelen özellik vektörünü alır. Sıralı verileri işleyen bir derin ağ olan RNN mimarilerinde her bir çıktı, tekrarlayan şekilde her bir dizi örneği üzerinden aynı fonksiyon işlenerek hesaplanır (Zhang et al., 2018). RNN mimarisi kelimeleri vektör olarak ifade eden gömme katmanı, işlenen dizilerin her sözcüğünün tahmin edilmesi için eğitilen tekrarlayan gizli katmanlardan ve görüntü özniteliklerine karşılık gelen en uygun sözcüğü tahmin eden tam bağlı katmandan oluşmaktadır. Geleneksel RNN yapılarında karşılaşılan kaybolan gradyan problemine çözüm olarak geliştirilen LSTM ağları, birleştirme modelinde kullanılmıştır. LSTM, yapısında uzun süreli hafıza yetisi barındırdığından metin üretmek ve kelimeleri daha doğru sıralamak için tercih edilmiştir. LSTM gizli ve hücre durumları olmak üzere iki durum vektörü içerir (Wang, Wang, & Xu, 2020).



Şekil 2. Birleştirme Modeli (Brownlee, 2019)

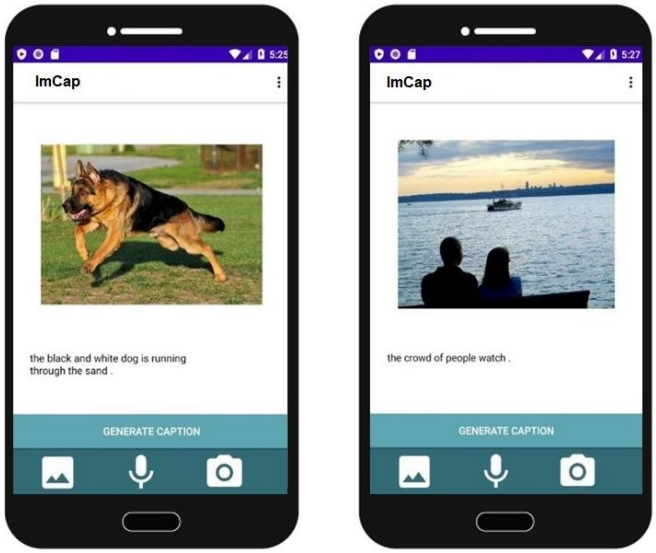
LSTM yapısı hücre, giriş, çıkış ve unutma kapısı olmak üzere dört kapıdan oluşur. Hücre kapısı, belirli zaman aralıklarındaki değerleri hatırlar. Giriş, çıkış ve unutma kapıları, hücrenin içine ve dışına bilgi akışını düzenler. LSTM yapısı, bir zaman serisindeki önemli olaylar arasında bilinmeyen süre gecikmeleri olabileceğinden, zaman serisi verilerine dayalı olarak sınıflandırma, işleme ve tahminler yapmak için uygundur (Zhang et al., 2018). Geleneksel görüntü altyazılamada kullanılan sistem Şekil 1'de gösterilmiştir.

Bu çalışmada, 16 katmandan oluşan VGG16 mimarisi kodlayıcı olarak çalıştırılmıştır. Bu mimari, 3 evrişim ve 2 tam bağlı katman, ardından yumuşama katmanı ile bir çıktı üreten bir CNN mimarisidir. Bu mimarinin sınıflandırma katmanı çıkarılarak, önceki katmanlarla girdi görüntüsünün özellik vektörü çıkarılır. Kod çözücü kısmında ise birleştirme modeli kullanılarak, kodlayıcıyla çıkarılan görüntü özellikleri ile referans altyazılarının LSTM ile kodlanmış biçimi iki girdi olarak alınır. Birleştirme modeli Şekil 2'de gösterilmiştir. Birleştirilen bu iki kodlanmış girdi, daha sonra kod çözücü modeli tarafından dizideki bir sonraki sözcüğü üretmek için kullanılmıştır. RNN yapısı yalnızca üretilen metni kodlamak için kullanılmıştır. Birleştirme modelinde, görüntü RNN yapısının dışında bırakılmıştır. RNN, sadece görüntü altyazısı önekini yani tamamen dilsel bilgiyi işler. Önek vektörleştirildikten sonra, görüntü öznitelik vektörü, LSTM ağından sonra gelen ayrı bir katmanda önek vektörüyle birleştirilir. Çalışmada, birleştirme modeli tabanlı LSTM yapısının görüntü altyazısı üretme başarısı değerlendirilmiştir.

