



Image captioning in Turkish language: Database and model

Tuğba Yıldız*^{ORCID}, Elena Battini Sönmez^{ORCID}, Berk Dursun Yılmaz^{ORCID}, Ali Emre Demir^{ORCID}

Department of Computer Engineering, Istanbul Bilgi University, Istanbul, 34060, Turkey

Highlights:

- Presentation of the MS COCO database captioned in Turkish language
- A model for Turkish image captioning
- Introduction of a Web-app for crowd sourcing the Turkish MS COCO database

Keywords:

- Turkish image captioning
- Turkish MS COCO
- Computer vision
- Natural language processing
- CNN, RNN

Article Info:

Research Article
Received: 26.07.2019
Accepted: 17.05.2020

DOI:

10.17341/gazimmfd.597089

Correspondence:

Author:
Tuğba Yıldız
e-mail:
tugba.yildiz@bilgi.edu.tr
phone: +90 212 311 7506

Graphical/Tabular Abstract

The fast growth of images on the internet turns out an increasing demand to understand the image and generate a caption. Automated image captioning is the problem of generating textual description of an image. This paper uses the successful encoder-decoder technique for image captioning in Turkish language. Moreover, it produces and releases a database of images described in Turkish language, and it implements a Web platform, to allow the improvement of the newly released dataset via crowdsourcing.

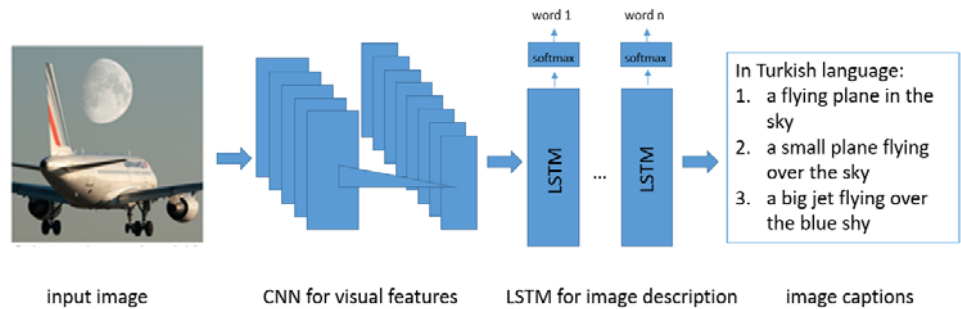


Figure A. System architecture

Purpose: Image captioning is the art of describing an image with sentences. The explanation of an image requires several tasks including the recognition of salient objects present in the picture, the understanding of their semantic relationships, the comprehension of the scene represented in the background and the capability to convert that knowledge into a syntactically correct sentence. A practical application of this research is also to support blind people with description of the surrounding environment. To date, the number of studies in Turkish language is still too limited and requires further investigation.

Theory and Methods:

This work used Python code to automatize Yandex translation API and convert all captions of the MS COCO (Lin et al. [20]) database from English language into Turkish language. The resulting Turkish captioned MS COCO database was used to test the proposed model for image captioning in Turkish language. Considering the recent developments in the machine translation field, the used image-captioning model employs an encoder-decoder framework, where a Convolutional Neural Network (CNN) encodes the image into a fixed-length vector representation, and a Long-Short Term Memory (LSTM) maps those vectors and generates image descriptions in Turkish language. In this study, we created two models. In Model-1 the weights of the used pre-trained CNN were frozen, while in Model-2 CNN and LSTM were fine-tuned together.

Results:

The proposed models were evaluated using both human based evaluations, and the most common metrics such as BLEU, METEOR, ROUGE and CIDEr. Both qualitative and quantitatively evaluations were satisfactory. In all cases, Model-2 had higher performance.

Conclusion:

This study introduces a novel Turkish captioned database together with a model to generate captions in Turkish language. The provided Web application will allow for crowd sourcing and the resulting Turkish captioned MS COCO database will be available for research purpose.



Türkçe dilinde görüntü altyazısı: Veritabanı ve model

Tuğba Yıldız*^{ID}, Elena Battini Sönmez^{ID}, Berk Dursun Yılmaz^{ID}, Ali Emre Demir^{ID}

İstanbul Bilgi Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, İstanbul, 34060, Turkey

Ö N E Ç I K A N L A R

- MS COCO veri tabanından Türkçe görüntü altyazısı oluşturma
- Türkçe görüntü altyazısı için bir model
- Türkçe MS COCO veri kümesine yönelik bir Web uygulaması

Makale Bilgileri

Araştırma Makalesi
Geliş:26.07.2019
Kabul:17.05.2020

DOI:

10.17341/gazimmfd.597089

Anahtar Kelimeler:

Türkçe görüntü altyazısı,
Türkçe MS COCO veri
kümesi,
bilgisayarla görme,
doğal dil işleme,
ESA, TSA

ÖZET

Otomatik görüntü altyazısı, yapay zekânın hem bilgisayarla görme hem de doğal dil işleme alanlarını kapsamaktadır. Makine çevirisi alanındaki gelişmelerden ilham alan ve bu alanda başarılı sonuçlar veren kodlayıcı-kod çözücü tekniği, özellikle İngilizce için otomatik görüntü altyazısı oluşturma konusunda kullanılan mevcut yöntemlerden biridir. Bu çalışmada ise, Türkçe dili için otomatik görüntü altyazısı oluşturan bir model sunulmaktadır. Bu çalışma, verilen görüntülerin özelliklerini çıkarmaktan sorumlu olan, Evrişimsel Sinir Ağı (ESA) mimarisine sahip bir kodlayıcıyı, altyazı oluşturmaktan sorumlu olan, Tekrarlayan Sinir Ağı (TSA) mimarisine sahip bir kod çözücüsü ile birleştirerek, Türkçe MS COCO veri kümesi üzerinde Türkçe görüntü altyazısı kodlayıcı-kod çözücü modelini test etmektedir. Modelin performansı, yeni oluşturulan veri kümesinde insanlar tarafından değerlendirilirken, bir taraftan da BLEU, METEOR, ROUGE ve CIDEr gibi en yaygın değerlendirme ölçütleri kullanılarak değerlendirilmiştir. Sonuçlar, önerilen modelin performansının hem niteliksel hem de niceliksel olarak tatmin edici olduğunu göstermektedir. Çalışma sonunda hazırlanan, herkesin kullanımına açık bir Web uygulaması (<http://mscoco-contributor.herokuapp.com/website/>) sayesinde Türkçe dili için MS COCO görüntülerine ait Türkçe girişlerin yapıldığı bir ortam kullanıcıya sunulmuştur. Tüm görüntüler tamamlandığında, Türkçe diline özgü, karşılaştırmalı çalışmaların yapılabileceği bir veri kümesi tamamlanmış olacaktır.

Image captioning in Turkish language: database and model

H I G H L I G H T S

- Presentation of the MS COCO database captioned in Turkish language
- A model for Turkish image captioning
- Introduction of a Web-app for crowd sourcing the Turkish MS COCO dataset

Article Info

Research Article
Received: 26.07.2019
Accepted: 17.05.2020

DOI:

10.17341/gazimmfd.597089

Keywords:

Turkish image captioning,
Turkish MS COCO database,
computer vision,
natural language proc.
CNN, RNN.

