TEKRARLAYAN SİNİR AĞLARI & TRANSFORMERS İLE TÜRKÇE GÖRÜNTÜ ALTYAZILAMA

Berkay Mayalı

ÖZETÇE

Görüntüler üzerinde altyazı oluşturma, görüntüde bulunan nesnelerin ve bunların çevreyle olan ilişkilerinin doğal bir dil ile anlamlı açıklamalarının otomatik olarak üretilmesini amaçlar. Son zamanlarda artan ilgiyle birlikte önemli bir alan olmuştur. Bilgisayarlı görü ve doğal dil işleme alanlarını birleştirmektedir. Makine çevirisi alanındaki gelişmelerden sonra bu alanda başarılı sonuçlar veren kodlayıcı-kod çözücü tekniği, özellikle İngilizce için otomatik görüntü altyazısı oluşturma konusunda kullanılan yöntemlerden biridir. Bu çalışmada ise, Türkçe dili için otomatik görüntü altyazısı oluşturan iki farklı kod çözücü model sunulmaktadır. Bu çalışma, verilen görüntülerin özniteliklerini çıkaran, evrişimsel sinir ağına sahip bir kodlayıcıyı, altyazı oluşturan, tekrarlayan sinir ağı ve transformer mimarisine sahip iki farklı kod çözücü ile birleştirerek, Türkçe MS-COCO veri kümesi üzerinde iki farklı kodlayıcı-kod çözücü modelini test etmektedir. Modellerin performansları oldukça sınırlı ve gürültülü bir veride görece iyi sonuçlar vermektedir.

Anahtar kelimeler— Görüntü altyazılama, evrişimli sinir ağları, tekrarlayan sinir ağları, transformer, bilgisayarlı görü, doğal dil işleme

1. GİRİŞ

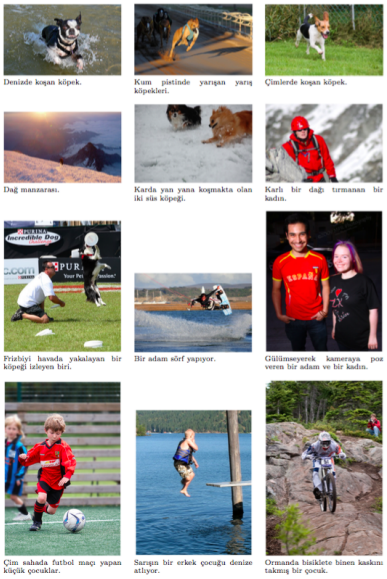
Görüntü altyazılama, verilen görüntüler için metin bazlı bir açıklama oluşturma işlemidir. Derin öğrenme alanında çok önemli ve temel bir görevi çözmek için sunulmuştur. Görme yeteneği az olan veya hiç görmeyen kişilere yardımcı olacak bir uygulama tasarlamak amacıyla farklı yaklaşımlar geliştirilmiştir. İnternetteki görüntülerin hızlı büyümesi de görüntü altyazılamaya artan bir talep yaratmıştır. Ayrıca bir görüntünün doğal dil tanımının otomatik olarak oluşturulması, iki ana yapay zeka alanını birbirine bağlaması nedeniyle oldukça ilgi çekicidir. Yalnızca bir görüntünün içeriği tek başına bir görüntü altyazısı oluşturmak için yeterli değildir, bunun yanı sıra nesneler arasındaki anlamsal ilişkileri eylemleriyle birleştirmek ve sonunda dilsel bir model tarafından üretilen bir görüntü altyazısı oluşturmak önemlidir. Ancak bilgisayarlar için son derece zor bir görevdir. Literatürde görüntü altyazılama problemi için çeşitli yaklaşımlar önerilmiştir. Piksel dizisi olarak kabul edilen görüntüleri bir sözcük dizisine dönüştürmek amaçlandığından, yakın zamanda oluşturulan çalışmalar uçtan uca bir sinir ağı geliştirilmesini benimsemiştir. Bunun için Karpathy, doğal dil işleme kısmında tekrarlayan sinir ağlarını, bilgisayarlı görü kısmı için ise sırasıyla öznitelik vektörlerini elde etmek için evrişimli sinir ağlarını kullanmayı önerdi [1]. Ayrıca evrişimli sinir ağlarının görüntü özniteliklerini sabit uzunlukta bir vektör ile temsil edebildiğine dikkat çekti. Vinyals ve ekibi, görüntü altyazılamada en iyi tahmini elde etmek için Karpathy'nin mimarisine ek olarak bir ışın arama yöntemi kullandılar. Bu yöntem ile her adımda k adet olası en iyi cümle adayları bulunur ve bu cümlelerden en muhtemel kelime adayları seçilir. Bu sayede başarıyı kayda değer oranda arttırmışlardır [2]. Altyazı oluşturma sırasında genellikle görüntüler oldukça karmaşık ve zorludur. Microsoft bu gibi durumlarda, güven puanı mekanizması kullanarak, oluşturulan altyazılara ne kadar güvenebileceğimizi gösteren bir güven puanı tahmin sınıflandırıcısı ek katmanı eklemişlerdir [3]. You ve arkadaşları, [4] çalışmalarında görüntü altyazılarını geliştirmek için anlamsal dikkat mekanizması kullanmışlardır. Yao ve arkadaşları ise hem görüntü temsili hem de üst düzey özellikleri kullanan uzun-kısa süreli bellek mimarisi önermişlerdir [5].

