Altsözcük Öğeleri ile Türkçe Görüntü Altyazılama
Image Captioning in Turkish with Subword Units

Menekşe Kuyu, Aykut Erdem, Erkut Erdem
Bilgisayar Mühendisliği Bölümü, Hacettepe Üniversitesi, Ankara, Türkiye
menekse.kuyu@hacettepe.edu.tr, {aykut,erkut}@cs.hacettepe.edu.tr

***Özetçe* —Görüntü altyazılama olarak da bilinen görüntülerin doğal cümlelerle açıklamalarının otomatik olarak üretilmesi, bilgisayarla görme ve doğal dil işlemenin kesişiminde yer alan ve son zamanlarda literatürde oldukça ilgi görmeye başlamış zorlu bir araştırma problemidir. Derin öğrenme alanında yaşanan gelişmelerle birlikte, görüntü altyazılama için önerilmiş yakın tarihli yaklaşımların tamamı derin yapay sinir ağlarına dayanmaktadır. Ancak bu yöntemlerin çoğu İngilizce dili üzerine odaklanmıştır ve bu durum Türkçe için kullanımlarını büyük ölçüde kısıtlamaktadır. Türkçe sondan eklemeli bir dil olduğu ve sözcüklere eklenen her ek sözcüğün anlamını değiştirebildiği için Türkçe'ye özgü geliştirilecek bir altyazılama yaklaşımının dilin bu özelliklerini göz önüne alması gerekmektedir. Bu çalışmamızda, bu eksikliği kapatmak adına, kelimeler yerine altsözcük öğeleri kullanan bu tarz bir altyazılama modeli önerilmektedir. Deneysel sonuçlarımız, bu modelin sözcük tabanlı modelden çok daha iyi sonuç verdiğini göstermektedir.**

***Anahtar Kelimeler*—Görüntü altyazılama, bütünleştirilmiş görme ve dil**

***Abstract*—Automatically describing images with natural sentences, also known as image captioning, is a challenging research problem at the intersection of computer vision and natural language processing which has recently become very popular in the literature. With the advances in deep learning, recently proposed image captioning approaches are all based on deep artificial neural networks. However, most of these methods focus on the English language, which greatly restricts their use for Turkish. Turkish is an agglutinative language and suffixes might change the meaning of a word entirely, hence an image captioning approach specifically designed for Turkish should consider the characteristics of the language. In this study, we propose such an image captioning model, which utilizes subword units. Our experimental results show that this model provides results which are much better than the word-based model.**

***Keywords*—Image captioning, integrated vision and language**

I. Giriş

Son yıllarda oldukça popülerlik kazanan bir bütünleşik görme ve dil problemi olan görüntü altyazılamada amaç, verilen bir görüntünün doğal dilde bir açıklamasını otomatik olarak üretilmesidir [1]. Literatürde yakın tarihte önerilmiş başarılı görüntü altyazılama yaklaşımlarının tamamına yakını derin öğrenmeye dayalıdır ve bu modeller basitçe görüntü içeriğini bir evrişimsel sinir ağı (*convolutional neural networks*) ile kodladıktan sonra ilgili açıklamayı bir dil modeline karşılık gelen bir yinelemeli sinir ağı (*recurrent neural networks - RNN*) kullanarak üretmektedirler. Bu yöndeki mevcut çalışmalar incelendiğinde bu yöntemlerin ağırlıkla İngilizce'ye yoğunlaştıkları görülmektedir ve dolayısıyla İngilizce'den yapısal farklılık gösteren diğer doğal diller için bu çalışmaların ne derecede uygun oldukları bir soru olarak karşımızda durmaktadır.