ABSTRACT

Automatic image captioning is a challenging issue in artificial intelligence, which covers both the fields of computer vision and natural language processing. Inspired by the later advances in machine translation, a successful encoder-decoder technique is currently the state-of-the-art in English language captioning. In this study, we proposed an image captioning model for Turkish Language. This paper evaluates the encoder-decoder model on MS COCO database by coupling an encoder Convolutional Neural Network (CNN) -the component that is responsible for extracting the features of the given images-, with a decoder Recurrent Neural Network (RNN) -the component that is responsible for generating captions using the given inputs- to generate Turkish captions. We conducted the experiments using the most common evaluation metrics such as BLEU, METEOR, ROUGE and CIDEr. Results show that the performance of the proposed model is satisfactory in both qualitative and quantitative evaluations. Finally, this study introduces a Web platform (<http://mscoco-contributor.herokuapp.com/website/>), which is proposed to improve the dataset via crowd-sourcing and free to use. The Turkish MS COCO dataset is available for research purpose. When all the images are completed, a Turkish dataset will be available for comparative studies.

*Sorumlu Yazar/Corresponding Author: *tugba.yildiz@bilgi.edu.tr, elena.sonmez@bilgi.edu.tr, dursun.yilmaz@bilgi.edu.tr, emre.demir05@bilgi.edu.tr / Tel: +90 212 311 7506

1. GİRİŞ (INTRODUCTION)

İnternet üzerinden paylaşılan görüntü sayısındaki hızlı artış, görüntüyü anlama ve görüntü altyazısı (image captioning) oluşturma konusundaki çalışmaların da artmasına sebep olmuştur. Bir görüntüye ait altyazının otomatik olarak oluşturulması, verilen bir görüntünün metinsel tanımını oluşturma problemidir. Bu nedenle, görüntü altyazısı oluşturma problemi hem bilgisayarlı görme hem de doğal dil işleme alanının bir parçası olarak görülmektedir. Bu bağlamda, bir görüntü altyazısı oluşturmak için bir görüntünün sadece içeriğini anlamak yeterli değildir. İçeriğe ek olarak; söz dizimsel ve anlamsal olarak doğru bir altyazı bulabilmek için nesnel arasındaki anlamsal ilişkileri ve aktiviteleri çıkarmak, nitelikleri tespit etmek, insanların kullandığı cümleler gibi tanımlamalar yapmak önemlidir.

Görüntü altyazısı oluşturmaya dayalı çalışmalar, kabaca ikiye ayrılır: şablon-tabanlı (template-based) ve erişim-tabanlı (retrieval-based). Üretici-tabanlı (generative-based) yaklaşımı benimseyen şablon-tabanlı yöntemler, belirtilen bir dizi görsel kavramı algılar ve bir cümle şablonuna veya dilbilgisi kurallarına dayanarak altyazı oluştururken [1-4], erişim-tabanlı yöntemler ise bir dizi önceden belirlenmiş cümleden altyazı oluştururlar [5-9]. Her iki yöntem de dilbilgisi açısından doğru ve akıcı cümleler üretse de, konu ile alakalı olmayan açıklamaları, kapsamı, yaratıcılığı ve ortaya çıkan cümlelerin karmaşıklığı gibi bazı dezavantajlar ve kısıtlamalara sahiptir. Şablon-tabanlı yöntemlerle oluşturulan başlıklar doğal görünmezken, çok katı ve çeşitlilik eksikliklerine sahipken, erişim-tabanlı yöntemlerle oluşturulan başlıklar, doğrudan aktarılması nedeniyle bir görüntünün içeriğini doğru şekilde tanımlayamayabilmektedir. Bu sebepten son teknoloji ürünü çözümlerin hala insanlar tarafından oluşturulan başlıklardan uzak olduğu da yadsınamaz bir gerçektir.

Son yıllarda, derin sinir ağları oldukça popüler olmuş ve farklı alanlarda başarılar göstermişlerdir. Evrimsel Sinir Ağları (ESA) – Convolutional Neural Networks [10] görüntüler üzerinde önemli performans gösterirken [11, 12], Tekrarlayan Sinir Ağları (TSA) – Recurrent Neural Network [13, 14] doğal dil işlemede önemli bir rol oynamıştır. İlk olarak makine çevirisi probleminde [15-17], Sıradan-Sıraya (Seq2Seq) kodlayıcı-kod çözücü mimarisinin başarı ile çalışması, görüntü altyazısı üzerine yapılan çalışmalar için de ilham kaynağı olmuştur. Vinyals vd. [18] çalışması, görüntü özelliklerini çıkarmak için ESA mimarisi üzerine kurulu bir kodlayıcı ve görüntü altyazılarını oluşturmak için bir kod çözücü olarak Uzun Kısa Süreli Bellek (UKSB) [19] kullanan ilk çalışma olması açısından önem taşımaktadır. Vinyals vd. [18] çalışmasından esinlenerek yapılan bu çalışmada, Türkçe dili için otomatik görüntü altyazısı oluşturan bir model için, görüntüyü sabit uzunlukta bir vektör temsili olarak kodlamak için ESA yapısı kullanılmış, Türkçe için görüntü altyazısı ve bu vektörleri haritalamak ve Türkçe dilinde bir görüntü açıklaması oluşturmak için ise TSA'nın bir çeşidi olan UKSB'den faydalanılmıştır. Deneysel

çalışmalar, MS COCO veri kümesi [20] üzerinde gerçekleştirilmiştir. Bu çalışmanın ana katkıları:

- MS COCO veri kümesi kullanılarak, herkesin kullanımına açık, görüntü ve Türkçe altyazı veri kümesinin oluşturulması
- Türkçe altyazı modelinin oluşturulması
- Bir web uygulamasının (<http://mscoco-contributor.herokuapp.com/website/>) üzerinden veri kümesinin oluşturulması

Literatürde görüntü altyazısını otomatik olarak oluşturmaya yönelik çeşitli çalışmalar önerilmiştir. Bu çalışmalar, en kesin ve en iyi bilinen yöntem olarak benimsenmiş olan sinir ağlarına dayanmaktadır. Son çalışmalar özellikle makine çevirisi problemine çözüm olarak sunulan Sıradan-Sıraya kodlayıcı-kod çözücü çerçevelerinin [15-17] başarısından ilham almıştır. Makine çevirisi probleminde kodlayıcı, sabit uzunlukta bir vektöre bir giriş cümlesi kodlarken, kod çözücü, kodlanmış vektörden bir çeviri ortaya çıkarır. Benzer yaklaşım, verilen görüntüyü doğal dil cümlelerine çevirmek için de kullanılabilir.

