

Derin Öğrenme ile Otomatik Olarak Üretilen Resim Alt Yazısındaki Hataların Tespiti

Detecting Errors in Automatic Image Captioning by Deep Learning

Murat Karakaya
Computer Engineering Dept.
Atılım University
Ankara, Turkey
murat.karakaya@atilim.edu.tr

Öz—Resimlerin otomatik olarak etiketlenmesi görüntü işleme alanında önemli araştırma konularından biridir. Bu konuya benzer bir diğer alan ise resimleri tasvir eden resim alt yazılarının otomatik olarak üretilmesidir. Bu çalışmamızda, resimlere otomatik olarak alt yazı üreten modellerin ürettikleri alt yazılardaki hataların tespiti için resimleri otomatik olarak etiketleyen bir derin öğrenme modeli önerilmiştir. Yapılan ilk deneyler sonucunda önerilen sistemin hataları %80 başarıyla tespit edebildiği görülmüştür.

Anahtar Sözcükler—görüntü işleme, etiketleme, alt yazı üretimi, hata tespiti, derin öğrenme, derin sinir ağları

Abstract— Automatic tagging of images is an important research topic in the field of image processing. Another area similar to this is the automatic generation of picture captions. In this study, a deep learning model that automatically tags the pictures is used to detect errors in image captions. As a result of the initial experiments, it is observed that the proposed system can find up to 80% of the errors in the image captions.

Keywords—image processing, tagging, captioning, error detection, deep learning, deep neural networks

I. GİRİŞ

Bu çalışmada son yıllarda yaygınlık kazanan otomatik olarak üretilen Resim Alt Yazısında (RAY) yapılabilecek hataların tespitine yönelik bir çözüm geliştirilmiştir. Geliştirilen çözüm, Derin Öğrenme yaklaşımını kullanarak otomatik olarak resim etiketlerinin üretimine ve üretilen bu etiketler ile önerilen RAY içeriğinin karşılaştırılmasına dayanmaktadır. Bu çözüm; RAY’da geçen ancak üretilen etiketlerde var olmayan kelimelerin hata olarak değerlendirilmesini ve raporlamasını kapsamaktadır. Daha iyi içerik oluşturmaya yönelik literatürde çalışmaların varlığına rağmen [1-5], üretilen alt yazılardaki hataların otomatik olarak tespitine yönelik çok az sayıda çalışma mevcuttur [6]. Bu nedenle, çalışmamız bu konuda yapılan ilk çalışmalardan biri olmaktadır.

Son yıllarda artan taşınabilir cihaz kullanımı, bulut üzerinden çalışan yedekleme hizmetleri sayesinde çoklu ortam verilerinde büyük bir artış sağlamıştır. Örneğin, en son istatistiklere göre; şimdiye kadar Instagram adlı sosyal medya ortamına 50 milyardan fazla resim yüklenmiştir ve her saniye bu rakama 995 yeni resim de eklenmektedir [7]. Yüklenen bu fotoğrafların aranmasında en yoğun olarak anahtar kelimeler kullanılmaktadır. Dolayısıyla, fotoğrafların doğru ve kapsayıcı nitelikte etiketlenmesi, arama sonuçlarını doğrudan etkilemektedir. Bu etiketlemenin kullanıcıların kendisi tarafından manuel olarak yapılması, istenilen kalitede

sonuçları doğurmadığından bu işlemin otomatik olarak yapılması hedeflenmiştir. Bu maksatla, yazında bir çok resim etiketleme (image tagging) metodu önerilmiştir [8, 9, 10].

Resimlerin sadece etiketlenmesi de bazı durumlarda yetersiz kalmaktadır. Resim içeriklerinin uygun bir cümle yapısı içerisinde ifade edilmesi de önem kazanmaktadır. Bu cümlelerin otomatik olarak yaratılması problemine Resim Alt Yazısı (RAY) Üretimi (Image Captions Generation) adı verilmiştir [9]. Üretilen RAY’ların, hem cümle yapılarının doğru hem de söz konusu resmi, içerik olarak yeterli seviyede ve doğru tasvir etmesi önemli bir başarımlı ölçütüdür.