Türkçe görüntü altyazılama için önerilmiş olan ilk veri kümesi, kitlekaynak yaklaşımı izlenerek oluşturulan TasvirEt veri kümesidir [2]. Bu veri kümesinde, daha önce İngilizce için oluşturulmuş Flickr8k veri kümesindeki [4] tüm görüntüler için Türkçe altyazılar toplanmış durumdadır. TasvirEt veri kümesi, toplam 8 bin görüntü ve her görüntü için 2 altyazıdan oluşmaktadır. Türkçe altyazılama için kullanılabilecek ikinci bir veri kümesi [3] tarafından geçtiğimiz yıl önerilmiştir. Yazarlar, bu çalışmalarında MS-COCO [5] veri kümesini Türkçe açıklamalar ile zenginleştirme yoluna gitmişlerdir. Burada Türkçe kavramsal açıklamaları kitlekaynak yaklaşımı ile toplamak yerine mevcut olan İngilizce açıklamaları otomatik bir tercüme aracı (Google Translate) kullanarak Türkçe'ye çevirmişlerdir. TasvirEt verikümesine oranla daha büyük hacimli bir verikümesi oluşturulmuş olsa da otomatik tercüme ile elde edilen Türkçe açıklamaların kullanılan tercüme sisteminin getirdiği dilbiçimsel ve anlamsal açıdan hayli gürültülü açıklamalar içerebilmektedir. Bu nedenle geliştirilecek olan derin modellerin eğitiminin kötü yönde etkilenme potansiyeli mevcuttur.

Türkçe dil olarak sondan eklemeli bir dildir ve eklenen her ek eklendiği sözcüğün anlamını değiştirmektedir. Bu nedenle İngilizce için önerilmiş olan yaklaşımların Türkçe altyazılama için doğrudan kullanım olanakları çok kısıtlıdır. Bu durum, mevcut görüntü altyazılama modellerinin farklı bir gözle yeniden değerlendirilmesini ve görsel veriden açıklamalar yaratılırken eklenen eklerle değişen anlamları dikkate alacak şekilde Türkçe'ye özgü olarak geliştirilmeleri gerekliliğini doğurmaktadır. Bu çalışmamızda, Türkçe eğitim verisindeki sözcüklerden, n-gram istatistiklerilerine bağlı olarak Byte Pair Encoding (BPE) algoritması [6], [7] kullanılarak oluşturulan altsözcüklere dayalı, Türkçeye özel bir görüntü altyazılama modeli önerilmektedir.

II. Altsözcük Modeli

Daha önce belirttiğimiz üzere Türkçe sondan eklemeli bir dildir ve bundan ötürü kullanılan farklı ekler üzerinden istenildiği kadar farklı sözcük türetmek mümkündür. İngilizce ile kıyaslandığında Türkçe için önerilecek nöral görüntü altyazılama modelleri için üstesinden gelinmesi gereken ciddi zorluklar bulunmaktadır. Öncelikle Türkçe bir sözlük İngilizce'ye oranla çok daha fazla sayıda kelime içermekte ve bu modellerin bellek kullanımını ve çalışma süresini arttırmaktadır. Yine bu durum ile ilişkili bir diğer güçlük literatürde seyrek kelime

978-1-5386-1501-0/18/$31.00

problemi olarak geçen ve içinde eğitim kümesinde çok az geçen kelimeleri barındıran açıklamaların öğrenimi kötü yönde etkilemesi sorunudur. Burada çözüm olarak sözlüğün sadece sık geçen kelimelerden oluşturulması yoluna gidilmektedir. Türkçe özelinde böyle bir yaklaşım birçok kelimenin açıklamalarda kullanılamayacak olması demektir. Aşağıda bu iki soruna basit ve doğal bir çözüm getiren altsözük modelinin detayları anlatılmaktadır.

Eğitim kümesindeki sözcükler ilk olarak n-gram istatistikleri kullanılarak Pair Encoding (BPE) algoritması [6], [7] kullanılarak alt sözcüklere ayrıştırılmakta ve bu alt sözcüklere dayalı sözlük temsilleri kullanılmaktadır. Bu algoritmanın diğer kodlama algoritmalarından en büyük farkı, sözcüklerden oluşturulan değişken uzunluklu karakter dizilerinin hala alt sözcük birimleri (subword units) olarak yorumlanabilmesidir. Bu sözcük birimleri kullanılarak, bir dil modelinin eğitim aşamasında karşılaşmadığı yeni sözcükler üretebilmesi sağlanabilmektedir. BPE yönteminde ilk olarak karakter alfabesi dikkate alınarak bir sembol sözlüğü elde edilmektedir. Bu sayede her sözcük, sembol adı verilen farklı uzunluklardaki karakter dizinleriyle ifade edilebilmektedir. Sözcük sonunu belirlemek için "</w>", altsözcüklerin sonunu belirlemek için ise "@@" özel sembolü kullanılmaktadır. Sözcük ve altsözcük sonunu belirlemenin ana nedeni, daha sonra bir dil modeli kullanılarak bir araya getirilecek alt sözcüklerden bir sözcüğün elde edilmesini kolaylaştırmaktır.

