Görüntü Altyazılama için Otomatik Tercümeyle Eğitim Kümesi Oluşturulabilir mi? Could We Create A Training Set For Image Captioning Using Automatic Translation?

Nermin Samet*, Samet Hiçsönmez†, Pınar Duygulu†, Emre Akbaş*
*Bilgisayar Mühendisliği, Orta Doğu Teknik Üniversitesi, Ankara, Türkiye
†Bilgisayar Mühendisliği, Hacettepe Üniversitesi, Ankara, Türkiye
Email: nermin.samet@metu.edu.tr, samethicsonmez@hacettepe.edu.tr, pinar@cs.hacettepe.edu.tr, emre@ceng.metu.edu.tr

***Özetçe* —Otomatik görüntü altyazılama, son yıllarda artan bir şekilde ilgi görmeye başlamış bir problemdir. Bu problem için geliştirilmiş birçok İngilizce veri kümesi olmasına rağmen, sadece bir Türkçe veri kümesi vardır ve bu veri kümesi İngilizce akranlarına göre çok daha küçüktür. Görüntü altyazılama için yeni bir veri kümesi oluşturmak oldukça masraflı ve zaman alıcı bir iştir. Bu çalışma, varolan büyük İngilizce veri kümelerinin mümkün olan en az çabayla Türkçe'ye kazandırılması yönünde atılmış bir ilk adımdır. İngilizce eğitim kümelerini otomatik bir tercüme aracı kullanarak Türkçe'ye çevirdik ve ortaya çıkan Türkçe altyazılarla bir görüntü altyazılama modeli eğittik. Yaptığımız deneyler, bu modelin Türkçe altyazılama problemi için şimdiye kadar alınmış en yüksek başarımı verdiğini gösterdi.**

Anahtar Kelimeler*—*Görüntü altyazılama, bilgisayarlı görü, makina çevrimi

***Abstract*—Automatic image captioning has received increasing attention in recent years. Although there are many English datasets developed for this problem, there is only one Turkish dataset and it is very small compared to its English counterparts. Creating a new dataset for image captioning is a very costly and time consuming task. This work is a first step towards transferring the available, large English datasets into Turkish. We translated English captioning datasets into Turkish by using an automated translation tool and we trained an image captioning model on the automatically obtained Turkish captions. Our experiments show that this model yields the best performance so far on Turkish captioning.**

Keywords*—*Image captioning, computer vision, machine translation

I. GİRİŞ

Günümüzde görsel ve metinsel verideki artışa paralel olarak derin öğrenme modelleri, hem bilgisayarlı görü alanında hem de doğal dil işleme alanında birçok farklı probleme uygulanmış olup oldukça başarılı sonuçlar vermiştir. Örneğin bilgisayarlı görünün en önemli problemlerinden olan görüntü sınıflandırma ve nesne algılama problemleri için evrişimsel sinir ağları artık standart yöntem olarak kullanılmaktadır [1], [2]. Benzer şekilde, doğal dil işleme alanının önemli problemlerinden makine çevrimi için yinelgeli sinir ağları kullanılmaktadır [3]. Tüm bunlara paralel olarak başta görüntü altyazılama olmak üzere metinsel ve görsel bilginin beraber kullanıldığı problemler önem kazanarak popüler olmuşlardır [4]–[8].