Yukarıda tanıtılan her iki problemin çözümünde de geliştirilen en son çözümler, Derin Öğrenme (Deep Learning) ya da diğer bir adlandırma ile Derin Sinir Ağlarını (Deep Neural Networks) kullanmaktadırlar [10, 11, 12, 13, 16, 18].

A. Derin Öğrenme ile Resim Etiketleme

Derin Sinir Ağlarını (DSA) kullanarak verilen veri setinden kuralların veya örüntülerin öğrenilmesine Derin Öğrenme adı verilmektedir [14]. Resim Etiketleme probleminin çözümü için son yıllarda en sık kullanılan yöntemler Derin Öğrenmeye dayanmaktadır [8]. Bu yöntemlerin ortak noktası, resim ve bu resme ait insan tarafından yazılmış etiketlerden oluşan veri setlerinin eğitimde kullanılmasıdır. Tüm etiketler incelenerek bir sözlük oluşturulur. Bu sözlük bir vektör ile gösterilir. DSA kullanılarak bu resimler girdi, muhtemel etiketler ise vektör gösterimde çıktı olarak kullanılır. DSA, eğitim esnasında her bir resim ve etiket vektörünü dikkate alarak kendi ağırlıklarını düzenleyip resim ve etiketler arasındaki ilişkiyi öğrenmeye çalışır. Eğitim belli bir başarıya ulaşıncaya durdurulur. Test aşamasında ise eğitilen DSA’ya daha önceden görmediği bir resim verilir ve çıktı olarak etiket vektörünü üretmesi beklenir.

B. Derin Öğrenme ile Resim Alt yazısı Oluşturma

RAY üretmede resim etiketleri üretmeden farklı olarak, verilen resmi tasvir eden düzgün bir cümle üretilmesi hedeflenir [9]. Son yapılan çalışmalarda düzgün cümle üretilmesinde gösterdiği başarıdan dolayı Özyinelemeli Sinir Ağları (RNN) kullanılmaktadır. Bu çözüm önerilerinde ortak yaklaşım, eğitim esnasında verilen her resim için bir DSA kullanılarak resme ait özet bir bilgi içeren vektör oluşturulması ve vektörle birlikte verilen eğitim setindeki alt yazının kullanılarak ayrı bir RNN eğitilmesidir. Böylece sistem hem resim vektörünü hem de cümlede geçen kelimelerin sıralamasını dikkate alarak cümle üretmeyi öğrenebilir.

C. Derin Öğrenme ile Resim Alt Yazısındaki Hataların Tespiti

Yukarıda özetlenen her iki problem de aslında birbirlerinden farklı oldukları için çözüm yöntemleri de farklı olmuştur. Göreceli olarak, RAY probleminin doğru çözümü daha zor ve karmaşık olarak görülebilir. Bu nedenle, yazın incelendiğinde, RAY probleminin çözümü için önerilen bir çok makalede verilen farklı ölçütlere göre başarımının çok yüksek olmadığı görülebilir. Örneğin, Flcikr8k veri seti kullanılarak yapılan bir çok çalışmada elde edilen sonuçlar farklı ölçütlerle incelendiğinde başarımın %70 seviyesini geçmediği raporlanmıştır [9]. Çalışmalar genelde resim alt yazını üretmeyi hedeflemişler ancak üretilen bu alt yazılardaki hataları otomatik olarak tahminleyip düzeltmeyi hedeflememişlerdir. Söz konusu çalışmaların başarımının ölçümü için genellikle insan gücü kullanılmıştır. Üretilen alt yazılardaki hataların otomatik olarak tespitine yönelik bizim bilginiz dahilinde olan bir çalışma mevcut değildir. Bu nedenle, çalışmamızda aşağıda detayları açıklanan çözümü önermekteyiz.

