TasvirEt: Görüntülerden Otomatik Türkçe Açıklama Oluşturma İçin Bir Denektaşı Veri Kümesi
TasvirEt: A Benchmark Dataset for Automatic Turkish Description Generation from Images

Mesut Erhan Unal[1], Begum Citamak[1], Semih Yagcioglu[1], Aykut Erdem[1],
Erkut Erdem[1], Nazli Ikizler Cinbis[1], Ruket Cakici[2]
[1]Hacettepe Bilgisayarlı Görü Laboratuvarı (HUCVL), Hacettepe Üniversitesi
erhan.unal@hacettepe.edu.tr, b21228194@cs.hacettepe.edu.tr, semih.yagcioglu@hacettepe.edu.tr,
{aykut,erkut,nazli}@cs.hacettepe.edu.tr
[2]Bilgisayar Müh. Bölümü, Orta Doğu Teknik Üniversitesi
ruken@ceng.metu.edu.tr

***Özetçe* —Görüntülerin doğal cümlelerle otomatik olarak tasvir edilmesi literatürde çok yakın zamanlarda incelenmeye başlanmış olan ve son derece zorlu kabul edilen bir araştırma problemidir. Bu problemin çözümüne yönelik ortaya konan yaklaşımların sayısının giderek artmasına rağmen bu alanda yaygın olarak kullanılan veri kümelerinin sadece İngilizce açıklamalar içermeleri nedeniyle bu çalışmalar büyük ölçüde tek dillidir ve İngilizce ile kısıtlı kalmıştır. Bu çalışmada, literatürde ilk kez görüntülerden Türkçe açıklamalar yaratmaya imkan veren ve bu amaçla denektaşı olarak kullanılabilecek yeni bir veri kümesi sunulmaktadır. TasvirEt adını verdiğimiz bu veri kümesi üzerinde, yine literatürde ilk kez Türkçe görüntü altyazılama amacıyla kullanılabilecek iki yaklaşım da önerilmektedir. Elde edilen deneysel sonuçlar bu veri kümesinin ve önerilen yaklaşımların görüntülerin otomatik olarak Türkçe tasvir edilmesinde başarılı şekilde kullanılabileceğini göstermektedir.**

Anahtar Kelimeler*—*Görüntü altyazılama, bilgisayarlı görü, doğal dil işleme

***Abstract*—Automatically describing images with natural sentences is considered to be a challenging research problem that has recently been explored. Although the number of methods proposed to solve this problem increases over time, since the datasets used commonly in this field contain only English descriptions, the studies have mostly been limited to single language, namely English. In this study,for the first time in the literature, a new dataset is proposed which enables generating Turkish descriptions from images, which can be used as a benchmark for this purpose. Furthermore, two approaches are proposed, again for the first time in the literature, for image captioning in Turkish with the dataset we named as TasvirEt. Our findings indicate that the new Turkish dataset and the approaches used here can be successfully used for automatically describing images in Turkish.**

Keywords*—*Image captioning, computer vision, natural language processing

I. GİRİŞ

Bilgisayarlı görü ve doğal dil işlemenin ortak bir problemi olan görüntü altyazılama, her geçen gün hızla artmakta olan veri miktarı ile daha da önemli bir problem haline gelmiştir. Görüntülerin otomatik altyazılarını oluşturmak, görüntü anlamlandırma ve robotik uygulamaları bağlamında son derece büyük bir önem arz etmektedir. Literatürde bu kapsamda önerilmiş olan görüntü altyazılama yöntemlerini iki grupta incelememiz mümkündür. Bunlardan ilki oluşturma (*generative*) tabanlı yöntemler, ikincisi ise getirim (*retrieval*) tabanlı yöntemlerdir.

Yolda kayan yarı çıplak bir adam.