Görüntü altyazılama, verilen bir görüntü için otomatik olarak o görüntünün içeriğini açıklayan bir cümle ya da ifade üretme problemidir. Bu problemin çözümü, görüntü içeriğinin bir bilgisayarlı görü modeli ile tam olarak yakalanmasını ve bu içeriğin bir doğal dil işleme modeli tarafından hedef dilde doğru bir şekilde ifade edilmesini gerektirmektedir. Başarılı sonuçların alınması, eğitim kümesindeki görüntü açıklamalarının doğruluğuna, gürültü içermemesine ve görüntüyü yeterince iyi ifade ediyor olmasına doğrudan bağlıdır. Günümüzde görüntü altyazılama daha çok İngilizce üzerine yoğunlaşmış, Tablo I'de görüldüğü üzere görüntü altyazılamada kullanılmak üzere kontrollü olarak hazırlanan geniş veri kümeleri toplanmıştır. Ancak görüntülerin otomatik olarak altyazılanması evrensel bir problem ve ihtiyaç olmakla beraber, farklı diller için bu nitelikte ve kapsamda veri kümelerinin toplanması zaman ve insan gücü açısında oldukça zor ve maliyetlidir. Bu maliyeti bir örnekle somutlayabiliriz. Bu bildiride önerdiğimiz yöntemin başarımını değerlendirmek amacıyla, MSCOCO [9] doğrulama veri kümesinde yer alan 500 adet görüntünün her biri için 2 farklı Türkçe altyazı üretmemiz yaklaşık 10 adam-saat sürdü (kişi başı yaklaşık beş saat olmak üzere toplam iki kişi çalıştı). Bu veri kümesinde 120.000'den fazla görüntü olduğunu göz önüne alırsak ve her bir görüntü için, İngilizce veri kümesinde olduğu gibi, beş farklı Türkçe altyazı üretmeyi hedeflersek, gerekli olan iş gücü yaklaşık olarak 6000 adam-saat, yani 8.3 adam-ay olacaktır. Bunun yanında, hali hazırda kontrollü olarak altyazılanmış bu görüntülerin tekrardan farklı dillerde benzer şekilde altyazılanması, zaman ve iş gücü kaybı olacaktır. İngilizce için var olan bu altyazıların, otomatik olarak farklı dillere aktarılıp, kullanılabilmesi bu kayıpların önüne geçecektir.

Öte yandan derin öğrenme yöntemlerinin doğal dil işlemenin makine çevirimi (tercümesi) alanına uygulanmasıyla, Google Çeviri'nin son sürümünde kullanılan yöntem [3] aralarında Türkçe'nin de bulunduğu dokuz farklı dil için uygulanmış olup, tercüme doğruluğu açısından geleneksel yaklaşımları geride bırakmıştır [10].

Bu bildiriye konu olan çalışmamızın temel kalkış noktası şu soru oldu: Google Çeviri gibi bir otomatik tercüme aracını kullanarak İngilizce hazırlanmış olan büyük görüntü-altyazılama

978-1-5090-6494-6/17/$31.00

veri kümelerini Türkçe'ye kazandırmak mümkün olabilir mi? Yukarıda açıkladığımız gerekçeler ve makine çevirimi alanındaki gelişmeler [10] ışığında, bu sorunun olumlu yanıtını hipotezimiz olarak kabul ettik ve halihazırda İngilizce için mevcut olan görüntü altyazılarını, Google Çeviri ile Türkçe'ye çevirerek, görüntü altyazılama modellerini otomatik tercümeyle elde ettiğimiz yeni (Türkçe) altyazılarla baştan eğittik. Eğittiğimiz modelin başarımını ölçmek için kendi ürettiğimiz altyazılara ek olarak daha önce başka bir çalışmada [11] yine insanlar tarafından üretilen altyazıları kullandık.

Bu noktada, İngilizce'den Türkçe'ye otomatik olarak çevrilmiş altyazılarla bir model eğitmek yerine; İngilizce altyazılarla eğitilmiş bir modelin altyazılama çıktısını Google Çeviri ile Türkçe'ye çevirmenin daha pratik olacağı düşünülebilir. Ancak yaptığımız deneyler Türkçe ile eğitilen modelin, İngilizce ile eğitilip çıktısı Türkçe'ye çevrilen modelden daha başarılı olduğunu gösterdi.

Bu çalışmada, Google Çeviri'den elde edilen altyazıları Türkçe altyazılama problemini çözmede kullandık. Daha önce Türkçe altyazılama için yapılan tek çalışma olan TasvirEt isimli çalışmada [11], Ünal ve diğerleri iki farklı yöntem önermiş, Flickr8K [12] veri kümesindeki görüntüler için Türkçe altyazı veri kümesi hazırlamışlardır. Yaptığımız çeşitli deneylerde, önerdiğimiz yöntemin TasvirEt'te rapor edilen sonuçlardan daha iyi bir başarım elde ettiğini gösterdik (BLEU-1 skoru bazında $\%7$'lik bir artış). Böylece Türkçe altyazılama problemi için (bilgimiz dahilinde) şimdiye kadar elde edilmiş en yüksek başarımı elde ettik.