II. YÖNTEM

Üretilen Resim Alt Yazısındaki (RAY) muhtemel hataları tespit edebilmek için önerdiğimiz sistem Resim Etiketleme çözümünün, RAY çözümüne entegre edilmesine dayanmaktadır. Öncelikle, resimlerden ve onlara ait RAY'lardan oluşan eğitim veri setini kullanarak RAY üretmek için tasarlanmış Derin Sinir Ağı (RAY-DSA) eğitilir (Şekil 1). Daha sonra, veri setinde bulunan RAY'lar; dizgelere (token) ayrılır, bağlaçlardan (stop words) ve kelime eklentilerden temizlenerek her resim için etiket (tag) seti oluşturulur. Böylece, RAY için hazırlanmış veri seti etiketleme probleminde kullanılabilir hale dönüştürülmüş olur. Dönüştürülmüş olan bu veri seti kullanılarak başka bir derin sinir ağı (Etiket-DSA) eğitilir (Şekil 2). Sonuçta, önerilen çözümde, biri RAY üretmek diğeri etiket üretmek üzere aynı veri seti ile eğitilmiş iki farklı derin sinir ağından oluşan bir model hazırlanmış olur.

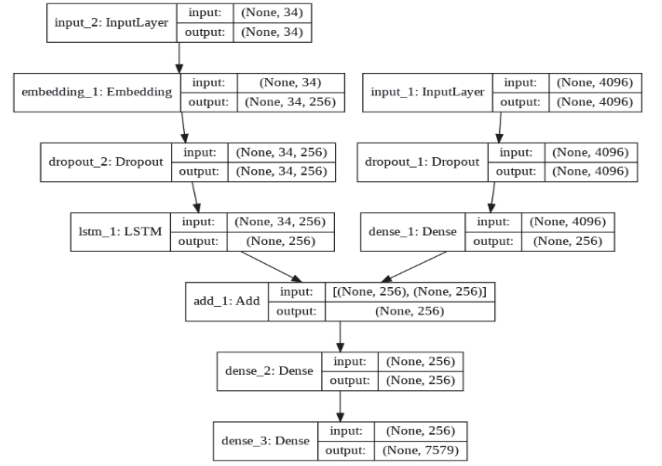
Test aşamasında resim öncelikle RAY-DSA'ya verilerek bu resme ait resim alt yazısının oluşturulması sağlanır. Aynı resim etiket üretecek Etiket-DSA'ya verilerek bu resme ait etiket kümesi elde edilir. Üretilmiş olan RAY dizgelere (token) ayrılıp bağlaçlardan (stop words) ve kelime eklentilerden temizlendikten sonra resim için üretilmiş olan etiket seti ile karşılaştırılır. Üretilmiş etiket setinde bulunmayan kelimeler RAY içinde işaretlenerek muhtemel hata olarak belirlenir.

Yukarıda detayları açıklanan sistem kullanılarak yapılan deney ve elde edilen sonuçlar aşağıda sunulmuştur.