Bu problem oldukça büyük veri kümeleriyle görece iyi sonuçlar verebilir. Fakat görüntülere altyazı/açıklama eklemek büyük miktarda emek ve zaman gerektirir. Ayrıca yukarıda belirtilen çalışmaların tümünde İngilizce dili üzerine odaklanılmıştır. Türkçe dili için Ünal ve arkadaşları tarafından görüntü altyazısı veri kümesi sunulmuştur [6]. TasvirEt veri kümesi olarak adlandırılmıştır. TasvirEt'in oluşturulması oldukça sınırlı koşullarda gerçekleştirildiği için, oluşturulan görüntü altyazıları güvenilir olsa da, veri kümesinin boyutu oldukça sınırlıdır. Başka bir çalışma [7] ise MS-COCO [8] ve Flickr30k [9] veri kümelerinin altyazılarını otomatik çeviri (Google Çeviri API) sistemi kullanılarak İngilizce'den Türkçe diline çevirmiştir. Bu veri

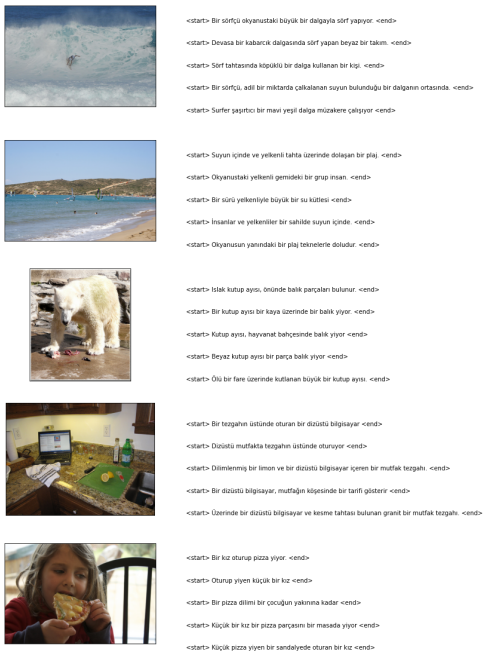
kümesinden Yıldız ve arkadaşları kodlayıcı-kod çözücü mimarisi için yararlandılar. Ayrıca sonuçlarını TasvirEt veri kümesi ile karşılaştırdılar ve önerilen yöntemin BLEU-1 skorunda yaklaşık % 7 artışla TasvirEt'i geride bıraktıklarını göstermişlerdir. Son zamanlarda ise transformer [10] tabanlı mimarilerin doğal dil işleme sistemlerinde son teknoloji performansları göstermesi nedeniyle görüntü altyazılama alanında da popülerlik kazanmıştır. Bu çalışmada, kodlayıcı-kod çözücü mimarisini kullanan iki farklı yöntem Türkçe dili için analiz edilmiştir. Hem görsel hem de sayısal sonuçlar sunulmuştur.