BPE modeli, eğitim kümesinde yer alan sözcüklerden alt sözcük kümesi oluşturma işleminde yinelemeli bir yöntem kullanmaktadır. Eğitim kümesindeki sözcükler öncelikle, bir sözcüğün en küçük parçası olan karakterlerine ayrıştırılmaktadır. Ayrıştırılan her bir karakter, bir sembol olarak düşünülerek veri kümesinde ikili karakterlerin yan yana geçme sıklığı hesaplanmakta ve bunlar arasında en sık geçen iki karakter birleştirilerek, iki karakter uzunluğunda bir sembol yaratılmaktadır. Bu birleştirme işlemi, yinelemeli olarak tüm semboller için belirli sıklıkta ve belirli sayıda alt sözcükler elde edilene kadar tekrarlanmaktadır. Bu işlemin ana amacı, eğitim kümesinde en sık görülen karakter n-gramlarını en sonunda tek bir sembole veya bir başka deyişle bir altsözcüğe dönüştürebilmektir.

BPE modelinin işleyişini bir örnekle açıklamak gerekirse, altsözcükleri çıkartmak istediğimiz veri kümesinin; "*Bir adam duruyor.*", "*Biri su içiyor.*" ve "*Biri yolda duruyor.*" cümlelerinden oluştuğunu varsayalım. Bu veriden yukarıdaki yöntem kullanılarak öğrenilen örnek altsözcük dizinleri ve sıklıkları Tablo 1'de gösterilmektedir. Bu tablodan da görülebileceği üzere sonuçta elde edilen altsözcükler eğitim verisinde sıklıkla görülen karakter n-gram'larını yansıtmaktadır. Bu bakımdan Türkçenin dilbilimsel özellikleri açısından bu altsözcükler doğrudan hecelere karşılık gelmeyebilmektedir.

Bu çalışmada, BPE modelini Türkçe sözcüklerin altsözcüklerine ayrıştırılmasının öğrenilmesinde www.tr.wikipedia.org sayfasından toplanan Türkçe metinlerin bulunduğu bir veri kümesi kullanılmıştır. Bu veri kümesinden toplam 30 bin farklı altsözcük öğrenilmiştir ve öğrenilen alt sözcük birimleri, görüntü betimlemelerini ayrıştırmak için kullanılmıştır. Şekil 1'de örnek bir görüntü için mevcut olan betimlemelerin altsözcük birimleri kullanılarak nasıl ayrıştırıldıkları gösterilmektedir. Bu çalışmada Türkçe eğitim verisinden elde edilen bu tarz altsözcük birimlerinden oluşan bir sözlüğe bağlı olarak Türkçeye özel bir dil modelinin eğitimi sağlanmaktadır. Görüntü altyazılama için eğitilen böyle bir dil modeli girdi görüntüsünü bir şart olarak ele alarak çalışmaktadır.

TABLO I: "*Bir adam duruyor.*", "*Biri su içiyor.*" ve "*Biri yolda duruyor.*" cümlelerinden BPE modeli kullanılarak çıkartılan altsözcük örnekleri. "</w>" işareti kelime bitişini ifade etmektedir.

Altsözcük dizini	Sıklık	Altsözcük dizini	Sıklık
`yo`	3	`ur`	2
`yor`	3	`uruyor</w>`	2
`yor.</w>`	3	`ri</w>`	3
`Bi`	3	`duruyor.</w>`	2
`uyor.</w>`	2	`Biri</w>`	3

BPE modeli uygulanmadan önceki orjinal altyazı:
`Karlarla kaplı bir dağda bir grup insan yürüyor.`

BPE modeli uygulandıktan sonraki altyazı:
`Kar larla kaplı bir dağ da bir grup insan yürüy or.`

Şekil 1: BPE modeli kullanılarak elde edilen sözcük ayrıştırma işlemi için örnek bir sonuç.