2.2. Android Tabanlı Uygulama

Önerilen model, internet bağlantısı olmadan görüntü altyazılama yapabilen geliştirdiğimiz *ImCap* Android uygulama ile birleştirilmiştir. Bulut sistemi üzerinden haberleşen benzer uygulamalardan (Çaylı et al., 2020; Makav & Kılıç, 2019b) farklı olarak *ImCap* çevrimdışı çalışmaktadır.

Geliştirilen uygulama Java Programlama dili ile Android Studio platformunda yazılmıştır. Geliştirilen derin öğrenme modelinin Android cihazlarda verimli bir şekilde çalışması için TensorFlow kütüphanesinden yararlanılmıştır. Tensorflow, model kodlayıcısında görüntü öznitelik çıkarımı için dondurulmuş CNN mimarilerini sunmaktadır. Inception-v3, diğer CNN mimarilerinden farklı olarak dondurulmuş model dosyasına sahip olduğu için birleştirme modeli ile kodlayıcı-kod çözücü yapısında kullanılmıştır. Gömülü modelin eğitimi Flickr8k/30k veri kümeleri üzerinde gerçekleştirilmiş ve eğitim sonucunda elde edilen tüm parametre değerleri, CNN model koduna gerek kalmadan Android tabanlı akıllı telefona gömülmüştür.



Şekil 3. ImCap Android Uygulamaya Gömülen Model ve Üretilen Altyazılar

Üretilen model boyutunun büyük olması ve modelin çalıştırılacağı akıllı telefonun düşük işlem gücü ve düşük bellek alanına sahip olması nedeniyle model boyutunda optimizasyon yapılır. Optimizasyon için eğitim sonrası niceleme yöntemi (Leon et al., 2020) kullanılarak model eğitiminde veri kaybı olmadan model boyutunun dörtte birine düşürülmesi sağlanmıştır. Optimize edilmiş gömülü model ve kelime öznitelik dosyası geliştirdiğimiz *ImCap* uygulaması ile entegre edilmiştir. *ImCap* uygulaması ile son kullanıcı, kamera ve galeriden görüntü seçerek görüntü altyazısını bir butona tıklayarak üretebilmektedir ve sesli anlatım seçenekleriyle de görüntü altyazısını otomatik duyabilmektedir. Android uygulama üzerindeki çıktılar Şekil 3'te verilmiştir.

3. Deneysel Sonuçlar

Bu bölümde görüntü altyazılamada kullanılan veri kümeleri ve performans metrikleri ve önerilen yöntem ile mevcut çalışmalar arasındaki karşılaştırmalı sonuçlar verilmiştir.

3.1. Veri Kümeleri ve Performans Metrikleri

Görüntü altyazılama sistemlerinde kullanılan veri kümeleri önemli rol oynamaktadır. Değerlendirme yapılabilmesi için geniş sayıda görüntü içeren ve referans altyazısına sahip veri kümeleri gerekmektedir. Flickr (Hodosh et al., 2013), MSCOCO (Lin et al., 2014) ve VizWiz (Bigham et al., 2010; Chen et al., 2015) görüntü altyazılama sistemlerinde kullanılan yaygın veri kümelerindendir. Bu çalışmada, Flickr8k ve Flickr30k veri kümeleri kullanılmıştır. Flickr8k ve Flickr30k veri kümeleri sırasıyla 8,092 ve 31,783 adet görüntü içermektedir. Tablo 1'de verildiği gibi her görüntü 5 referans altyazı ile tasvir edilmiştir. Flickr8k 6,000 eğitim, 1,000 test ve 1,000 değerlendirme görüntüsü içerirken, Flickr30k 29,783 eğitim, 1,000 test ve 1,000 değerlendirme görüntüsüne sahiptir. Önerilen yöntemle üretilen görüntü altyazısının performansını ölçmek için BLEUn (n = 1, 2, 3, 4) (Papineni et al., 2002) metriği kullanılmıştır. BLEUn metriği, referans altyazılar ve üretilen altyazının benzerlik oranını kıyaslayarak benzerlik yüzdesini sunar. BLEU metriğinin 4 farklı seviyesi üzerinden sonuç alınmıştır.