II. Görüntü Altyazılama İçin Derin Öğrenme Modeli

Bu çalışmamızda Google tarafından önerilen uçtanuca (*İng.* end-to-end) altyazılama modelini [13] kullandık. Kodlayıcı-kod çözücü (*İng.* encoder-decoder) tabanlı olan bu model ilk defa makine çevriminde kullanılmıştır [14]. Makine çevriminde alınan başarılı sonuçlar ile birlikte görüntü altyazılama problemine uygulanmış [13] ve burada da başarılı sonuçlar vermiştir.

Model, kodlayıcı olarak çalışan bir evrişimsel sinir ağından (*İng.* convolutional neural network) ve kod çözücü olarak görev yapan bir yinelgeli sinir ağından (*İng.* recurrent neural network) oluşmaktadır. Evrişimsel sinir ağı (ESA), girdi görüntülerini belirli boyuttaki bir vektöre dönüştürerek kodlayıcı görevini yerine getirmektedir. Derin evrişimsel sinir ağını, yinelgeli sinir ağı (YSA) yapısı takip etmekte ve ESA çıktısını girdi olarak alarak cümle oluşturmaktadır. Model temelde bir görüntü için, doğru altyazının olasılığını maksimize etmektedir:

$$\theta^* = \arg\max_{\theta} \sum_{(G,A)} \log p(A|G;\theta), \qquad (1)$$

burada θ model parametrelerini, G görüntüyü ve A da doğru altyazıyı tanımlamaktadır.

Modelde, ESA olarak en sondaki softmax katmanı çıkarılmış bir VGG16 [15] ağı kullandık. YSA olarak da Uzun Kısa Süreli Bellek (Long Short Term Memory) (UKSB) kullanılmıştır [16]. UKSB modeli, bir görüntüye ait ESA modeli ile çıkartılan öznitelikleri ve ilgili altyazıyı girdi olarak almakta ve negatif log likelihood hata fonksiyonunu altyazının her kelimesini tahmin edecek şekilde minimize etmektedir.

III. Deneysel Sonuçlar

Türkçe görüntü altyazılama problemi için mevcut tek veri kümesi TasvirEt [11] çalışmasında sunulmuştur. Bu veri kümesinde her görüntüye ait iki farklı altyazı mevcuttur. Ancak Bölüm II'de açıklanan yöntem için bu sayı yetersizdir ve Türkçe altyazıya sahip daha büyük veri kümelerinin kullanılması gerekmektedir. Bu problem MSCOCO [9] ve Flickr30k [17] veri kümeleri için Türkçe altyazılar hazırlanarak aşılabilir. Son zamanlarda Google Çeviri'nin başarılı sonuçlar vermesi, bizi elle altyazılama yapmak yerine bu büyük veri kümelerinde hali hazırda bulunan İngilizce altyazıları Türkçe'ye çevirmeye yöneltti. Bu amaçla MSCOCO ve Flickr30k veri kümelerindeki tüm altyazılar Google Translate API kullanılarak Türkçe'ye çevrildi[1].

TABLO I: Deneylerimizde kullandığımız veri kümeleri ve bu kümelerdeki görüntü ve altyazı sayıları.

Veri Kümesi	Görüntü Sayısı			Görüntü başına düşen altyazı sayısı
	Eğitme	Doğrulama	Test	
Flickr8k [12]	6000	1000	1000	5 adet İngilizce
Flickr30k [17]	31783	-	-	5 adet İngilizce
MSCOCO [9]	82783	40504	40775	5 adet İngilizce
TasvirEt [11]	6000	1000	1000	2 adet Türkçe

A. Türkçe Altyazılama Modelinin Eğitilmesi ve Test Edilmesi

Deneylerimizde kullandığımız veri kümelerine ait bilgiler Tablo I'de verilmiştir. Altyazılama modeli olarak Bölüm II'de detayları açıklanan yöntemin açık olarak paylaşılan bir implementasyonunu kullandık [18]. Modeli eğitirken iki aşamalı bir süreç izledik. Öncelikle modelin ESA kısmı değiştirilmeden, Imagenet ağırlıkları olduğu gibi kullanılarak, sadece YSA (yinelgeli sinir ağı) kısmını eğittik. Daha sonra modelin hem ESA (CNN fine-tune) hem de YSA kısmını birlikte eğittik. Modelin kodlayıcı (CNN) ve kod çözücü (RNN) aşamalarında optimizasyon yöntemi olarak *Adam* kullanılmıştır. Kodlayıcıda öğrenme oranı 0.00001, kod çözücüde ise 0.0004 olarak seçilmiştir. *Alfa* ve *Beta* değerleri iki adımda da aynı olacak şekilde sırası ile 0.8 ve 0.999'dur.