III. Görüntü Altyazılama İçin Yinelemeli Sinir Ağ Modeli

Bu çalışmada, görüntü altyazılama için önerilen, kodlayıcı-kod çözücü tabanlı Uzun Kısa Süreli Bellek (*Long Short-Term Memory* - LSTM) mimarisi [8] kullanılmıştır. Geliştirilen modelde, öncelikle görüntü içeriğini belirlemek adına evrişimsel sinir ağları ile girdi görüntüsünün anlamsal bir gösterimi oluşturulmakta; ardından da LSTM tabanlı bir dil modeli kullanmaktadırlar. LSTM mimarisi, RNN'lerin uzun süreli bağlılıkları yakalamak için varolan sinir ağı yapısına ek olarak, bellek hücresi (*memory cell*) vektörü ile güçlendirilmesine dayanmaktadır. LSTM, her bir zaman adımında girdi olarak x_t, h_{t-1}, c_{t-1} vektörlerini almakta ve çıktı olarak h_t, c_t vektörleri, aşağıda verilen formüllerle üretilmektedir:

$$i_t = \sigma(W^i x_t + U^i h_{t-1} + b^i) \quad (1)$$
$$f_t = \sigma(W^f x_t + U^f h_{t-1} + b^f) \quad (2)$$
$$o_t = \sigma(W^o x_t + U^o h_{t-1} + b^o) \quad (3)$$
$$g_t = tanh(\sigma(W^g x_t + U^g h_{t-1} + b^g)) \quad (4)$$
$$c_t = f_t \cdot x_t + i_t \cdot g_t \quad (5)$$
$$h_t = o_t \cdot tanh(c_t) \quad (6)$$

Burada i_t, f_t, o_t sırasıyla girdi, unut ve çıktı kapılarına karşılık gelmektedir. Uzun-Kısa Süreli Bellek modelinin temeli, her adımda gözlemlenen girdileri kodlayan hafıza hücresidir. Bu hafıza hücresine bilgi yazılıp silinmesi, "kapı" adı verilen yapılar tarafından düzenli olarak kontrol edilmektedir. Öyle ki

Kapı 1 değerini aldığında, veri hafıza hücresinde tutulur, 0 değerini aldığında ise hafıza hücresinden silinir. Bu modelin toplam 3 kapısı bulunmaktadır; girdi kapısı i_t (input gate), çıktı kapısı o_t (output gate) ve unutulma kapısı f_t (forget gate). Bu sayede, model hem bir önceki girdiyi, hem de şimdiki girdiyi dikkate alarak karar vermektedir.

Şaşırtıcı seviyede başarılı sonuçlar üretebilen bu modellerin en büyük eksikliği, sabit bir sözlüğe dayalı olarak sözcük seviyesinde tahmin yapmalarıdır. Bu yüzden daha önce karşılaşılmamış, sözlüklerinde bulunmayan bir sözcüğü (out-of-dictionary word) bir açıklamada kullanma şansları yoktur. Bu durum, özellikle Türkçe gibi sondan ekleme bakımından zengin dillerde ciddi problemler çıkarabilmektedir zira bu özelliklere sahip bir dil tam anlamıyla modellenmek istendiğinde sözlük boyutu işin içinden çıkılmaz bir hale gelmektedir. Bu bağlamda, projemizde varolan görüntü altyazılama yaklaşımlarının doğrudan uyarlanması yapılmayarak, yani Türkçedeki mevcut olan ve eklerle genişletilmiş sözcükler kullanmak yerine,bunun yerine literatürde yakın zamanda karşımıza çıkmaya başlayan ve bir önceki bölümde özetlediğimiz altsözcüklere bir başka deyişle karakter tabanlı ngram'lara dayalı, Türkçeye özel bir görüntü altyazılama modeli geliştirilmiştir.

Önerilen derin öğrenme modelini eğitmek için, [9] çalışmasında önerilmiş olan görüntü altyazılama modelinin açık kaynak kodlu bir gerçekleştiriminden yararlanılmıştır. Modeli eğitmeye başlamadan önce, önişlem olarak veri kümesindeki altyazılar BPE sözcük kodlama modeli ile altsözcüklerine ayrıştırılmış, ve bu altyazılara dayalı bir sözlük oluşturulmuştur. Daha sonra görüntüler ve altsözcüklerine ayrıştırılmış altyazılar yinelemeli derin ağ modeline beslenmiştir. Böylelikle geliştirilmiş olan derin ağ modeli her bir adımda yeni bir altsözcük üretmektedir. Test aşamasında üretilen çıktılar, aynı BPE sözcük kodlama modeli kullanılarak ardışık altsözcüklerin belirlenmesinde kullanılmış olup çıktı cümlesi bu altsözcüklerin bütünleştirilmesiyle elde edilmektedir.