Sinir ağları kullanılarak altyazı kalitesini iyileştirmeyi ve yüksek performans elde etmeyi amaçlayan çeşitli çalışmalar literatürde mevcuttur. Görüntü altyazısı için sinir ağları kullanmaya yönelik ilk girişim Kiros vd. [21] tarafından sunulmuştur. Çalışmada, derin sinir ağlarından öğrenen kelime temsilleri ve görüntü özelliklerine dayanan, log-bilinear prototipini benimseyen iki çok kipli (multimodal) dil modeli önerilmiştir. Deneysel çalışmalar, IAPR TC-12 ve Attributes Discovery veri kümeleri üzerinde gerçekleştirilmiştir. Diğer bir çalışmalarında [22] ise, görüntü metni gömme prototiplerini birleştirmek için bir model sunmuşlar ve bir önceki sonuçları önemli ölçüde geliştirmişlerdir [21]. Mao vd. [23], görüntü altyazısı oluşturma sorununu gidermek için çok kipli Tekrarlayan Sinir Ağı (m-TSA) modelini sunmuşlardır. Önerilen prototip iki alt ağdan oluşmaktadır: cümleler için derin bir TSA ve görüntüler için derin bir ESA. Her iki alt ağın bağlantısı çok kipli bir katmanla sağlanmaktadır. Prototip, IAPR TC-12, Flickr8K [24], Flickr30K [25] ve MS COCO [20] olmak üzere dört temel veri kümesi üzerinde değerlendirilmiştir.

Socher vd. [26] cümleler için vektör gösterimlerini öğrenen bağımlılık ağaçlarına dayalı Özyinelemeli Sinir Ağı modelini sunmuşlardır. Vinyals vd. [18], Sinirsel Görüntü Başlığı (SGB) adında, kodlayıcı olarak bir TSA'na ve ardından ilgili cümleyi oluşturmak için bir UKSB'e dayanan bir prototip tasarlamışlardır. Benzer şekilde, Donahue vd. [27], video tanıma, görüntü altyazı oluşturma ve video açıklama görevleri için Uzun Dönemli Tekrarlayan Evrimsel Ağ (UDES) mimarisi önermişlerdir. Çoklu kipli TSA mimarisine dayanan bir başka model Karpathy vd. [28] tarafından önerilmiştir. Sonuçlar önceki çalışmalarla [18, 22, 23, 26, 27] karşılaştırılmış ve önerilen prototipin diğerlerinden daha iyi performans gösterdiği kanıtlanmıştır. Yapılan benzer çalışmalar [29, 30] cümle oluşturmak için

kodlayıcı-kod çözücü yapısına anlamsal bilgiler ekleyerek çözüm önermişlerdir. Dikkat temelli yöntemlerin (attention-based) de, görüntü altyazısı oluşturmada iyi performans sergilediği gösterilmiştir. Xu vd. [31] dikkat mekanizması önererek [18]'in çalışmalarını geliştirmişlerdir. Çalışmada, görüntü altyazısı oluşturmak için yumuşak ve sert olmak üzere iki farklı dikkat mekanizmasına sahip bir model ortaya koymuşlardır. Dikkat temelli modellerin performansı, üç karşılaştırmalı veri kümesi, Flickr8K, Flickr30K ve MS COCO kullanılarak doğrulanmıştır. Yang vd. [30] çalışmasında, görüntü altyazısı ve kaynak kod yazısı problemlerine yönelik olarak, ESA ve TSA kullanan konvansiyonel dikkat mekanizmalı kodlayıcı-kod çözücülerin, standart kodlayıcı-kod çözücülerden daha üstün olduğunu göstermişlerdir. Literatürdeki diğer çalışmalar da [32-36] farklı dikkat temelli görüntü altyazı yöntemini kullanarak probleme çözüm sunmuşlardır.

Anlamsal kavram-temelli yöntemler, görüntülerden anlamsal kavramları çıkarmak için de kullanılır. Bu kavramlar, gizli durumlar (hidden states) ve TSA'nın çıktıları ile birleştirilebilir. You vd. [37] çalışmasında, görüntü altyazısını geliştirmek için anlamsal dikkat olarak sinir temelli yaklaşımı kullanmıştır. Yao vd. [38], hem görüntü temsili hem de üst düzey özellikleri kullanan UKSB mimarisi önermişlerdir.

Shetty vd. [39], bir çift ağ kullanarak tek bir görüntü için birden fazla görüntü altyazısı oluşturmak üzere Üremsel Çekişmeli Ağ (ÜÇA) - Generative Adversarial Networks tabanlı bir görüntü altyazı yöntemi kullanmıştır. Dai vd. [40], ÜÇA tabanlı bir görüntü altyazı yöntemi önermişlerdir.

Son zamanlarda, Aneja vd. [41] mekânsal görüntü özelliklerinden yararlanmak için dikkat mekanizmasına sahip evrimsel bir mimari önermişlerdir. Wang vd. [42] çalışmasında, Aneja vd. [41] çalışmasına dayanan ESA bazlı görüntü altyazısı metodunu hiyerarşik dikkat modülü kullanarak geliştirmişlerdir. Aynı makale hiperparametrelerin görüntü altyazı performansına etkisini de göstermiştir.

Yukarıda belirtilen görüntü altyazısı oluşturma alanındaki çalışmaların tümü İngilizce dili üzerine odaklanmış olsa da, Türkçe için önerilen az sayıda çalışma mevcuttur. Bu

çalışmalardan biri, Türkçe görüntü altyazısı veri kümesi oluşturma amacı taşıyan TasvirEt [43] çalışmasıdır. Ünal vd.[43], bu çalışmada Flickr8K veri kümesindeki görüntüleri kullanarak, görüntü altyazılarının en benzer görüntüler arasından aktarılması ve adaptif komşu temelli görüntü altyazısı olmak üzere iki farklı yöntem kullanmışlar ve 8091 görüntü için bir ve ya iki Türkçe açıklama eklemişlerdir. Başka bir çalışmada Samet vd. [44], MS COCO ve Flickr30k altyazılarını Google Çeviri Uygulama Program Arayüzü (Google API) kullanarak İngilizce'den Türkçe'ye çevirmeye yönelik bir yöntem sunmuşlardır. Kodlayıcı-kod çözücü yapısı, modeli eğitmek için kullanılmıştır. Sonuçlar TasvirEt ile karşılaştırılmış ve önerilen yöntemin, BLEU-1 puanında %7'lik bir artışla TasvirEt'i geride bıraktığı sonucuna varmışlardır. Kuyu vd. [45] çalışmasında, Byte Çifti Kodlama (BÇK) algoritmasını kullanan alt kelimelere dayanan bir model önermiştir. TasvirEt sonuçları, MS COCO ve Flickr30k veri kümeleri, alt kelime temelli derin öğrenme modelinin eğitim aşamasında kullanılmıştır. BLEU, METEOR, ROUGE ve CIDER metriklerine ek olarak, model performanslarını değerlendirmek için Word Mover's Distance doküman metriğini kullanmışlardır.

2. DENEYSEL METOT (EXPERIMENTAL METHOD)

2.1. Veri Kümesi (Dataset)

Bu çalışmada, Tsung-Yi Lin vd. [20] tarafından oluşturulan Microsoft Common Objects in COntext (MS COCO) veri tabanındaki görüntülerden faydalanılmıştır. MS COCO nesne segmentasyonu, bağlam tanınması ve görüntü altyazısı oluşturma gibi çalışmalarda sıklıkla kullanılmaktadır. 80 ortak nesne kategorisi (örneğin kişi, araba, köpek vb.), 91 alan kategorisi (örneğin gökyüzü, sokak, deniz vb.) içermektedir. 160.000 (160K)'dan fazla etiketli resimle toplam 330K görüntüye sahiptir. Görüntüler Amazon Mechanical Turk (AMT) çalışanları tarafından İngilizce dilinde yaklaşık 5 görüntü altyazısı ile etiketlenmiştir. Şekil 1, MS COCO veri tabanındaki etiketli görüntülerden bir örnek içermektedir.