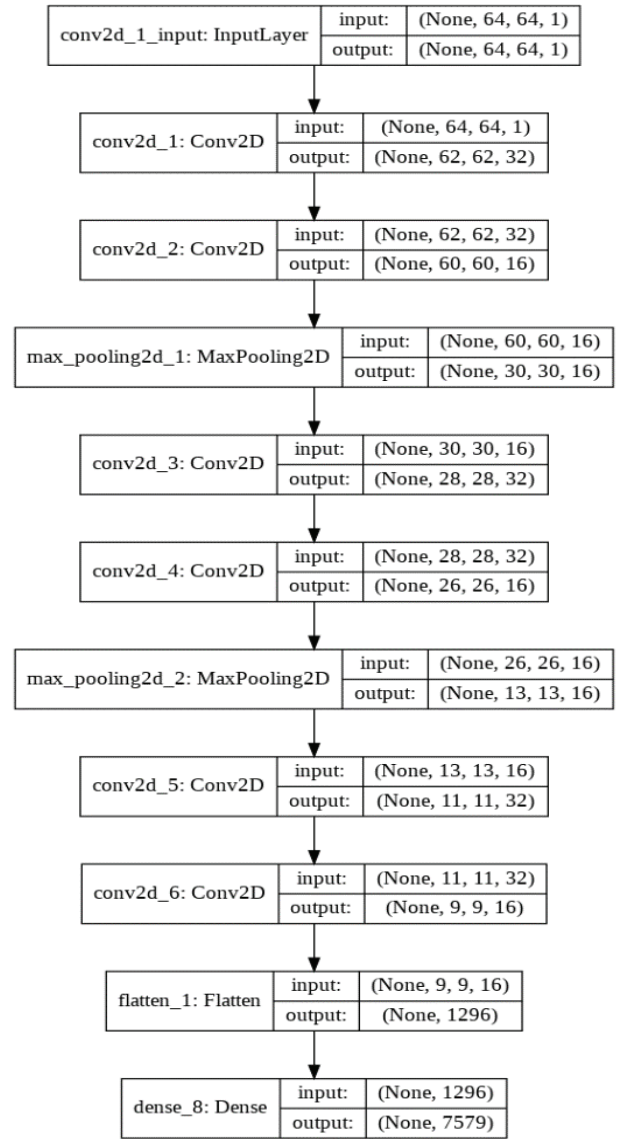
III. DENEY VE SONUÇLARI

A. Veri Seti ve Geliştirilen Derin Sinir Ağları

Resim Altı Yazı (RAY) üretmek için kullanılacak Derin Sinir Ağı (RAY-DSA) açık kaynak olarak yayınlanmış olan bir projeden alınarak geliştirilmiştir [15]. Kullanılan RAY-DSA aşağıdaki adımlar takip edilerek eğitilmiştir:



Şekil. 1. RAY-DSA: RAY üretmek için tasarlanmış Derin Sinir Ağı



Şekil.2. Etiket-DSA: Etiket üretmek için tasarlanmış Derin Sinir Ağı

- Önerilen sistemin denenmesi için Flickr8K veri seti kullanılmıştır [16]. Bu veri setinde 8000 adet resim ve her resme ait 4 farklı insan tarafından yazılmış 4 farklı alt yazı bulunmaktadır.
- VGG16 modeli [17] öğrenme aktarımı (transfer learning) yapılarak son sınıflandırma katmanı hariç ithal edilmiştir. Bu model kullanılarak her resim için 4096 veriden oluşan görüntü vektörü elde edilmiştir.
- Her resim için verilen 4 farklı alt yazı ön işlemlerden (küçük harfe çevrilmesi, noktalı işaretlerin, 1 karakter olan kelimelerin ve rakamların silinmesi, v.b.) geçirilmiştir. Bu kelimeler kullanılarak 7579 farklı kelimedenden oluşan bir sözlük oluşturulmuştur. Bu sözlük sayesinde kelime vektörü oluşturulmuştur.
- Şekil 1’de görüldüğü üzere, RAY-DSA eğitim aşamasında, iki girdi kullanılmaktadır: görüntü ve kelime vektörleri. Kelime vektörü kullanılarak her seferinde cümle içerisinde geçen ilk kelimedenden başlayarak sırasıyla bir kelime RAY-DSA’ya verilerek cümle yapısının öğrenilmesi sağlanır. Bu öğrenme sırasında, resim vektörü de her kelime ile birlikte RAY-DSA’ya verilmektedir. Her kelime vektörü bir RNN (LSTM) ile öncelikle işlenmekte daha sonra resim vektörü ile işleneceği İleri Beslemeli (Feed Forward FF) ağı yönlendirilmektedir. Bu eğitim esnasında, ağı çıktısı olarak bir sonraki kelimenin vektörü verilmektedir.
- Bu şekilde eğitilen ağı, deneyde kullanılacak RAY’ları üretecek hale gelmektedir.