IV. Deneysel Sonuçlar

Altsözcük tabanlı derin öğrenme modelinin eğitim aşamasında, TasvirEt [2], MS-COCO [5] ve Flickr30k [10] veri kümeleri kullanılmıştır. İlk olarak, eğitim kümesindeki altyazılara altsözcük modeli uygulanmadan, ham altyazılarla sözcük tabanlı bir derin öğrenme modeli eğitilmiştir. Daha sonraki aşamada, eğitim kümesindeki altyazılar altsözcüklerine ayrıştırılarak altsözcüklerden oluşan bir sözlük oluşturulmuştur ve derin öğrenme modeli, altsözcüklere dayalı olarak eğitilmiştir.

Sözcük ve altsözcük modelleri eğitilirken öncelikli olarak MS-COCO ve Flickr30k veri kümeleri kullanılmıştır. MS-COCO veri kümesi, Türkçe için bu zamana kadar hazırlanmış en fazla görüntü ve altyazı içeren veri kümesidir. Eğitim kümesi ortalama 80.000 görüntü içermektedir ve her görüntü için beş adet açıklama bulunmaktadır. MS-COCO çok büyük hacimli bir veri kümesi olmasına karşın, otomatik tercüme kullanılarak oluşturulmasından dolayı gürültülü altyazılar içerebilmektedir. Otomatik tercüme yöntemiyle Türkçe için oluşturulan bir diğer veri kümesi ise Flickr30k'dır ve toplamda 30000 resim ve her resim için 5 adet açıklama içermektedir. MS-COCO veri kümesinde yaşanan gürültülü altyazı problemi, bu veri kümesinde de mevcuttur. Bu sebeple, MS-COCO ve Flickr30k ile eğitilen derin öğrenme modeli, insanlar tarafından

Orijinal altyazılar:

- *Asker kıyafetiyle motor sporları yapan motorcu.*
- *Asker kıyafetli bir adam motoru ile taşlarda ilerliyor.*

Sözcük modeli ile tahmin edilen altyazı:
Bir adam bir bisikletin yanında bir bankta oturuyor.

Altsözcük modeli ile tahmin edilen altyazı:
Bir motosikletin arkasına binen bir adam.

Şekil 2: Flickr8k veri kümesinden alınmış bir örnek görüntüye ait orjinal ve tahmin edilen açıklamalar.

üretilen altyazılar içeren TasvirEt eğitim kümesi üzerinde ince ayara tabi tutulmuştur.

Modelin test aşamasında, TasvirEt test kümesi ve MS-COCO doğrulama kümesinden 500 görüntü kullanılmıştır. Her iki test kümesi de insan tarafından oluşturulmuş altyazılar içermektedir. Başarım ölçümü için ilk olarak, görüntü altyazılamada sıklıkla kullanılan BLEU, METEOR, Rouge-L ve CIDEr kullanılmıştır. Bu metriklere ek olarak, [12]çalışmasında görüntü altyazılama problemi için kullanılması önerilen, cümleler arasındaki anlamsal benzerlikleri baz alan Word Movers Distance (WMD) [11] başarım metriği ayrıca kullanılmıştır.

Şekil 2'de TasvirEt test kümesinden örnek bir görüntüye ait orijinal altyazı ve eğitilen iki farklı model için tahmin edilen altyazılar gösterilmiştir. Üretilen altyazılar incelendiğinde, altsözcük modelinin sözcük modeline göre anlamsal olarak orijinal altyazıya daha yakın sonuç ürettiği söylenebilir. Ancak BLEU gibi n-gram tabanlı metriklerde, yukarıda bahsedilen anlamsal yakınlık bilgisi ölçülememektedir [12]. Bu sebeple WMD metriği de görüntü altyazılama problemi için başarı ölçümü olarak kullanılmıştır. Elde edilen sayısal deney sonuçları Tablo 2'de gösterilmiştir.