Bu çalışma kapsamında, MS COCO veri kümesindeki tüm etiketli görüntüler kullanılmış ve görüntüye ait altyazılar ilk olarak Yandex Çeviri Uygulama Program Arayüzü (Yandex API) üzerinden Türkçe'ye çevrilmiştir. Yaklaşık 164K



a horse with a half white and half black face.
a horse in a pen stands on the side in partial shade.
a horse standing on a dirty floor next to a wooden fence.
an image of a horse that is inside of a coral
a gray horse wit a white muzzle is standing in a fenced enclosure.

Şekil 1. MS COCO veri kümesinden alınmış orijinal bir görüntü ve görüntüye ait 5 farklı İngilizce altyazı örneği
(An original image from MS COCO dataset with 5 different English captions)

görüntü için 616,767 altyazının ($\approx 36M$ karakter) çeviri işlemi, INTEL i7 6950X 3.00GHz 25M R2PA işlemci, ASUS STRIX-GTX1080-A8G-GAMING DVI 2HDMI 2DP 8GD5X ekran kartına sahip bir yüksek performanslı bilgisayar üzerinde ~ 7 günde tamamlanmıştır.

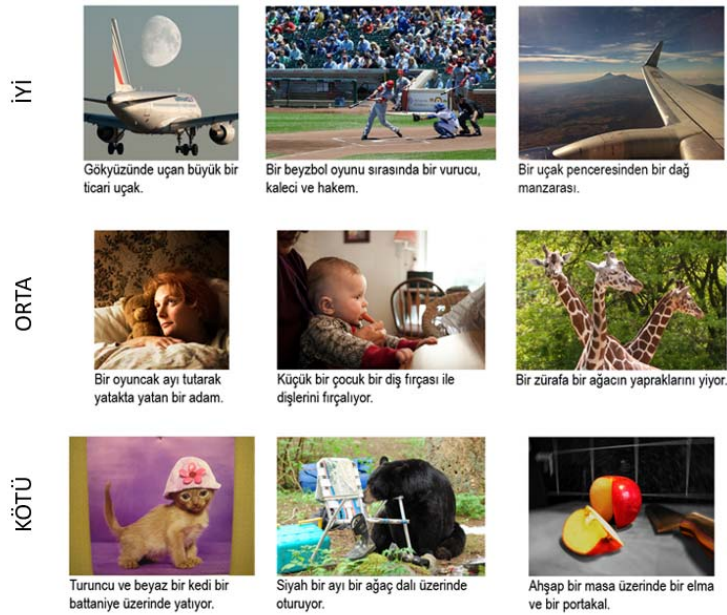
Şekil 2’de ise, Yandex API kullanılarak elde edilen Türkçe altyazı ile etiketlenmiş görüntülerin örnekleri gösterilmektedir. Her görüntü 5 altyazı içerirse de Şekil 2’de her görüntü için bir Türkçe altyazı eklenmiştir. Çevirinin kalitesi birinci satırdan üçüncü satıra doğru düşecek şekilde verilmiştir. Bu çeviriler, modelin ürettiği sonuçlarla karşılaştırma yapabilmek adına önemlidir.

2.2. Model (Model)

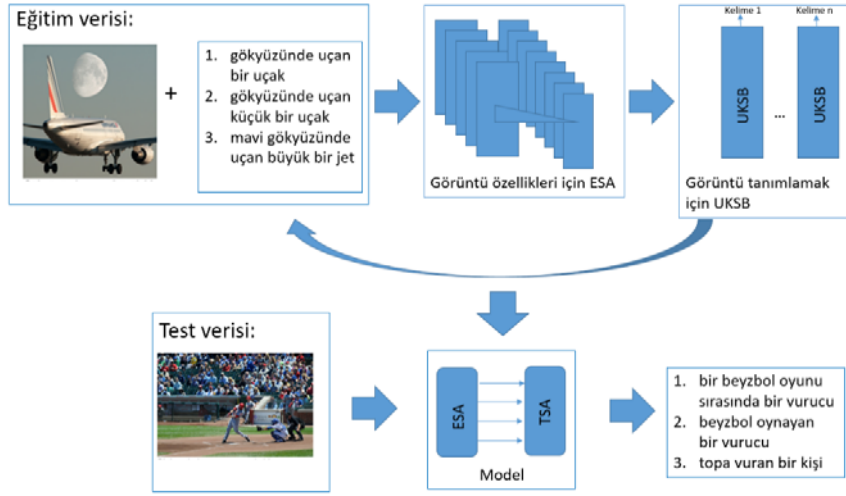
Cho vd. [16], makine çevirisi problemi için bir sinir ağı mimarisi önermiş, iki TSA içeren kodlayıcı-kod çözücü modelini kullanmışlardır. Kodlayıcı TSA, bir giriş dili dizisini sabit uzunluklu vektör temsiline kodlamayı öğrenirken, ikinci TSA, bir kod çözücü olarak sabit uzunluklu vektör gösterimini hedef dil dizisine eşlemektedir. Benzer yaklaşım Vinyals vd. [18] tarafından görüntüden altyazı çıkarma problemi için de kullanılmıştır. Kodlayıcı kısmında TSA yerine ESA kullanılarak verilen görüntülerin özelliklerinden sabit uzunluktaki bir vektör çıkartılarak görüntü altyazısı oluşturulmuş, görüntü altyazılama problemine önerilen tek bir boru hattı (pipeline) ağı kurulmuştur. Sistem mimarisi Şekil 3’de verilmiştir.

Birkaç sinir ağı katmanından oluşan ESA’nın birinci katmanı, özellikleri açısından incelenecek olan görüntünün her pikseli tarafından beslenir. Özellikler, görüntüleri evriştirmek için çeşitli filtreler kullanılarak elde edilir.