Verilen bir resmi etiketleyecek derin sinir ağını (Etiket DSA) eğitmek içinse aşağıdaki adımlar izlenmiştir:







- Öncelikle Flickr8K veri setinde yukarıda açıklanan işleme tabi tutulmuş RAY’lar üzerinde bir çalışma yapılır. Bu çalışmada, cümlelerdeki bağlaçlar (stop words) silinir, kalan kelimelerden eklentiler temizlenerek dizgeler (token) bulunur. Bu dizgeler kullanılarak her resim için etiket (tag) vektörü oluşturulur.
- Alt yazıları üretecek olan derin sinir ağı için hazırlanan resim vektörü bu ağı içinde aynı şekilde kullanılır.
- Şekil 2’de görüldüğü üzere, eğitim için Etiket-DSA’ya iki girdi verilir: görüntü ve etiket vektörleri. Resim vektörü tasarlanan İleri Beslemeli (Feed Forward FF) bir ağı girdi ve Etiket vektörü ise bu ağı beklenen çıktı olarak verilerek ağı eğitilir.

Çalışmamızda, yukarıda açıklanan her iki ağı da TensorFlow/Keras kütüphanesi kullanılarak geliştirilmiş ve eğitilmiştir. Daha sonra, test setindeki resimler bu ağlara girdi olarak verilip üretilen RAY ve etiketler karşılaştırılmıştır.

B. Sonuçlar

Tablo 1’de örnek sonuçların bir kısmı verilmiştir. Tablonun ilk sütununda verilen resim eğitilen her iki ağı verildiğinde elde edilen; resim alt yazısı (RAY) ikinci sütunda, etiketler üçüncü sütunda, RAY içinde tahmin edilen hatalar ise son sütunda verilmiştir. Son sütunda **koyu** ile yazılan kelimeler; üretilen etiketler arasında olmayan ve bu

TABLE I. ÜRETİLEN ÖRNEK ÇIKTILAR VE TESPİT EDİLEN HATALAR

Resim	Üretilen RAY	Üretilen Etiketler	Tespit Edilen Hatalar
	man in black shirt and black pants is standing in front of the wall	man, boy, girl, woman, play, wear, stand, young, sit, blue, shirt, hold, littl, child, street, outsid, smile, hat, orang, next	man in black shirt and black pants is standing in front of the wall
	young boy in red shirt is standing on the grass	man, boy, girl, white, play, stand, young, blue, shirt, hold, littl, child, front, small, children, pink	young boy in red shirt is standing on the grass
	two dogs are playing with ball in the grass	dog, white, black, run, play, stand, jump, water, brown, ball, grass, field, larg, one, air, near, toy, grassi	two dogs are playing with ball in the grass
	man in red shirt is sitting on bench	dog, man, white, black, run, stand, blue, grass, person, field, front, larg, air, rock, near, climb, mountain, tree, background, hill, grassi, path	man in red shirt is sitting on bench
	man with black hair and black shirt is sitting on the street	man, white, black, wear, sit, blue, shirt, hold, look, next, face	man with black hair and black shirt is sitting on the street
	man is standing in front of crowd	man, white, black, woman, wear, stand, peopl, blue, shirt, hold, look, front, group, larg, men, near, street, anoth, smile, hat, next, hand, crowd, glass	man is standing in front of crowd
	dog is running through the grass	dog, white, black, run, play, brown, ball, grass, field, yellow, through, green, mouth, air, toy, catch, grassi, collar	dog is running through the grass

nedenle RAY’da bulunması hata olarak tespit edilen ve resim gözle incelendiğinde de hata olarak teyit edilmiş olan (true negative) kelimelerdir. Son sütunda verilen **koyu ve alt çizgili** kelimeler ise üretilen etiketler arasında olmayan ve bu nedenle RAY’da bulunması hata olarak tahmin edilen ancak resim gözle incelendiğinde hata olmadığı teyit edilmiş olan (false negative) kelimelerdir.

Tablo 1’de üretilen etiketler incelendiğinde etiketlerin yüksek çoğunluğunun resimle ilgili olduğu görülmektedir. Bu sonuç, çalışmanın temelini oluşturan, “RAY’da geçen ancak etiketlerde olmayan kelimelerin hatalı olabileceği” önermesini desteklemektedir.