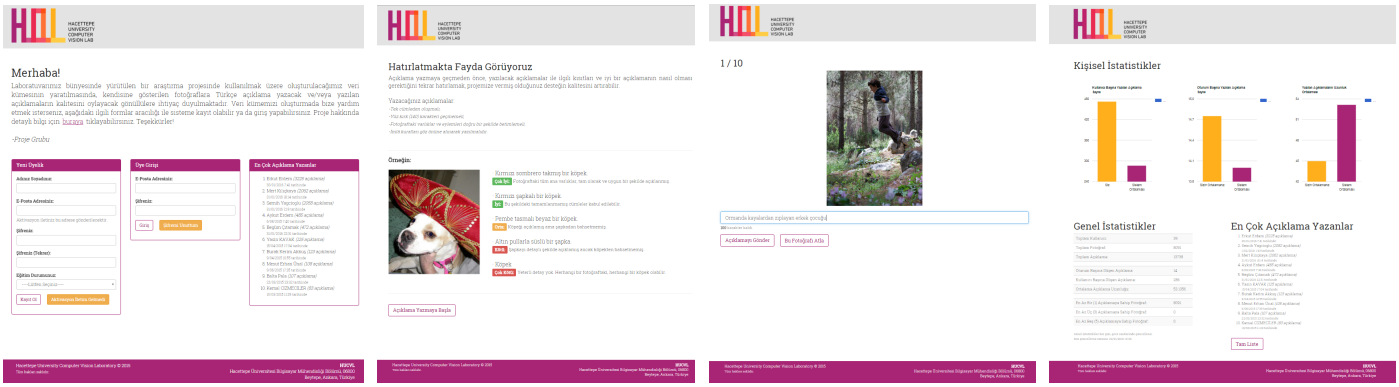
Şekil 1: TasvirEt veri kümesinden alınmış örnek bir görüntü ve bu görüntüye ait kitle-kaynaklı şekilde toplanan bir altyazı.

İlk grupta yer alan oluşturma tabanlı yöntemlerde temel yaklaşım, görüntü içeriğindeki bilgi kullanılarak altyazıların doğal dil oluşturma teknikleri ile oluşturulmasıdır [1]–[4]. Bu yöntemlerde genel olarak görüntülerin içinde yer alan nesneleri, sahneleri ve eylemleri belirlemek için bilgisayarlı görü teknikleri kullanılır ve altyazılar bu alt bileşenler kullanılarak oluşturulur. İkinci grupta yer alan getirim tabanlı çalışmalar ise ilk gruptaki çalışmalardan farklı olarak görüntülerin görsel benzerliklerinden ve altyazıların metinsel benzerliklerinden aynı anda yararlanarak görüntüler için uygunluğu en olası altyazıyı var olan altyazıların arasından seçerler [5]–[8].

Bu çalışmada, Türkçe altyazılama için bir veri kümesi önerilmektedir. Ek olarak, bu veri kümesi kullanarak literatürdeki ilk otomatik Türkçe altyazı üreten yöntemler sunulmaktadır. Bu yöntemler, getirim tabanlı görüntü altyazılama yöntemleri arasında yer almaktadır. Temel olarak, birinci yöntem, sadece görüntüler arasındaki görsel benzerliğe dayanırken, ikinci yöntem adaptif komşuluk içindeki aday görüntü kümesi alt yazıları üzerinde, Türkçe köklerine ayırma işlemi uygulanması ve oluşturulan kök tabanlı gösterim üzerinden BLEU metriği ile en yakın görüntülerin ve alt yazıların getirilmesi şeklinde işlemektedir.

Bu çerçevede, yaptığımız katkılar (1) Türkçe açıklamalar

978-1-5090-1679-2/16/$31.00



Şekil 2: Soldan sağa: Kullanıcı giriş ekranı, hatırlatmalar, altyazı giriş ekranı, istatistiksel durum ekranı.

içeren TasvirEt veri kümesinin oluşturulması ve (2) literatürde görüntüler için otomatik olarak Türkçe açıklamalar üreten ilk yöntemlerin önerilmesi olarak özetlenebilir. Yapılan deneyler, önerilen yöntemlerin görüntülerin Türkçe tasviri için başarılı olarak kullanabileceğini göstermektedir.

II. İLGİLİ ÇALIŞMALAR

Doğal dil işleme ve bilgisayarlı görünün bir arada çalışması yapay zekanın önemli hedeflerinden biridir. Son yıllarda bu alanda yapılan çalışmaların sayısında çok ciddi bir artış yaşanmıştır. Görüntü altyazılama konusunda son yıllarda ciddi gelişmelerin kaydedilmesini ve bu alandaki çalışmaların yaygınlaşmasını sağlayan en önemli unsur, araştırmacılara açık olarak sunulan veri kümeleridir.