Katmanların her biri birbirine ağırlıklı kenarlarla bağlanmıştır. Buradaki asıl amaç, sistemi anlama ve bu nedenle içeriği sınıfları temsil eden sabit uzunluklu bir vektör üzerinde sınıflandıracak hale getirecek makul ağırlıkları bulmaktır. Ağırlık ayarı geri yayılma teknikleri kullanılarak gerçekleştirilir. Modelin TSA bileşeni, UKSB olarak tercih edilmiştir. Bu tercihin arkasındaki ana fikir, Kaçış Gradyan Problemini (Vanishing Gradient Problem) önlemektir. Ağdaki her bir UKSB birimi bir hücre, bir giriş kapısı, bir çıkış kapısı ve bir unutma kapısından oluşur. Bu kapılar, mevcut değeri unutmak veya yeni hücre değerini çıkarmak gibi hücre davranışını kontrol etmek için kullanılmaktadır. Mevcut aşamaya kadar gözlenen girdilerin bilgisinin kodlanmasından hücreler sorumludur. Sonunda, hücrenin çıkışı, UKSB ünitesinin çıkışını alan ve sözcük tahmini için ilgili sınıfa bağlayan bir fonksiyon olan Düzgeli Üstel Fonksiyon (Softmax) ile beslenir. ESA, TSA’nın girişini beslemek için önerilen sinir ağının hayati bileşenlerinden biridir. ESA’nı elde etmenin birçok yolu vardı. Bunlardan biri ESA’larını sıfırdan eğitmektir. Bir diğeri ise, önceden eğitilmiş olanları kullanmaktır. Alternatiflerden bazıları şunlardır: ResNet, Inception v3. ImageNet veri seti ile eğitilmiş olan Inception v3’ün son sürümünün oldukça iyi performans gösterdiği kanıtlanmış olduğundan, önerilen sinir ağı için kodlayıcı bileşeni olarak da kabul edilmektedir. Bu çalışmada, Vinyals vd. [18] tarafından önerilen SGB prototipinden ilham alınmıştır. SGB, bir görüntüyü temsil etmek için ESA ve İngilizce olarak altyazıyı tanımlamak için UKSB kullanmıştır. Çalışmamızda, eğitim kümesi çiftler (I, S)’den oluşmaktadır. I, MS COCO veri kümesinde bir görüntü ve S ise o görüntüyü tanımlayan bir Türkçe cümledir. Eğitim süresinde, algoritma formülü en üst düzeye çıkarabilecek en iyi θ^* parametresini Eş. 1’i kullanarak arar:



Şekil 2. MS COCO veri tabanından alınan görüntülerin Yandex API kullanılarak Türkçe altyazı ile örnekleme (Samples of images from MS COCO database with Turkish subtitles using Yandex API)



Şekil 3. Sistem mimarisi (System architecture)

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{(I,S)} \log p(S|I, \theta) \quad (1)$$

Zincir kuralı denklemleri kullanılarak formül yeniden yazılabilir:

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_{(I,S), t=0}^N \log p(S_t | I, \theta, S_0, \dots, S_{t-1}) \quad (2)$$

Bu formülde N, Türkçe başlıkta kullanılan kelime sayısı ve S_0, \dots, S_N model tarafından üretilen kelimelerdir. Bu durumda (I, S), eğitim kümesinin bir örneğidir ve model, tüm eğitim kümesine stokastik gradyan inişi uygulayarak en iyi parameter θ^* yi hedeflemektedir.

Formül 2'deki $p(S_t | I, \theta, S_0, \dots, S_{t-1})$ kodlayıcı-kod çözücü çifti ile sağlanmıştır. Kodlayıcı, giriş görüntüsü olan I'yi ilgili başlığın oluşturulması için kod çözücü tarafından kullanılan bir dizi sabit uzunluklu özelliğe dönüştürür. Daha ayrıntılı olarak 2D görüntüsünün parçalarından özellik vektörlerini çıkarmak için bir ESA yapısı ve görüntü altyazısı oluşturmak için UKSB yapısı kullanılmıştır. Görüntü altyazısı, bir kelime dizisi; t zamanda oluşturulan kelimenin önceden oluşturulan tüm kelimeler, gizli durum ve bağlam vektörü tarafından koşullandırılmıştır. Vinyals vd. [18] önerdiği boru hattı modelinin sistem bileşenlerini birleştirme prosedüründe bazı ayarlamalar dikkate alınmalıdır. Yukarıda açıklandığı gibi, ESA bileşeni Inception v3 olarak seçilebilir ve olumsuz bir etkiden kaçınmak için ağırlıkları değişmeden kalabilir. Bunun aksine TSA bileşeni rastgele ağırlıklarla seçilebilir. ESA'nda ince ayar (fine-tuning) yapılması, UKSB parametrelerinin geri yayılması ve belirlenmesine kadar askıya alınmalıdır, çünkü UKSB'lerle aynı anda ayarlama denemesi olumsuz sonuçlara yol açabilir. Bu sebepten, bu çalışma iki farklı şekilde değerlendirilmiştir. Birinci model (Model-1) için TSA tek başına, ESA'nı dondurmak suretiyle yaklaşık 1M epok (epoch) için eğitilmiştir. Ardından ikinci model (Model-2), ek bir 2M epok için ortaklaşa (ESA-UKSB) eğitilmiştir. Böylelikle Model-1'e kıyasla performansı artırarak, MS COCO olarak seçilen veri

kümesinin kapsamına odaklanma olanağını bulunmuştur. Bu çalışmada, yığın boyutu 128, gömme boyutu 512, TSA gizli durum boyutu 512 ve öğrenme oranı $1e-3$ olarak belirlenmiştir. Tablo 1, kullanılan görüntülerin dağılımını detaylı şekilde göstermektedir. Deneysel çalışmalar, Python'da yazılmış ve TensorFlow üzerinde çalışabilen yüksek seviyeli bir yapay ağ API'si olan Keras (<https://keras.io/>) ile gerçekleştirilmiştir. Değerlendirme sonuçları bir sonraki bölümde sunulmuştur.

Tablo 1. Kullanılan görüntülere ait bilgiler
(Information of the used images)

	Train	Validation	Test	Total
Görüntü sayısı	83K	41K	41K	165K

3. SONUÇLAR VE TARTIŞMALAR (RESULTS AND DISCUSSIONS)

Görüntü altyazılarının başarısını değerlendirmek için çeşitli değerlendirme ölçümleri kullanılmaktadır. Temel olarak kullanılan iki tür değerlendirme kategorisi vardır: insan temelli değerlendirme ve otomatik değerlendirme. Otomatik değerlendirme için kullanılan ölçütlerinden bazıları şunlardır: BLUE, ROUGE, METEOR ve CIDEr. Tüm otomatik metriklerin olası skorları 0 ile 1 arasında değişmektedir; burada 1'e yakın bir skor, makine çevirisi ile elde edilen çevirinin bir insan çevirisi kadar başarılı olduğu anlamına gelir. Genel olarak, çalışmaların çoğu, modellerinin başarısını değerlendirmek için bu yöntemleri tercih etmektedir. Bu fikrin arkasındaki ana sebep, değerlendirmelerin objektif olarak yapılmasını sağlamaktır. Kılıçkaya vd. [46], görüntü altyazısı için otomatik metrikler üzerine kapsamlı bir araştırma yapmış, nasıl yorumlanacağına dair kapsamlı bir bakış açısı ortaya koymuşlardır. Bu çalışmada da, iki tür değerlendirme yapılmıştır: insan temelli ve otomatik değerlendirme. Otomatik değerlendirme metriklerinin tümü kullanılmış, karşılaştırmalar yapılmıştır.