Tablo 1’de verilen örnek sonuçlar incelendiğinde önerilen sistem ile üretilen RAY’da tespit edilen hataların bir çoğunun doğru olarak tespit edildiği gözlemlenebilir. İlk üç satırda

TABLE II. HATA TESPİT SONUCU

Ölçüt	Göze m	Açıklama
Doğru Negatif sayısı	909	True Negative TN
Yanlış Negatif sayısı	255	False Negative FN
Toplam Tahmin Edilen Negatif sayısı	1164	Predicted Condition Negative
Gerçek Hata Sayısı	1128	Condition Negative
Doğru Negatif oranı (True negative rate TNR)	%80	$\frac{\Sigma \text{True negative}}{\Sigma \text{Condition negative}}$
Hatalı Yanlış oranı (False omission rate)	%22	$\frac{\Sigma \text{False negative}}{\Sigma \text{Predicted Condition negative}}$

üretilen RAY'da hata olarak tespit edilen tüm kelimelerin (**black, pants, front, wall, red, grass, two**) resimler gözle incelendiğinde doğru hata (*true negative*) olduğu görülmektedir. Ancak ikinci ve üçüncü satırlar incelendiğinde ise RAY'da hatalı olan *boy* ve *ball* kelimelerinin hata olarak tespit edilemediği görülmektedir. Dördüncü ve beşinci satırlar incelendiğinde ise hatalı olarak tespit edilen (**sitting, hair**) kelimelerin ise resimler gözle incelendiğinde hatalı olmadığı (*false negative*) görülmektedir. Son iki satırdaki örnekler incelendiğinde de, üretilen RAY'da hiç hata bulunamadığını ve gözle yapılan kontrolde de RAY'ların doğru olduğu görülmektedir.

Yukarıda verilen örnek sonuçlardaki başarımlar gözlemlendikten sonra rastgele seçilen 300 adet resim test veri seti olarak hazırlanmıştır. RAY ve etiketlerin birbirleri ve resim ile karşılaştırılması gözle yapıldığından, mevcut sınırlı zaman ve insan gücü sebebiyle ilk sonuçların alınabilmesi için 300 resim ile yetinilmek durumunda kalmıştır.

Tablo II, test verisi incelendiğinde farklı ölçütler için elde edilen sonuçları özetlemektedir. Önerdiğimiz yöntem sonucu tespit edilen hatalar, üretilen RAY'lardaki hataların %80'ini kapsayabilmektedir (*True negative rate*). Yanlış olduğu tahmin edilen kelimeler içerisindeki hata oranı (*False omission rate*) ise %22 seviyesindedir. Diğer bir deyişle hata olarak tahmin edilen kelimelerin yaklaşık %80'ni doğru olarak tespit edilmiştir.

IV. GELECEK ÇALIŞMA

Yukarıda özetlenen sonuçlar ışığında önerilen sistemin otomatik olarak yaratılan resim alt yazılarında (RAY) oluşan kelime hatalarını %80 seviyede yakalayabildiği görülmektedir. Sunulan çalışmanın bu *ilk ve sınırlı* sonuçları önerilen sistemin başarısı için umut vericidir.

Çalışmanın bu ilk adımından sonra yapılacak iyileştirmelerin başında içinde daha fazla resim ve resim alt yazısı olan veri setlerinin kullanımı, RAY ve etiket üreten derin yapay sinir ağ modellerinin daha da geliştirilmesi, testte kullanılan ve gözle kontrol edilen örnek sayısının eğitimde kullanılan örnek sayısının %10'una tekabül edecek şekilde artırılması bulunmaktadır. Ancak, resimlerin ve çıktıların karşılaştırılması için insan gücü gerekmektedir. Şu an için araştırmacı tarafından sınırlı zaman içinde 300 resim incelenebilmiştir. İleride oluşturulacak gönüllüler ile bu sayı artırıldığında çalışmanın başarımlarını ölçmek için diğer metriklerin (*hatırlama- recall, hassaslık-sensitivity, doğruluk-precision, vb.*) kullanılabilmesi de söz konusu olacaktır. Böylelikle önerilen mimarinin doğruluğunun tespiti için çapraz doğrulama yapılmış olacaktır.