Görüntü altyazılama için yaygın olarak kullanılan veri kümelerini iki ana başlık altında incelemek mümkündür. Bunlardan ilki SBU Captioned Photo Dataset [9] ve YFCC-100M [10] gibi sosyal fotoğraf paylaşım sitelerinden direkt olarak toplanan veri kümeleridir. Bu veri kümelerinin en büyük dezavantajı, kullanıcılar görüntüleri yüklerken altyazılarını serbestçe oluşturdukları için varolan açıklamalar yapısal olarak çok tutarlı olmayıp fazlaca gürültü içermektedir. PASCAL [1], Flickr8K [11], Flickr30K [12] gibi ikinci tür veri kümelerinde ise görüntülerin açıklamaları kitle-kaynaklı (*crowd-sourced*) şeklinde toplandığı için bu açıklamalar görüntü içeriği ile çok daha tutarlı ve çok daha az gürültü içermektedir. Öte yandan bu veri kümelerinin oluşturulmaları oldukça maliyetli olabilmektedir.

Yakın zamanda önerilen birkaç çalışma ve veri kümesi de bilgisayarlı görü ve doğal dil işlemeyi birlikte kullanıyor olması açısından konu ile yakından ilgilidir. Bu çalışmalardan birkaçı görsel soru cevaplama [13], boşlukları doldurma ve görsel yeniden ifade [14] çalışmalarıdır. Bu alandaki çalışmaların ilerleyişi açısından veri kümeleri son derece büyük bir öneme sahiptir. Dolayısıyla bu bağlamda yapılacak katkı, başka çalışmaların da önünü açması açısından ayrıca önem arz etmektedir. Çok dilli görüntü altyazılama veri kümelerinin bildiğimiz kadarıyla henüz önerilmiş bir örneği bulunmamaktadır. Konuyla ilgili bilinen en yakın çalışma görsel soru cevaplama için yakın zamanda önerilmiş olan Çince/İngilizce çok dilli görsel soru cevaplama veri kümesidir [15]. Bu kapsamda literatürde görüntü altyazılama için Türkçe olarak yapılmış mevcut bir çalışma bulunmamaktadır.

Tablo I: TasvirEt istatistikleri.

Görüntü Sayısı	8091
Toplanan Açıklama Sayısı	12222
Açıklamalardaki Ortalama Kelime Sayısı	8
Görülme Sıklığı 1 Olan Kelime Sayısı	2104

III. TASVİRET VERİ KÜMESİ

Bu bölümde Türkçe açıklamalar içeren TasvirEt veri kümesi[1] ile ilgili bilgiler ve verilerin nasıl toplandığı konusundaki yaklaşım anlatılmaktadır.

A. *Verinin Toplanması*

TasvirEt veri kümesi, Flickr8K veri kümesindeki görüntüler üzerine Türkçe altyazıların toplanması ile oluşturulmuştur. Bu kapsamda altyazıların toplanabilmesi için bir sistem geliştirilmiş, toplanacak olan verilerin bu sistem aracılığıyla kitle-kaynaklı şekilde toplanması amaçlanmıştır. Bu doğrultuda veri girişi halka açık olarak herkesin katkı sağlayabileceği şekilde tasarlanmıştır. Bu sistemde kullanıcılar belirli bir hesaptan giriş yaparak Flickr8K görüntü kümesine ait görüntüleri yani yaklaşık 8000 görüntüye onluk kümeler halinde açıklama yazabilmektedir. Kullanıcıların imla kurallarına uygun bir şekilde altyazıları yazabilmesi için başlangıçta bir sınama bölümü yer almaktadır. Bu sisteme ait görseller, Şekil 2'de sunulmaktadır.

B. *Veri Kümesi İstatistikleri*

TasvirEt veri kümesinde bu çalışmanın yapıldığı tarihte toplamda 8091 adet görüntü, bu görüntülere ait toplam 12222 açıklama bulunmaktadır. Açıklamaların ortalama uzunluğu ise yaklaşık olarak 8 kelimedir. Açıklamaları oluşturan kelimelerin görülme sıklığı açısından bakılacak olursa, veri kümesinde 2104 kelimenin en olası kökünün görülme sıklığı 1'dir. Ayrıca görüntü kümesindeki her bir görüntü için yazılan altyazı sayısı 1 ile 2 arasında değişmektedir. TasvirEt veri kümesi ile ilgili istatistikler Tablo I'de detaylı olarak verilmektedir.

IV. TÜRKÇE GÖRÜNTÜ ALTYAZILAMA

Çalışmamız kapsamında Türkçe görüntü altyazılama probleminin çözümüne yönelik iki veri güdümlü yaklaşım denenmiştir. Bu yöntemlerden birincisi, sadece görsel içerik üzerinden en yakın görüntünün getirilip onun altyazısının transfer edilmesi üzerinedir. İkinci yöntem, adaptif komşuluk

[1] Veri kümesine tasviret.cs.hacettepe.edu.tr adresinden veri girişi için halka açık olarak erişilebilmektedir.