Modeli test ederken hem Flickr hem MSCOCO veri kümeleri için dört farklı deney konfigürasyonu ile sonuçlar elde ettik. İlk deney konfigürasyonunda yukarıda belirtilen ilk eğitme yöntemini kullandık. İkinci deney konfigürasyonunda ise VGG16 modelini ilgili veri kümesi üzerinde *finetune* edip (ikinci eğitme yöntemi), elde ettiğimiz modelleri karşılık geldikleri veri kümelerine uygulayıp sonuçlar elde ettik. Sonraki deneysel çalışmamızda, elde ettiğimiz altyazılardaki kelimeleri, anlamlı olarak kalabilen en küçük parçalarına yani köklerine ayırarak ilgili skorları hesapladık. Son olarak ise öğrenme aktarma (*İng.* transfer learning) deneyleri yaptık. Burada bir veri kümesi üzerinde eğitilmiş model (ikinci yöntem kullanılarak), diğer veri kümesi üzerinde test edilmiştir.

Başarım ölçümü için BLEU [19], METEOR [20], Rouge-L [21] ve CIDEr [22] skorlarını kullandık. Bu metriklere ait skorları hesaplayabilmek adına açık kaynak kodlu yazılım kullanılmıştır [23].

B. Flickr Deneyleri

Flickr8k ve Flickr30k, görüntü altyazılama problemi için hazırlanmış öncü veri kümeleridir. Flickr30k, Tablo I'de veril-

[1] Bildiride kullanılan kodlara ve çevrilmiş Türkçe altyazılara https://github.com/giddyyupp/turkish-image-captioning adresinden ulaşılabilir.

Bir grup genç erkek basketbol oynuyor.

İki motosikletçi bir pistte yarışıyor.

Frizbi oyunu oynayan bir grup insan

Takım elbiseli ve kravatlı bir adam

İki küçük kız bir ağacın yanında duruyor.

İki futbolcu bir futbol sahasında duruyor.

Küçük bir kedi bir tuvaletin yanında duruyor.

Bir pazarda sergilenen çeşitli meyveler

Bir adam bir grup insanın önünde bir gitar çalıyor.

Bir basketbol oyuncusu bir basket yapmaya çalışıyor.

Bir çift makas ve bir çift makas

Bir kadın ve bir kadın bir trenin yanında duruyor.

Şekil 1: Farklı başarı kategorilerinde görsel altyazılama sonuçları. İlk satırda başarılı, ikinci satırda küçük hatalara sahip, son satırda ise görüntü ile alakasız altyazılama sonuçları verilmiştir. İlk iki sütun Flickr8k, ikinci iki sütun MSCOCO veri kümesine ait test görüntüleridir.

diği üzere yaklaşık 32 bin görüntüden oluşmaktadır ve aynı zamanda Flickr8k veri kümesindeki görüntüleri de içermektedir. Ancak Flick8k'nın aksine veri kümesindeki resimler eğitim ve test şeklinde işaretlenmemiştir. Flickr8k ve Flickr30k veri kümelerinde her görüntü için beş adet altyazı mevcuttur. [11] çalışması ile doğrudan karşılaştırma yapabilmek adına, Flickr8k veri kümesinde test ve doğrulama olarak işaretlenmiş görüntüler dışında kalan tüm Flickr30k görüntülerini, modeli eğitmek için kullandık.

C. *MSCOCO Deneyleri*

MSCOCO mevcut durumda görüntü ve altyazı sayısı olarak en büyük veri kümesidir. Tablo I'de görüldüğü üzere eğitme ve doğrulama kümeleri toplam 120 binden fazla görüntü içermektedir ve her görüntü için beş adet altyazı bulunmaktadır. Deneylerimizde eğitim kümesindeki 80.000 görüntü, modeli eğitmek için kullanılmıştır. Yöntemin başarımını ölçmek üzere doğrulama kümesinden rastgele seçilen 500 adet test görüntüsü iki farklı kişi tarafından dayanak oluşturmak amacıyla elle altyazılanmıştır.