Çalışmanın ileri safhalarında ise RAY ve etiket üreten derin sinir ağlarının birbirlerine tam olarak entegre edilerek üretilen RAY'larda minimum sayıda hata yapacak yeni bir derin öğrenme modelinin geliştirilmesi hedeflenmektedir.

KAYNAKÇA

- [1] Xinwei He, Yang Yang, Baoguang Shi, Xiang Bai, VD-SAN: Visual-Densely Semantic Attention Network for Image Caption Generation, *Neurocomputing*, Volume 328, 2019, Pages 48-55, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2018.02.106>.
- [2] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: a neural image caption generator, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156-3164.
- [3] Jing Zhang, Kangkang Li, Zhe Wang, Parallel-fusion LSTM with synchronous semantic and visual information for image captioning, *Journal of Visual Communication and Image Representation*, Volume 75, 2021, 103044, ISSN 1047-3203, <https://doi.org/10.1016/j.jvcir.2021.103044>.
- [4] Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell, neural image caption generation with visual attention, in: *IEEE International Conference on Machine Learning*, pp. 2048-2057.
- [5] Zhou Yu, Nanjia Han, Accelerated masked transformer for dense video captioning, *Neurocomputing*, Volume 445, 2021, Pages 72-80, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2021.03.026>.
- [6] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, Cornelia Fermüller, Image Understanding using vision and reasoning through Scene Description Graph, *Computer Vision and Image Understanding* 173 (2018) 33-4534.
- [7] Omnicore web site, <https://www.omnicoreagency.com/instagram-statistics/#:~:text=More%20than%2050%20billion%20photos,Location%20Get%2079%25%20more%20engagement> 27 Mayıs 2021.
- [8] Fu, J., and Rui, Y. "Advances in Deep Learning Approaches for Image Tagging", *APSIPA Transactions on Signal and Information Processing*, 6, 2017.
- [9] Wang, M., Ni, B., Hua, X., "Chua, T.-S.: Assistive tagging: a survey of multimedia tagging with human-computer joint exploration", *ACM Comput. Surv.*, 44 (4), 25:1-25:24, 2012.
- [10] Üstümkök, T., Acar, O. C., and Karakaya, M. "Image Tag Refinement with Self Organizing Maps", *2019 1st International Informatics and Software Engineering Conference (UBYMK)*, 348-353, IEEE, 2019.
- [11] Bai, S., and An, S. "A survey on automatic image caption generation", *Neurocomputing*, 311, 291-304, 2018.
- [12] Hu, B., Lu, Z., Li, H., and Chen, Q., "Convolutional neural network architectures for matching natural language sentences", *Proceedings of the Twenty Seventh International Conference on Neural Information Processing Systems*, 2042-2050, 2014.
- [13] Kalchbrenner, N., Grefenstette, E., Blunsom, P., "A convolutional neural network for modelling sentences", *arXiv: 1404.2188v1*, 2014.
- [14] Wang, X., Zhang, L., Liu, M., Li, Y., Ma, W., "Arista - image search to annotation on billions of web photos", *Conf. on Computer Vision and Pattern Recognition*, 2987-2994, 2010.
- [15] Fu, J., Wang, J., Rui, Y., Wang, X., Mei, T., Lu, H., "Image tag refinement with view-dependent concept representations", *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE, 2014.
- [16] LeCun, Y., Bengio, Y., Hinton, G., "Deep learning", *Nature*, 521 (7553) 436-444, 2015.
- [17] Brownlee, J., How to Develop a Deep Learning Photo Caption Generator from Scratch, *Machine Learning Mastery*, <https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/>, 2019.
- [18] Rashtchian, C., Young, P., Hodosh, M., and J. Hockenmaier. "Collecting image annotations using amazon's mechanical turk", *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, p. 139-147, 2010.
- [19] Simonyan, K., and Zisserman, A. "Very deep convolutional networks for large-scale image recognition", *Int. Conf. on Learning Representations*, 2015..