3.1. İnsan temelli değerlendirme (Human-based Evaluation)

Modelimizin başarısını değerlendirmek için, örnek görüntü ve altyazılara ihtiyaç duyulmuştur. Bu sebepten öncelikle MS COCO veri kümesinde bulunan görüntülere ait altyazılar Yandex API kullanılarak Türkçe'ye çevrilmiştir. Bu çevirinin başarısını değerlendirmek için Şekil 4'de ekran görüntüsü verilen "Çeviri Değerlendirme Formları" oluşturulmuş, toplamda 800 rastgele görüntü üzerinde, çevirilerin başarısının değerlendirilmesi için, üniversitede çalışan öğrencilerden formu kullanarak sonuçları değerlendirmeleri beklenmiştir. İnsan temelli bu değerlendirmenin sonucunda, Yandex API kullanılarak, Türkçe MS COCO veri kümesi üzerinde yapılan çevirilerin doğruluğunun %51 olduğu gösterilmiştir. Sonrasında aynı veri kümesi üzerinde Bölüm 2.2'de önerilen modeller çalıştırılmıştır. Modelin ürettiği Türkçe altyazıların başarısını değerlendirmek amacı ile Şekil 5'de görüldüğü gibi "Altyazı Değerlendirme Formları" hazırlanmış, değerlendirme yapacak insanlara 800 görüntü ve bu görüntülere ait üç altyazı dağıtılmıştır. Önerilen modellerin ürettiği altyazıların insan gücü ile yapılan değerlendirmesi neticesinde Model-1 için doğruluk %58,5 iken Model-2 için %68,2 olduğu görülmüştür. Modelin ürettiği altyazıların, insan gücüne dayalı bu değerlendirme sonucunda Yandex API kullanılarak elde edilmiş çevirilerden daha iyi sonuçlar elde ettiği görülmüştür. Yandex API ile değerlendirmelerin doğruluğu %51 iken, Model-1'de bu oran %7,5, Model-2'de ise %17,2 artmıştır.

3.2. Otomatik değerlendirme (Automatic Evaluation)

Çalışmada önerilen modelin başarısını, otomatikleştirilmiş ölçüm araçları yardımıyla değerlendirmek için, örnek görüntüler ve görüntülere ait gerçek altyazılar gerekmektedir. Bunun için 800 görüntü içeren, Şekil 6'da görülen "Gerçek Referans Değeri Formları (Ground-Truth Form)" oluşturulmuştur. Nasıl doldurulması gerektiği ve gerekli uyarılar kullanıcılara yapılmıştır. Kullanıcıların her bir görüntü için üç altyazı girmesi istenmiştir.



Her iki model, oluşturulan bu veri kümesi üzerinde çalıştırılmıştır. Bu işlem sonrası iyi bilinen değerlendirme ölçütleri kullanılarak sonuçlar değerlendirilmiştir. Tablo 2'de otomatik değerlendirme sonuçları verilmiştir.

Tablo 2'den de görüldüğü üzere sonuçlar tatmin edicidir. Model-2, Model-1'den her ölçüt için daha iyi sonuç vermiştir. Model-1'de ESA'nı dondurarak sadece TSA'nın eğitilmesi sağlanırken, Model-2'de ESA-UKSB'in birlikte kullanılmasının performansı artırdığı görülmüştür. Sonuçları kendi arasında karşılaştırdığımızda ise BLUE-1 ve CIDEr sonuçları diğer metriklerden daha iyidir. MS COCO veri kümesinden direk yapılmış çeviriler olduğundan, cümlelerin daha uzun bölümlerini inceleyen metriklerin puanları daha düşüktür. Diğerleri arasında, CIDEr oluşan görüntü tanımlarının kalitesini değerlendirmek için önerilen en yeni metriktir. Model-2'deki CIDEr'in performans puanı 0,5'ten

Sample Records from MSCOCO Turkish Dataset

WARNING: Please peek the images for the ONLY purpose of getting an intuition regarding the context of the images. However, please score the sentences ONLY grammatically (DO NOT CONSIDER PUNCTUATION or LOWER&UPPER-CASE LETTERS)!

UYARI: Lütfen fotoğraflara YALNIZCA fotoğrafın içeriği hakkında fikir sahibi olmak için bakın! Cümleleri YALNIZCA ve YALNIZCA Türkçe dil bilgisi açısından (NOKTALAMA İŞARETLERİNİ veya KÜÇÜK&BÜYÜK-HARF AYRIMINI DİKKATE ALMAYIN) değerlendirin!


Number	Photo	Caption	Score
1		raketi tenis oynarken, kadın bir rekabet içinde .	Select ▾
2		küçük ev yatak odası sarı bir yatak.	Select ▾

Şekil 4. Çeviri değerlendirme form görüntüsü (Image of the form for translation rating)

Sample Captions Generated by Turkish-Supported Show and Tell Model

WARNING: You will find three captions per image down below. Please score the captions, considering all of three, based on their relevance to the respective images.

UYARI: Aşağıda her bir imaj başına 3 farklı altyazı göreceksiniz. Lütfen altyazıları, tümünü göz önünde bulundurarak, ait oldukları imajlarla olan ilişkileri açısından değerlendirin.

Number	Photo	Caption	Score
1		<p>1. Bir saat ile büyük bir kule.</p> <p>2. Bir saat ile bir kule üzerinde bir saat.</p> <p>3. Bir saat ile büyük bir kule üzerinde bir saat.</p>	Select ▾

Şekil 5. Altyazı değerlendirme form görüntüsü (View of the evaluation form for captions)

fazladır, bu da yapılan işin zorluğu göz önüne alındığında oldukça iyidir.

Tablo 2. Otomatik değerlendirme sonuçları
(Automatic evaluation results)

Algorithm	Model-1	Model-2
BLEU-1	0,288	0,297
BLEU-2	0,155	0,164
BLEU-3	0,071	0,076
BLEU-4	0,030	0,035
ROUGE_L	0,266	0,272
METEOR	0,125	0,129
CIDEr	0,479	0,528

Başka bir sonuç ise, önerilen model, görüntülerin içeriğini tanımakta yetkin olsa bile, dilbilgisi açısından doğru cümleler kurmada çoğu zaman yetersiz kalmıştır. Nispeten kısa sekanslar üretmek için UKSB mimarisinin aşırı karmaşık olabileceği sonucuna varılmıştır. Bu sebepten, bu çalışmanın sonraki adımlarında, farklı gizli durum boyutları ile çalıştırılabilir ve sayıdaki artışın performansı ne şekilde etkilediği gözlemlenebilir. Buna ek olarak, kod çözücü kısmında standart TSA, Geçitli Tekrarlayan Birim (GTB) ve UKSB kullanılıp, karşılaştırma yapılabilir. UKSB, GTB ile karşılaştırıldığında daha eğitilebilir parametrelere sahip olduğundan, UKSB tabanlı ağlar ile eğitim prosedürünün uzatılmasının, GTB tabanlı ağlardan aynı veya daha iyi