Elde ettiğimiz sonuçlar Tablo II'de verilmiştir. Şekil 1'de F2 ve M2 yöntemleri ile elde edilen bazı sonuçlar niteliksel değerlendirme için görsel olarak sunulmuştur. Şekil 1 incelendiğinde yöntemin oluşturduğu altyazıların gramatik açıdan başarılı olduğunu söyleyebiliriz.

Flickr veri kümesi üzerinde alınan sonuçlar MSCOCO'ya kıyasla beklenildiği üzere daha düşük çıkmıştır. Bunun temel sebebi MSCOCO veri kümesinin Flickr'a göre iki kattan daha fazla görüntü ve altyazı içermesidir. Benzer şekilde Flickr ile eğitilen model, verinin azlığı sebebi ile çok çabuk ezberlemeye (İng. overfit) başlamıştır. Flickr30k veri kümesinde eğitilmiş modeller, [11] çalışmasındaki referans yöntemlerin ikisinden de daha yüksek başarım elde etmiştir. [11] çalışmasında Bleu skorları hesaplanırken, Türkçe'nin sondan eklemeli dil yapısı göz önüne alınarak tüm kelimeler köklerine ayrılarak skorlar hesaplanmıştır. Sonuçların karşılaştırılabilir olması adına, elde edilen altyazılardaki kelimeleri Zemberek [24] ile köklerine ayırarak da skor hesapladık ve başarımın daha da arttığını gözlemledik.

Öğrenme aktarma testlerinin her ikisinde de başarımın düştüğü gözlemlenmiştir. Bunun temel sebebi eğitme ve test aşamalarında kullanılan altyazıların farklı kaynaklardan gelmesi olarak açıklanabilir. Bu duruma [13] çalışmasında da değinilmiş ve özellikle çok daha fazla veriye sahip MSCOCO üzerinde eğitilmiş modelin Flickr üzerindeki başarımı artırmaması altyazıların farklı gruplar tarafından hazırlanmış olması ile açıklanmıştır. Benzer durum, eğittiğimiz Flickr modellerinin çıktılarını TasvirEt altyazıları ile test ettiğimizde de oluşmaktadır. Tablo II'deki son satır F2 modelinin, dayanak olarak TasvirEt altyazıları yerine, aynı görüntülerin, İngilizce dayanak altyazılarının Google Çeviri kullanılarak elde edilmiş Türkçe karşılıklarıyla test edildiğinde elde edilen başarımını temsil etmektedir. Görüldüğü üzere aynı grup tarafından hazırlanmış altyazılar başarımı artırmıştır.

Test sonuçlarındaki bir başka önemli nokta ise İngilizce altyazılar ile eğitilmiş bir model kullanarak elde edilen test altyazılarının Google Çeviri ile Türkçe'ye çevrilmesi ile hesaplanan skorların, Türkçe ile eğitilmiş modelden daha düşük çıkmasıdır. MSCOCO veri kümesindeki tüm eğitme ve doğrulama gruplarındaki görüntüler ve İngilizce altyazıları kullanılarak eğitilmiş bir modelin Tasviret (Flickr8k) test görüntüleri için ürettiği altyazıları Türkçe'ye çevirdik ve TasvirEt altyazılarını kullanarak skor hesapladık. Tablo II'ın 11. satırında verilen skorlar ("MSCOCO İng. Model"), 8. satırda ("MSCOCO

TABLO II: DENEYSEL SONUÇLAR.

Model	Veri Kümeleri		Metrikler						
	Eğitme	Test	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L	CIDEr
No Finetune (M1)	MSCOCO Eğitme	MSCOCO Doğrulama 500 Görüntü	0.282	0.150	0.073	0.035	0.141	0.288	0.360
CNN Finetune (M2)			0.313	0.173	0.085	**0.044**	0.154	0.308	0.450
CNN + Kök Alma (M3)			**0.342**	**0.181**	**0.085**	0.039	**0.157**	**0.322**	**0.461**
Flickr30k (F2)	Flickr30k - Eğitme		0.195	0.077	0.024	0.011	0.099	0.207	0.094
No Finetune (F1)	Flickr30k - Eğitme	Tasviret - Test	0.288	0.145	0.066	0.029	0.106	0.244	0.146
CNN Finetune (F2)			0.307	0.159	0.074	0.030	0.115	0.255	0.158
CNN + Kök Alma (F3)			**0.335**	**0.172**	**0.082**	**0.035**	**0.123**	**0.268**	**0.236**
MSCOCO (M2)	MSCOCO - Eğitme		0.291	0.143	0.059	0.022	0.107	0.242	0.119
TasvirEt [11] Yöntem 1 (T1)	Tasviret - Eğitme	Tasviret - Test	0.211	0.072	0.020	-	-	-	-
TasvirEt [11] Yöntem 2 (T2)			0.260	0.102	0.034	-	-	-	-
MSCOCO İng. Model	MSCOCO	Tasviret - Test	0.264	0.136	0.057	0.021	0.097	0.228	0.125
F2	Flickr30k - Eğitme	Flickr30k - Test	0.341	0.192	0.107	0.053	0.132	0.304	0.213