Tablo II: Üstte örnek bir girdi için Zemberek ile elde edilen kelime köklerinin gösterimi. Her bir kelime kökleri ile ifade edilmekte, bir altyazı kelime sayısı boyutunda bir küme ile gösterilmektedir.

Altyazı	Kucakladığı	sörf	tahtasıyla	okyanusa	giren	bir	adam
Kökler	kucakla	sörf	tahta	okyanus	gir	bir	adam

çerçevesinde, Türkçe kök benzerlikleri üzerinden konsensus altyazının [18] bulunması suretiyle çalışmaktadır. Altta bu yaklaşımların detayları verilmektedir.

A. En Yakın Görüntü Altyazısının Transferi

Yaklaşımımızı gerçekleştirmek için ilk olarak görüntünün görsel olarak bir tanımının yapılması gerekmektedir. Bunun için her görüntü için ImageNet veri kümesi üzerinde eğitilmiş bir derin öğrenme modeli olan 16-katmanlı Konvolüsyonel Yapay Sinir Ağının $fc7$ aktivasyonları ile 4096 boyutlu vektörler oluşturulmuştur [16]. Bu gösterim üzerinden, sorgu görüntüsü ve diğer görüntüler arasındaki uzaklık öklid uzaklığı cinsinden ölçülerek sorgu görüntüsüne en yakın N[2] aday görüntü getirilmiştir. Bu kümeden öklid uzaklığı en düşük olan görüntü ise görsel olarak en yakın görüntüyü oluşturmaktadır. Bu yaklaşım, sorgu görüntüsünün altyazısı olarak, en yakın görüntünün altyazısının transfer edilmesi ile çalışan yaklaşımdır ve altyazılama için kullanılabilecek en temel yaklaşımdır.

B. Konsensus Altyazının Bulunması

Adaptif Komşuluk. Getirilen aday görüntülerin uygunluğunun arttırılması için Denklem 1'de belirtilen formül uygulanarak aday görüntü kümesi yeniden oluşturulmuştur.

$$N(I_q) = \{(I_i, c_i) | dist(I_q, I_i) \leq (1+\epsilon) dist(I_q, I_c), \\ I_c = \arg\min dist(I_q, I_i), I_i \in \tau\} \quad (1)$$

Denklemde $dist$ ile sorgu ve aday görüntüler arasındaki öklid uzaklığı, N ile adaptif aday görüntü kümesi, τ ile eğitim görüntü kümesi, ϵ[3] ile de pozitif skaler bir değer ifade edilmekte, c I görüntüsüne ait altyazıyı I_q sorgu görüntüsünü I_c ise görsel olarak en yakın görüntüyü belirtmektedir. Gerçekleştirilen bu işlemler ile en yakın komşu görüntü kümesi bulunmaktadır. Bu görüntülerin altyazıları bizim alt yazılarımızı oluşturmaktadır. Sonrasında yapılan işlemler bu altyazılar üzerinde gerçekleştirilmiştir.

Köklerine Ayırma. Türkçenin sondan eklemeli yapısı göz önüne alındığında, altyazıların benzerliklerinin verimli bir şekilde tanımalanabilmesi için, kelime benzerliklerine bakılırken, kök benzerlikleri üzerinden gitmek çok önemlidir. Bu nedenle, benzer Türkçe alt yazıların bulunabilmesi amacı ile, kök bilgisinden faydalanılması gerekmektedir. Köklerine ayrılmadan altyazıdaki kelimelerin doğrudan kullanılması performansı yüksek ölçüde düşürmektedir. Bu sebeple altyazıyı oluşturan kelimelerin en olası köküne ayrılması Zemberek [17] ile gerçekleştirilmiştir. Zemberek ile örnek bir girdi için oluşturulan kelime kökleri Tablo II'te gösterilmektedir. Örnekte de görüldüğü üzere altyazıları oluşturan kelimelerin daha basit düzeye indirgenerek oluşturulan altyazıların daha tanımlayıcı olması sağlanmıştır.

[2] $N = 100$ olarak seçilmiştir

[3] Adaptif komşuluk parametresi $\epsilon = 1.15$ olarak seçilmiştir.