performans elde etmelerini sağlayıp sağlamadığını görmek bu çalışmanın devamı olabilir. Bunlara ek olarak, çalışmada kullanılan model ve oluşan Türkçe MS COCO veri kümesi, çoklu ortam (multimedia) belgeleri üzerinde de denenebilir [47]. Önerilen model, daha önce yapılan Türkçe çalışmalar [44, 45] ile karşılaştırıldığında sonuçlar arasında büyük farklar saptanmamıştır. Samet vd.'nin çalışmasında [44] önerilen modelin MS COCO veri kümesinin İngilizce altyazılar için eğitilmiş durumunda CIDEr sonucunun 0,8 civarında olduğunu belirtilmiş, Türkçe üzerine yaptığı çalışmada ince ayar yapılarak çalıştırdıkları modelden 0,450 sonucunu almışlardır. Bizim çalışmamızda ise, ince ayar yapılmış modelin (Model-2) CIDEr sonucu 0,528'dir. Diğer Türkçe çalışmalar [44, 45] ile karşılaştırma yapıldığında, diğer bir sonuç ise önerilen bu çalışmanın sonunda Türkçe görüntü altyazı veri kümesinin oluşturulmasıdır. Türkçe'de görüntü altyazısı ile ilgili çalışmaların sayısı hala çok sınırlıdır. Bu sebepten, önerilen bu çalışma, literatüre olumlu katkı sağlaması açısından da önemlidir. Çalışma sonucunda ayrıca Şekil 7'de görüldüğü üzere bir Web uygulaması hazırlanmıştır. Her bir görüntü için kullanıcının en fazla 5 altyazı gireceği bir ortam sunulmuştur. Bu altyazılar, kullanıcı tarafından sisteme girildikten sonra, uygunluğu, doğruluğu kontrol edilmek üzere uygulama yöneticisine yönlendirilmektedir. Uygunluk durumunda, o görüntüye ait altyazı oluşturularak, sonraki çalışmalarda karşılaştırma yapabilmek adına bir veri kümesinin oluşturulması sağlanmıştır. Böylelikle araştırmacıların karşılaştırma yapabilecek ve farklı problemlerde kullanılacak, herkesin kullanımına açık Türkçe veri kümesi oluşturulmaya başlanmıştır.

Sample Captions to be Ground Truthed for Turkish-Supported Show and Tell Model


WARNING: Please describe the images as accurate as possible with only one sentence.

UYARI: Lütfen imajları mümkün olduğunca alakalı bir formatta ve yalnızca tek bir cümleyle açıklayın.

	Bir tekne bir iskelede park etmiş.
	Karda oturan bir grup insan.

ÖRNEK

1. Lütfen verilen imajlara yukarıda sağlanan örnekteki gibi altyazılar üretin.
2. Altyazılarınızın kapsayıcı olmasına özen gösterin.
3. İmla kurallarına ve noktalama işaretlerine dikkat etmeyi unutmayın.

Number	Photo	Caption
1		

Şekil 6. Gerçek referans değeri form görüntüsü (View of the form for ground truth)



1	<input type="text" value="Bir kopek demirliklerin arkasında duruyor."/>
2	<input type="text" value="Kopek sokagi gozluyor."/>
3	<input type="text" value="Pencerenin yanındaki kopek arabalara bakıyor."/>
4	<input type="text" value="Siyah kopek demirliklere yaslanıyor."/>
5	<input type="text" value="Bir sokakta arabaları izleyen bir kopek var."/>

Gönder

Şekil 7. Web uygulaması görüntüsü (Image of the Web application)

4. SİMGELER (SYMBOLS)

ESA	: Evrişimsel Sinir Ağları
TSA	: Tekrarlayan Sinir Ağları
UKSB	: Uzun Kısa Süreli Bellek
GTB	: Geçitli Tekrarlayan Birim
SGB	: Sinirsel Görüntü Başlığı
UDESİA	: Uzun Dönemli Tekrarlayan Evrişimsel Ağ
ÜÇA	: Üremsel Çekişmeli Ağ
API	: Uygulama Program Arayüzü
MS COCO	: Microsoft Common Objects in Context
I	: MS COCO veri tabandan bir görüntü
S	: I görüntüyü tanımlayan bir Türkçe cümledir
AMT	: Amazon Mechanical Turk
Model-1	: Birinci model
Model-2	: İkinci model
BÇK	: Byte Çifti Kodlama

5. SONUÇLAR (CONCLUSIONS)

Bu çalışma, görüntü altyazısı problemine çözüm getirmek amacı ile mevcut yaklaşımları incelemiş, güncel olmayan yöntemleri eleyerek, en modern yöntemleri benimsemiştir. Literatürde kullanılan veri kümelerini ve değerlendirme ölçütlerini incelemenin yanı sıra, görüntülemeye ilişkin son teknoloji çözümlerin kapsamlı ve ayrıntılı bir görünümü sunulmuştur. Görüntü altyazısı oluşturma alanındaki son çalışmalar, sinir ağları tabanlı yöntemlerin diğerlerine göre daha üstün olduğunu göstermektedir. Tüm bunların ışığında, Vinyals vd. tarafından öne sürülen model, Türkçe için MS COCO veri kümesi kullanılarak geliştirilmiştir.

İnsan temelli ve otomatik olmak üzere iki farklı değerlendirme yapılmıştır. Sonuçlar önerilen modelin ürettiği skorların tatmin edici olduğunu göstermiştir. Buna ek olarak, Türkçe görüntü altyazısı geliştirmek adına bir Web uygulama (<http://mscoco-contributor.herokuapp.com/website/>) desteği sağlanmıştır. Böylelikle, daha başarılı bir altyazı modelinin elde edilmesini mümkün kılmak için kitle kaynaklı veri kümesini iyileştiren bir web platformu sunarken, tatmin edici bir Türkçe görüntü altyazısı modeli ve Türkçe altyazısı veri kümesi elde edilmiştir.

KAYNAKLAR (REFERENCES)

1. Yang, Y., Teo, C.L., Daume, H., Aloimono, Y., Corpus-Guided Sentence Generation of Natural Images, Conference on Empirical Methods in Natural Language Processing, Edinburgh - United Kingdom, 444-454, 27 - 31 Temmuz, 2011.
2. Mitchell, M., Dodge, J., Goyal, A., Yamaguchi, K., Stratos, K., Han, X., Mensch, A., Berg, A. Berg, H., Daume, H., Generating Image Descriptions from Computer Vision Detections, 13th Conference of the European Chapter of the Association for Computational Linguistics, Avignon - France, 747-756, Nisan, 2012.
3. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L., Baby talk: Understanding and Generating Simple Image Descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (12), 2891-2903, 2013.
4. Ushiku, Y., Yamaguchi, M., Mukuta, Y., Harada, T., Common Subspace for Model and Similarity: Phrase Learning for Caption Generation from Images, IEEE International Conference on Computer Vision, Washington DC - USA, 2668-2676, 07-13 Aralık, 2015.
5. Ordonez, V., Kulkarni, G., Berg, T.L., Im2text: Describing Images Using 1 Million Captioned Photographs, Advances in Neural Information Processing Systems, 24, 1143-1151, 2011.
6. Gupta, A., Verma, Y., Jawahar, C.V., Choosing Linguistics over Vision to Describe Images, AAAI Conference on Artificial Intelligence, Toronto - Canada, 606-612, 22-26 Temmuz, 2012.
7. Farhadi, A. ve Sadeghi, M.A., Phrasal Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35 (12), 2854-2865, 2013.
8. Mason, R. ve Charniak, E., Nonparametric Method for Data-Driven Image Captioning, 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore - Maryland, 592-598, 22-27 Temmuz, 2014.
9. Kuznetsova, P., Ordonez, V., Berg, T., Choi, Y., Tree talk: Composition and Compression of Trees for Image Descriptions, Transaction of Association for Computational Linguistics, 2 (10), 351-362, 2014.
10. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., Gradient-based learning applied to document recognition, Proceedings of the IEEE, 86 (11), 2278-2324, 1998.
11. Yıldız O., Melanoma detection from dermoscopy images with deep learning methods: A comprehensive study, Journal of the Faculty of Engineering and Architecture of Gazi University, 34 (4), 2241-2260, 2019.
12. Hanbay K., Hyperspectral image classification using convolutional neural network and twodimensional complex Gabor transform, Journal of the Faculty of Engineering and Architecture of Gazi University, 35 (1), 443-456, 2019.
13. Elman, J.L., Finding structure in time, Cognitive Science, 14 (2), 179-212, 1990.
14. Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., A neural probabilistic language model, J. Mach. Learn. Res., 3, 1137-1155, 2003.
15. Kalchbrenner, N. ve Blunsom, P., Two Recurrent Continuous Translation Models, ACL Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Seattle - USA, 1700-1709, 18-21 Ekim, 2013.
16. Cho, K., van Merriënboer, B., Gülçehre, Ç., Bougares, F., Schwenk, H., Bengio, Y., Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, CoRR, abs/1406.1078, 2014.
17. Sutskever, I., Vinyals, O., Le, Q.V., Sequence to Sequence Learning with Neural Networks, 27th International Conference on Neural Information Processing Systems (NIPS'14), Editör: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D. ve Weinberger,