Veri kümesinden bağımsız olarak, önerilen altsözcük modelinin standart sözcük tabanlı modelden BLEU-2, BLEU-3, BLEU-4, CIDEr ve WMD metriklerinde daha başarılı olduğu gözlenmiştir. Bu sonuca göre altsözcük modeli kullanılarak elde edilen altyazıların dil bilgisi açısından diğer modellerden daha başarılı olduğu söylenebilir. MS-COCO ile eğitilip TasvirEt eğitim kümesi üzerinde ince ayara (finetuning) tabi tutulan altsözcük bazlı model, MS-COCO doğrulaması kümesindeki 500 görüntü için en başarılı sonuçları vermektedir. Daha önce belirttiğimiz üzere Türkçe MS-COCO veri kümesi otomatik olarak oluşturulduğu için çoğu gürültülü ve Türkçe'nin dil bilgisi kurallarına uymayan açıklamalar içerebilmektedir. Bu nedenle, MS-COCO veri kümesinden öğrenilen modelin, insanlar tarafından hazırlanmış TasvirEt eğitim kümesi üzerinde ince ayar çekilerek elde edilen son hali, deneysel sonuçlara göre en başarılı deney konfigürasyonu olmuştur. Şekil 3'te altsözcük bazlı bu model kullanılarak elde edilen örnek sonuçlara yer verilmiştir. Yukarıda bahsedilen ince ayar stratejisi, Flickr30k veri kümesi ile eğitilen modeller için de uygulanmıştır. Fakat bu veri kümesi üzerinde alınan sonuçlar, MS-COCO ile kıyaslandığında daha az başarılıdır. Bunun ana sebebi, Flickr30k

TABLO II: Deneysel sonuçlar.

Model	Eğitim Kümesi	Test Kümesi	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	WMD
Sözcük	MS-COCO	MS-COCO	0.274	0.148	0.069	0.033	**0.147**	0.292	0.485	0.050
Altsözcük			**0.293**	**0.165**	**0.088**	**0.053**	**0.147**	**0.302**	**0.567**	**0.058**
Sözcük	Flickr30k		0.180	0.075	0.026	0.012	0.089	0.190	0.084	0.024
Altsözcük			0.215	0.089	0.036	0.019	0.104	0.220	0.148	0.031
Sözcük	MS-COCO	TasvirEt	0.299	0.147	0.057	0.019	0.105	0.247	0.120	0.029
Sözcük+İnce ayar			0.295	**0.166**	0.088	0.041	0.105	0.264	0.251	0.042
Altsözcük			0.280	0.137	0.053	0.021	0.096	0.234	0.115	0.031
Altsözcük+İnce ayar			0.263	0.163	**0.097**	**0.055**	0.103	**0.267**	**0.320**	**0.058**
Sözcük	Flickr30k		**0.308**	0.161	0.072	0.027	**0.114**	0.258	0.179	0.031
Sözcük+İnce ayar			0.291	0.160	0.075	0.031	0.100	0.261	0,211	0.034
Altsözcük			0.297	0.165	0.079	0.038	0.112	0.265	0.214	0.038
Altsözcük+İnce ayar			0.239	0.147	0.082	0.045	0.100	0.261	0.303	0.042

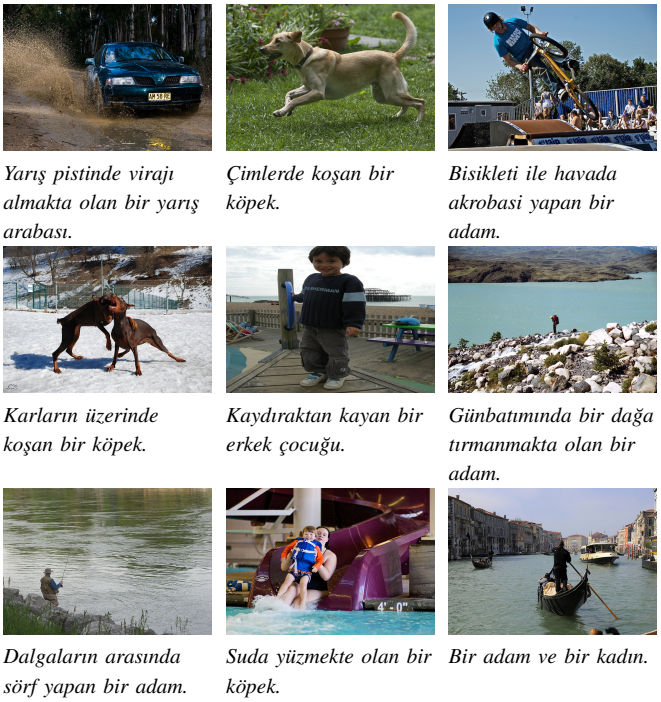
kümesinin MS-COCO'dan çok daha küçük hacimli olmasıdır.