Tablo 1: Flickr Veri Kümesinden Örnek Bir Resim ve Referans Altyazıları

Referans Altyazılar:

- Blonde horse and girl in black shirt are staring at fire in barrel
- Girl and her horse stand by fire
- Girl holding horse lead behind fire
- Man and girl and two horses are near contained fire
- Two people and two horses watching fire

3.2. Karşılaştırmalı Sonuçlar

Önerilen birleştirme modeli tabanlı LSTM kod çözücü, VGG16 mimarisi ile Flickr8k/30k veri kümeleri üzerinde test edilmiştir. Eğitilen iki farklı veri kümesi üzerinde, modelin doğruluk performansları ve önerilen sistemin Flickr veri kümesini kullanan diğer modellerle kıyaslaması sırasıyla Tablo 2 ve Tablo 3'te verilmiştir. Görüldüğü gibi önerilen birleştirme modeli yaklaşımı BLEU1 metriğinde en iyi performansı göstermiştir. Kodlayıcı-kod çözücü yaklaşımı üzerinde birleştirme modelinin başarısı, tekrarlayan sinir ağının rolünün çıktı üretmek yerine girdiyi kodlamak olduğunu göstermiştir. Android uygulamaya gömülen görüntü altyazılama modeli de başarıyla altyazı üretebilmiştir. Tensorflow verilerinin dondurulması bilgi kaybına sebep olduğu için aynı görüntünün bilgisayardaki ve Android uygulama üzerindeki çıktısı arasında az bir fark gözlemlenmiş olsa da bilgisayar ve *ImCap* uygulaması üzerindeki çıktıların görüntüyle oldukça uyumlu olduğu saptanmıştır. Veri kümeleri kıyaslanacak olursa Flickr30k daha fazla kelime ve görüntü barındırdığı için Flickr8k'ya göre daha iyi sonuçlar vermiştir. Tablo 4'te Flickr veri kümelerinden alınmış örnek resimler için üretilen altyazılar gösterilmektedir.

Tablo 2: Flickr8k Veri Kümesinde Performans Karşılaştırılması

Yöntem	BLEU1	BLEU2	BLEU3	BLEU4
(Chen & Zitnick, 2014)	22.5	-	-	-
(Mao et al., 2014)	43.8	18.5	13.4	-
(Hodosh et al., 2013)	48.0	-	-	-
(Kiros et al., 2014)	51.0	-	-	-
Önerilen Yöntem	**51.4**	**25.3**	**15.0**	**6.2**

Tablo 3: Flickr30k Veri Kümesinde Performans Karşılaştırılması

Yöntem	BLEU1	BLEU2	BLEU3	BLEU4
(Chen & Zitnick, 2014)	18.9	-	-	-
(Mao et al., 2014)	46.9	19.6	12.5	-
Çok Katmanlı RNN (Mao et al., 2014)	54.7	23.9	19.5	-
Önerilen Yöntem	**55.4**	**25.9**	**15.0**	**6.2**

4. Sonuç

Bu çalışmada VGG16 ve LSTM mimarileri kullanılarak kodlayıcı-kod çözücü yaklaşımına dayanan birleştirme modeli tabanlı bir görüntü altyazılama sistemi sunulmuştur. Önerilen sistem, Flickr8k ve Flickr30k veri kümeleri ile değerlendirilmiş ve birleştirme modeli tabanlı yöntemin görüntü altyazılamanın başarısına etkisi gözlemlenmiştir. Sistem, akıllı telefonlarda çevrimdışı çalışmasını sağlayan modele dönüştürülmüş ve *ImCap* adlı geliştirdiğimiz Android uygulama ile birleştirilmiştir. Bu uygulama ile, görme engelli bireylerin günlük hayatlarını kolaylaştıracak, seslendirme özelliğine sahip kullanıcı dostu bir platform sunulmuştur. Gelecek çalışmalarda, Android'e gömülü model oluşturmak için TensorFlow ve yüksek seviyeli yapay ağ API'sı olan Keras'ın birlikte kullanılması ve daha yüksek doğruluk performasına sahip modeller geliştirilmesi hedeflenmektedir.

Kaynakça

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., . . . White, S. (2010). *Vizwiz: nearly real-time answers to visual questions.* Paper presented at the Proceedings of the 23rd annual ACM symposium on User interface software and technology.