Altyazı Seçimi. Yöntemimizde, aday altyazıların kökleri ile gösterimi ile karşılaştırılabilir hale gelmesi sağlanmıştır. Belirlenen altyazılar arasındaki benzerliği hesaplamak için her birinin diğer alt yazılar ile olan BLEU skoru hesaplanarak en yüksek skora sahip altyazı en uygun altyazı olarak seçilmiştir [18]. Burada izlenen yaklaşıma göre adaptif komşuluğa göre belirlenen imgeler arasında sorgu imgesini dikkate almadan altyazılar arasından en merkezi altyazı seçilmektedir. Merkezilik tanımı ise getirilen kümede bulunan her bir altyazı için diğerlerine ortalamada ne kadar yakın olduğu üzerinden hesaplanmaktadır. Bunun için kullanılan formül, Denklem 2'de verilmiştir.

$$c* = \arg\max_{c \in C} \sum_{c' \in C} Sim(c, c') \quad (2)$$

Bu formülde $Sim(c, c')$ fonksiyonu iki altyazı arasındaki benzerlik fonksiyonunu, $c*$ ise C içerisindeki altyazılar arasında ortalama benzerliği en yüksek konsensus altyazısını ifade etmektedir. Benzerlik fonksiyonu 1-to-4-gram BLEU olarak kullanılmıştır ve bu benzerlik şu şekilde tanımlanmaktadır:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count_{clip}(n-gram')} \quad (3)$$

$$Count_{clip} = min(Count, Max\_Ref\_Count) \quad (4)$$

Burada başlangıçta cümle cümle n-gram eşleşmeleri bulunmakta, ardından da her aday cümle için $clipped\ n-gram$ sayısı toplanarak aday cümlelerin n-gram sayısına bölünmektedir. $Count_{clip}$ hesaplanırken karşılaştırılan cümlelerde en fazla gözlenme sayısını aşmaması sağlanmaktadır.

V. DENEYLER

Altyazıların doğruluğunu ölçümlemek amacı ile, çalışmamızda BLEU metriği kullanılmıştır. BLEU yöntemi makina tarafından çevrilmiş olan metinlerin insanlar tarafından yapılan çeviriler ile ne denli uyuştuğuna bakarak makina çevirisinin kalitesinin değerlendirildiği Papineni vd. [19] tarafından önerilmiş bir ölçümdür. Bu ölçümde çeviri, insan tarafından yapılmış olan çeviriye ne kadar yakınsa çevirinin kalitesi de o denli iyi kabul edilir. BLEU algoritması, aday çeviri ile insan çevirisi arasında kaç kelimenin örtüştüğü üzerinden bir hesap yapmaktadır ve sonuçları 0 ile 1 arasında değişmektedir. Her ne kadar BLEU yöntemi bazı çalışmalar tarafından eleştirilmekte olsa da Kuznetsova vd. [6] otomatik çeviri ölçümleri arasında insan çevirimi ile önemli ölçüde uyuşan en yaygın yöntemlerden biri olması, dilden bağımsız ve kullanımının kolay olması açısından Türkçe metinlerin çevirim kalitesini ölçmek açısından bu çalışma kapsamında otomatik çevirim değerlendirme yöntemi olarak tercih edilmiştir.

TasvirEt veri kümesi üzerinde, 1000 adet olan tüm test görüntüsü için getirdiğimiz altyazıları, görüntülerin gerçek altyazıları ile karşılaştırdığımızda elde ettiğimiz BLEU [19] skorlarının ortalaması Tablo III'de yer almaktadır. Burada, `Y1`, görsel olarak en yakın görüntünün altyazısını transfer etme işlemini, `Y2` ise, konsensus altyazılama yöntemini ifade etmektedir. Ayrıca elde edilen BLEU skorları $n = 1, 2, 3$ olacak şekilde belirlenmiştir. Tablo III'da görüldüğü gibi, konsensus yönteminin Türkçe altyazılama başarısı, temel alınan en yakın görüntü tabanlı altyazılamaya göre çok daha yüksektir.

				
G	Sarı renkli, iri bir bir köpek merdivenlerden koşarak iniyor.	Bir adam tekne ile su kayağı yapıyor.	Dalgadan havalanan bir sörfçü.	Tenis raketi ile topa vuran kadın tenisçi.
Y1	sari bir kopek cesmeden su iciyor.	Bir adam sörf yapıyor	El arabasına oturmuş bir kız çocuğunu çeken patenli bir bayan.	Bir adam elinde tenis raketiyle koşuyor.
Y2	Çimlerde koşan bir köpek.	Azgın dalgaların arasında sörf yapan bir adam.	Dalgada sörfle kayan bir adam	Tenis oynayan bir kadın

Şekil 3: Test görüntülerinden getirilen altyazılar.