- K.Q, MIT Press, Cambridge, MA, USA, 2, 3104-3112, 2014.
18. Vinyals, O., Alexander Toshev, A., Bengio, S., Erhan, D., Show and Tell: A Neural Image Caption Generator, CoRR, 2014.
 19. Hochreiter, S. ve Schmidhuber, J., Long Short-Term Memory, *Neural Computation*, 9 (8), 1735–1780, 1997.
 20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L., Microsoft COCO: Common Objects in Context, *Computer Vision*, Springer International Publishing, ECCV 2014, Zurich - Switzerland, 740-755, 6-12, Eylül, 2014.
 21. Kiros, R., Salakhutdinov, R., Zemel, R., Multimodal Neural Language Models, 31st International Conference on Machine Learning, *Proceedings of Machine Learning Research (PMLR)*, 32 (2), 595-603, 2014.
 22. Kiros, R., Salakhutdinov, R., Zemel, R.S., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, CoRR, abs/1411.2539, 2014.
 23. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.L., Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN), 3rd International Conference on Learning Representations (ICLR), San Diego - CA - USA, 7-9 Mayıs, 2015.
 24. 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.
 25. Young, P., Lai, A., Hodosh, M., Hockenmaier, J., From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions, *TACL*, 2, 67-78, 2014.
 26. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A., Grounded Compositional Semantics for Finding and Describing Images with Sentences, *Transactions of the Association for Computational Linguistics*, 2, 207-218, 2014.
 27. Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description, *IEEE Conference on Computer Vision and Pattern Recognition*, Boston - MA, 2625-2634, 7-12 Haziran, 2015.
 28. Karpathy, A. ve Fei-Fei, L., Deep Visual-Semantic Alignments for Generating Image Descriptions, *IEEE Trans. Pattern Anal. Mach. Intell.*, 39 (4), 664-676, Nisan 2017.
 29. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T., Guiding the Long-Short Term Memory Model for Image Caption Generation, *IEEE International Conference on Computer Vision*, Santiago - Chile, 2407-2415, 13-16 Aralık, 2015.
 30. Yang, Z., Yuan, Y., Wu, Y., Cohen, W.W., Salakhutdinov, R.R., Review Networks for Caption Generation, *Advances in Neural Information Processing Systems 29 (NIPS2016_6167)*, Editör: Lee D.D., Sugiyama, M., Luxburg, U.V., Guyon, I. ve Garnett, R., 2361-2369, 2016.
 31. Xu, K., Lei Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y., Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, *32nd International Conference on Machine Learning - Volume 37 (ICML'15)*, 37, Editör: Bach, F. ve David Blei, D, JMLR.org, 2048-2057, 2015.
 32. Park, C.C., Kim, B.G., Kim, G., Attend to You: Personalized Image Captioning with Context Sequence Memory Networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu - HI, 6432-6440, 2017.
 33. Tavakoli, H.R., Shetty, R., Borji, A., Laaksonen, J., Paying Attention to Descriptions Generated by Image Captioning Models, *IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii - United States, 2506-2515, 21- 26 Temmuz, 2017.
 34. Liu, C., Mao, J., Sha, F., Yuille, A.L., Attention Correctness in Neural Image Captioning, 31st AAAI Conference on Artificial Intelligence (AAAI'17), AAAI Press, California, USA, 4176-4182, 4-9 Şubat, 2017.
 35. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Chua, T.S., SCA-CNN: Spatial and Channel-wise Attention in Convolutional Networks for Image Captioning, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii - United States, 6298-6306, 21- 26 Temmuz, 2017.
 36. Lu, J., Xiong, C., Parikh, D., Socher, R., Knowing When to Look: Adaptive Attention via Avisual Sentinel for Image Captioning, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Hawaii - United States, 3242-3250, 21- 26 Temmuz, 2017.
 37. 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, Nevada-ABD, 4651-4659, 26 Haziran - 1 Temmuz, 2016.
 38. Yao, T., Pan, Y., Li, Y., Qiu, Z., Tao Mei, T., Boosting Image Captioning with Attributes, *IEEE International Conference on Computer Vision (ICCV)*, Venice - Italy, 4904-4912, 22 - 29 Ekim, 2017.
 39. Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M., Schiele, B., Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training, *IEEE International Conference on Computer Vision (ICCV)*, Venice - Italy, 4155-4164, 2017.
 40. Dai, B., Lin, D., Urtasun, R., Fidler, S., Towards Diverse and Natural Image Descriptions via a Conditional GAN, *IEEE conference on computer vision and pattern recognition (CVPR)*, Hawaii - United States, 2989-2998, 21- 26 Temmuz, 2017.
 41. Aneja, J., Deshpande, A., Schwing, A.G., Convolutional Image Captioning, *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City - UT, 5561-5570, 12-18 Haziran, 2018.
 42. Wang, Q. ve Chan, A.B., {CNN+CNN:} Convolutional Decoders for Image Captioning, CoRR, abs/1805.09019, 2018.
 43. Ünal, M.E., Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N.I., Cakici, R., TasvirEt: A Benchmark Dataset for Automatic Turkish Description Generation from Images, *24th Signal Processing and*

- Communication Application Conference (SIU), Zonguldak-Turkey, 16-19 Mayıs, 2016.
44. Samet, N., Hiçsönmez, S., Duygulu, P., Akbas, E., Could we Create A Training Set For Image Captioning Using Automatic Translation? 25th Signal Processing and Communications Applications Conference (SIU), Antalya-TR, 15-18 Mayıs, 2017.
 45. Kuyu, M., Erdem, A., Erdem, E., Image Captioning in Turkish with Subword Units, 26th Signal Processing and Communications Applications Conference (SIU), Izmir-TR, 2-5 Mayıs, 2018.
 46. Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., Erdem, E., Re-evaluating automatic metrics for image captioning, Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Long Papers, Valencia, Spain, 1, 199-209, 2017.
 47. Yüksek Y., Karasulu B., A review on semantic video analysis using multimedia ontologies, Journal of the Faculty of Engineering and Architecture of Gazi University, 25 (4), 719-739, 2010.