Brownlee, J. (2019). A gentle introduction to pooling layers for convolutional neural networks. *Machine Learning Mastery, 22*.

Çaylı, Ö., Makav, B., Kılıç, V., & Onan, A. (2020). *Mobile Application Based Automatic Caption Generation for Visually Impaired.* Paper presented at the International Conference on Intelligent and Fuzzy Systems.

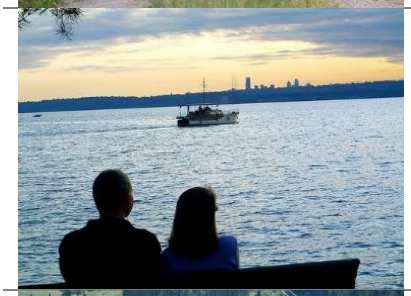
Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft coco captions: Data collection and evaluation server. *J arXiv preprint arXiv:.00325*.

Chen, X., & Zitnick, C. L. (2014). Learning a recurrent visual representation for image caption generation. *J arXiv preprint arXiv:1411.5654.*

Elliott, D., & Keller, F. (2013). *Image description using visual dependency representations.* Paper presented at the Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing.

Flair, D. (2019). Python based Project – Learn to Build Image Caption Generator with CNN and LSTM.

Tablo 4: Önerilen Altyazılama Yaklaşımı ile Üretilen Altyazılar

Flickr Görüntüleri	Referans Altyazılar	Üretilen Altyazılar
	*A big black and brown dog plays outdoor *A black and tan dog leaps over the green grass *A brown and black dog runs on the grass outdoors in front of a sidewalk *A dog runs *A German shepherd jumps left on patchy grass	***German sphered runs through the grass**
	*A boy does a skateboard trick *A man jumps in the air on his skateboard *A skateboarder doing a maneuver in the air the *Skateboarder jumps into the air *The young man catches some air while skateboarding	***Man is performing skateboard in the air**
	*Men participate in a soccer game *Two male soccer players in game action *Two men fall as they battle for a soccer ball *Two opposing footballers play on a field *Two soccer players playing soccer on a field	***Two man are playing soccer on field**
	*Two children carry flowers as they walk along a grassy track *Two children hold hands and flowers *Two children in jackets walking down a dirt road carrying flowers *Two children walk along a path holding purple flowers *Two girl walk town a dirt road, holding flowers	***Young boy in red shirt is walking through grassy field**
	*A couple watches a boat against a skyline *A man and woman sit on a bench watching a boat go *The sun is setting while a man and woman watch a boat go *Two people sit on a bench and watch a boat on the water *Two people watching a boat sail past	***Two people are sitting on rocks near water**
	*A boy wearing a red shirt and jeans is doing a flip on his bike *A person flipping a bicycle upside down *A person flips on a bike *A person in a red shirt doing tricks on a bicycle *A person is show upside down on his bicycle over a large field	***A men is performing trick on a bike**

Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). *Deep compositional captioning: Describing novel object categories without paired training data.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

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

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019). A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys, 51*(6), 1-36.

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *J arXiv preprint arXiv:.*

Kuznetsova, P., Ordonez, V., Berg, T. L., & Choi, Y. (2014). Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association for Computational Linguistics, 2*, 351-362.

Leon, V., Mouselinos, S., Koliogeorgi, K., Xydis, S., Soudris, D., & Pekmestzi, K. (2020). A tensorflow extension framework for optimized generation of hardware cnn inference engines. *J Technologies, 8*(1), 6.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., . . . Zitnick, C. L. (2014). *Microsoft coco: Common objects in context.* Paper presented at the European Conference on Computer Vision.

Makav, B., & Kılıç, V. (2019a). *A new image captioning approach for visually impaired people.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Makav, B., & Kılıç, V. (2019b). *Smartphone-based image captioning for visually and hearing impaired.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:.*

Mason, R., & Charniak, E. (2014). *Nonparametric method for data-driven image captioning.* Paper presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

Mathews, A., Xie, L., & He, X. (2018). *Semstyle: Learning to generate stylised image captions using unaligned text.* Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th Annual Meeting on Association for Computational Linguistics.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics, 2*, 207-218.

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

Wang, H., Wang, H., & Xu, K. (2020). Evolutionary Recurrent Neural Network for Image Captioning. *Neurocomputing*.

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Journal of Wiley Interdisciplinary Reviews: Data Mining Knowledge Discovery, 8*(4), e1253.