Tablo III: Y1 ve Y2 yöntemleri ile 1000 test görüntüsü için ortalama BLEU değerleri.

	BLEU-1	BLEU-2	BLEU-3
Y1	0.211	0.072	0.020
Y2	**0.260**	**0.102**	**0.034**

Yöntemlerin TasvirEt veri kümesi üzerinde uygulanması sonucunda test kümesinden birkaç örnek görüntü için getirilen altyazılar[4] Şekil 3'te yer almaktadır. Burada G ile gösterilen altyazılar görüntülere kullanıcıların yazdıkları altyazılardır. Görüldüğü gibi konsensus altyazılama yönteminin oluşturduğu altyazılar, sadece görüntü tabanlı altyazı transfer yöntemine göre görüntünün içeriğini anlatmada çok daha başarılıdır.

VI. SONUÇ

Bu çalışmamızın sonucunda görüntülerden otomatik olarak açıklama oluşturulması için TasvirEt isimli Türkçe bir veri kümesi oluşturulmuştur. Bu veri kümesi, halihazırda İngilizce açıklamaları olan Flickr8K veri kümesinin Türkçe altyazılar ile zenginleştirilmiş halidir. Bu veri kümesi üzerinde, Türkçe altyazılama problemi tanımlanmış, ve bu doğrultuda, Türkçe altyazılama amacı ile kullanılabilecek iki yöntem sunulmuştur. Veri güdümlü yaklaşımların başarısı, veri kümesindeki altyazı miktarı ile doğru orantılı olduğu için, veri girişinin yaygınlaştırılması ve altyazıların çoğaltılması faydalı olacaktır. TasvirEt veri kümesi, Türkçe altyazılama amacı ile kullanılabileceği gibi, çok dilli yapısı sayesinde, altyazılama probleminde dillerin kullanımındaki farklılıkların da araştırılmasına olanak sağlayacak bir veri kümesidir.

TEŞEKKÜR

Bu çalışma kısmen TÜBİTAK-COST 113E116 nolu proje tarafından desteklenmiştir.

KAYNAKÇA

[1] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: Generating sentences from images," in *ECCV*. Springer, 2010, pp. 15–29. 1, 2

[2] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg, "Babytalk: Understanding and generating simple image descriptions," *PAMI, IEEE Transactions on*, vol. 35, no. 12, pp. 2891–2903, 2013. 1

[3] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. Berg, K. Yamaguchi, T. Berg, K. Stratos, and H. Daumé III, "Midge: Generating image descriptions from computer vision detections," in *Proc. of EACL*, 2012, pp. 747–756. 1

[4] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. of CVPR*, 2015, pp. 3128–3137. 1

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

[6] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Proc. of ACL: Long Papers-Volume 1*, 2012, pp. 359–368. 1, 3

[7] R. Mason and E. Charniak, "Nonparametric Method for Data-driven Image Captioning," in *Proc. of ACL*, 2014. 1

[8] M. Kilickaya, E. Erdem, A. T. Erdem, N. I. Cinbis, and R. Cakici, "Data-driven image captioning with meta-class based retrieval," in *SIU*. IEEE, 2014, pp. 1922–1925. 1

[9] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: Describing images using 1 million captioned photographs," in *Advances in NIPS*, 2011, pp. 1143–1151. 2

[10] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "The new data and new challenges in multimedia research," *arXiv preprint arXiv:1503.01817*, 2015. 2

[11] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proc. of NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010, pp. 139–147. 2

[12] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, pp. 67–78, 2014. 2

[13] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proc. of ICCV*, 2015, pp. 2425–2433. 2

[14] X. Lin and D. Parikh, "Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks," in *Proc. of CVPR*, 2015, pp. 2984–2993. 2

[15] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question answering," *arXiv preprint arXiv:1505.05612*, 2015. 2

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. 3

[17] "Zemberek-NLP," https://github.com/ahmetaa/zemberek-nlp, 2015, [Online; accessed 7-Feb-2015]. 3

[18] J. Devlin, S. Gupta, R. Girshick, M. Mitchell, and C. L. Zitnick, "Exploring nearest neighbor approaches for image captioning," *arXiv preprint arXiv:1505.04467*, 2015. 3

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proc. of ACL*, 2002, pp. 311–318. 3

[4] İki farklı açıklamaya sahip görüntüler için, altyazılardan herhangi birine yer verilmiştir.