V. Sonuç

Bu çalışmamızda Türkçe görüntü altyazılama problemi için altsözcük tabanlı bir model önerilmiştir. Önerilen altsözcük modeli, RNN tabanlı derin öğrenme modelinin eğitim aşamasında kullanılmıştır. Eğitim kümesi olarak otomatik oluşturulmuş MS-COCO ve Flickr30k veri kümeleri kullanılmıştır ve bu veri kümeleri ile eğitilen modeller daha sonra TasvirEt eğitim kümesi üzerinde ince ayara tabi tutulmuştur. Deneysel sonuçlar incelendiğinde, bahsedilen ince ayar stratejisinin TasvirEt kümesinde başarıyı artırdığı ve önerilen altsözcük modelinin BLEU-2, BLEU-3, BLEU-4, CIDEr ve WMD metriklerinde daha iyi sonuç verdiği gözlenmiştir. Bu metriklerdeki başarı artışı, altsözcük modelinin Türkçe'nin dilbilgisi kurallarına uygun, daha anlamlı altyazılar üretildiğini göstermektedir.

TEŞEKKÜR

Bu çalışma, Hacettepe Üniversitesi Bilimsel Araştırma Projeleri Koordinasyon Koordinasyon Birimince FBB-2016-11653 nolu proje kapsamında desteklenmiştir. NVIDIA firmasına sağladıkları GPU kartı için teşekkür ederiz.



Yarış pistinde virajı almakta olan bir yarış arabası. *Çimlerde koşan bir köpek.* *Bisikleti ile havada akrobasi yapan bir adam.*

Karların üzerinde koşan bir köpek. *Kaydıraktan kayan bir erkek çocuğu.* *Günbatımında bir dağa tırmanmakta olan bir adam.*

Dalgaların arasında sörf yapan bir adam. *Suda yüzmekte olan bir köpek.* *Bir adam ve bir kadın.*

Şekil 3: Altyazılama sonuçları. İlk satırda başarılı, ikinci satırda kısmen başarılı, üçüncü satırda başarısız sonuçlar gösterilmektedir.

Kaynaklar

[1] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank, "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures", Journal of Artificial Intelligence Research, Vol. 55, pp. 409-442, 2016.

[2] M. E. Unal, B. Citamak, S. Yagcioglu, A. Erdem, E. Erdem, N. I. Cinbis, and R. Cakici, "Tasviret: A benchmark dataset for automatic turkish description generation from images," in Proc. SIU, pp.1977-1980, 2016.

[3] N. Samet, S. Hiçsönmez, P. Duygulu and E. Akbaş, "Could we create a training set for image captioning using automatic translation?," in Proc. SIU, pp. 1-4, 2017.

[4] M. Hodosh, and J. Hockenmaier, "Sentence-based image description with scalable, explicit models", in Proc. CVPRW, 2013

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in Proc. ECCV, pp. 740–755, 2014.

[6] P. Gage, "A New Algorithm for Data Compression", C Users J., Vol. 12, No. 2, pp.23–38, 1994.

[7] R. Sennrich, B. Haddow and A. Birch, "Neural Machine Translation of Rare Words with Subword Units.", in Proc. ACL, pp 1715-1725, 2016.

[8] S. Hochreiter, and J. Schmidhuber, "Long Short-Term Memory", Neural Computation, Vol. 9. pp. 1735-1780, 1997.

[9] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," in Proc. CVPR, pp. 3156-3164, 2015.

[10] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions", Transactions of the Association for Computational Linguistics, Vol. 2, pp. 67–78, 2014.

[11] M. J. Kusner, Y. Sun, N. I.Kolkin, and K. Q. Weinberger, "From word embeddings to document distances" in Proc. ICML, pp. 957–966, 2015.

[12] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis and E. Erdem, "Re-evaluating Automatic Metrics for Image Captioning", in Proc. EACL, pp. 199–209, 2017.