(M2)") verilen skorlardan CIDEr hariç daha düşük çıkmıştır.

MSCOCO veri kümesi ile eğitilen modelin başarımı ise hem Flickr modeli ile kıyaslandığında hem de İngilizce ile eğitilen modeller ile kıyaslandığında kabul edilebilir bir seviyededir. Örneğin aynı modelin İngilizce altyazılar ile eğitilmiş versiyonunun CIDEr skoru 0.8 civarında iken Türkçe çevirileri 0.5 civarında çıkmıştır. Bu sonuçtaki sebeplerden birisi Türkçe'nin İngilizceye göre daha zor bir gramatik yapıya sahip olmasıdır. Burada çeviriler üzerinde yapılacak iyileştirmeler ile İngilizce için elde edilen sonuçlara yaklaşılabilinir.

IV. SONUÇ VE GELECEK ÇALIŞMALAR

Bu çalışmada görüntü altyazılama için tercüme araçları kullanarak otomatik Türkçe veri kümesi oluşturulabilir mi sorusuna cevap aradık. Bu çerçevede İngilizce için var olan veri kümelerini Türkçe'ye çevirerek çeşitli deneyler gerçekleştirdik. Deneylerimizde Türkçe için var olan en yüksek başarımı elde ettik. Bu başarı, Google Çeviri ile elde ettiğimiz Türkçe altyazılama eğitim kümesinin geçerli bir veri kümesi olduğunu ve böylece Türkçe için var olan en büyük altyazılama veri kümesini oluşturmuş olduğumuzu göstermektedir. Elde ettiğimiz sonuçlar bize otomatik tercüme ile görüntü altyazılama için, iyileştirmelere açık bir veri kümesi oluşturulabileceğini göstermiştir.

Gelecekte, Google Çeviri tarafından otomatik olarak üretilen Türkçe altyazılardaki hataların, doğal dil işleme yöntemlerini kullanarak tespit edilmesi ve giderilmesi üzerinde çalışacağız.

TEŞEKKÜR

"NVIDIA Academic Hardware Grant" programı kapsamında araştırma grubumuza Tesla K40 GPU kartı hibe eden NVIDIA firmasına teşekkür ederiz.

KAYNAKLAR

[1] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NIPS*, 2015.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[3] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144*, 2016.

[4] X. Chen and C. L. Zitnick, "Mind's Eye : A Recurrent Visual Representation for Image Caption Generation," *CVPR*, 2015.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." in *ICML*, vol. 14, 2015, pp. 77–81.

[6] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.

[7] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *NIPS*, 2011, pp. 1143–1151.

[8] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: Understanding and generating simple image descriptions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2891–2903, 2013.

[9] 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 *ECCV*. Springer, 2014, pp. 740–755.

[10] B. Turovsky, "Found in translation: More accurate, fluent sentences in google translate," https://blog.google/products/translate/found-translation-more-accurate-fluent-sentences-google-translate/.

[11] 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 *2016 24th Signal Processing and Communication Application Conference (SIU)*, May 2016, pp. 1977–1980.

[12] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 139–147.

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

[14] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] 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.

[18] A. Karpathy, "Efficient image captioning code in torch, runs on gpu," https://github.com/karpathy/neuraltalk2.

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.

[20] M. D. A. Lavie, "Meteor universal: language specific translation evaluation for any target language," *ACL 2014*, p. 376, 2014.

[21] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries."

[22] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.

[23] "Microsoft coco caption evaluation," https://github.com/tylin/coco-caption.

[24] A. A. Akın and M. D. Akın, "Zemberek, an open source nlp framework for turkic languages."