2. VERİ SETİ

Görüntü altyazılama problemi için Türkçe dilinde oluşturulmuş veri kümesi oldukça az bulunmaktadır. Bu problem için önerilmiş olan ilk veri kümesi, TasvirEt veri kümesidir. Bu veri kümesi, Flickr8k veri kümesindeki toplam 8.000 görüntü ve her görüntü için 2 altyazı yani 16.000 adet tasvir edilmiş veri barındırır. Bir diğer veri kümesi ise MS-COCO veri kümesindeki görüntüler için sunulmuş ve onlara ait İngilizce açıklamaların Google Çeviri kullanılarak Türkçe'ye otomatik çevrilmiş halidir. 120.000 farklı görüntü ve her bir görüntüye ait 5 farklı açıklama barındırır. Bu veri kümesi, TasvirEt veri kümesine oranla daha büyük çaplı olmasına karşın otomatik çeviri sisteminin getirdiği dilbiçimsel ve anlamsal açıdan gürültülü altyazı açıklamaları içermektedir.



Şekil 2.1. TasvirEt veri kümesinden örnek görseller ve açıklamaları.



Şekil 2.2. MS-COCO veri kümesinden örnek görseller ve her bir görsel için 5 farklı otomatik çeviri açıklamaları.

MS-COCO veri kümesi büyük ve her bir görsel için 5 farklı açıklama içerdiğinden sözcük türetme konusunda büyük avantaj sağlar. Ancak doğal dil işleme özelinde, bu avantaj aşılması gereken ciddi zorlukları da beraberinde getirir. Türkçe için oluşturulan sözlüğün diğer birçok dile göre daha fazla sayıda kelime içermesi de bir diğer zorluktur. Çünkü daha çok kelime daha yüksek bellek kullanımı ve hesaplama karmaşıklığını beraberinde getirir. Bunun yanında eğitim kümesinde çok az sıklıkla geçen kelimeleri ve yetersiz açıklama barındıran TasvirEt veri kümesi ise öğrenimi kötü yönde etkilemektedir. Çözüm için ise sözlüğün sadece sık kullanılan kelimelerden oluşturulması yoluna gidilmektedir. Ancak Türkçe gibi bir dilde bu yaklaşımın kullanılması birçok kelimenin açıklamalarda kullanılamayacağına dolayısıyla anlamsal açıdan zayıf açıklamalar çıkarılmasına neden olur. Bu tür nedenlerle çalışmada MS-COCO otomatik çeviri veri kümesi kullanılmıştır.

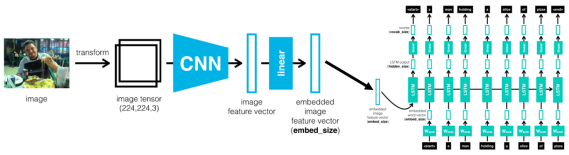
3. MODEL

Klasik görüntü altyazılama modeli, Vinyals ve arkadaşları tarafından 2015'de önerilmiştir. Bu mimari, önceden eğitilmiş bir evrişimli sinir ağının görüntünün sabit bir vektör gösterimini (tanımını) oluşturmak için kodlayıcı olarak kullanıldığı kodlayıcı-kod çözücü mimarisidir. Daha sonra kodlanan görüntü özniteliklerini anlamlandırmak için kod çözücü olarak uzun-kısa süreli bellek (LSTM) kullanılmıştır ve görüntü açıklamaları için her bir sözcük birer birer oluşturulmuştur. Fakat bu model, sözcükleri oluştururken görüntünün tüm gösterimini kullanır ve bu

nedenle farklı sözcükler oluşturmak için görüntünün farklı bölümlerini inceleyemez. Bu soruna çözüm bulmak amacıyla Xu ve arkadaşları dikkat mekanizmalarını kullanarak bir görüntünün bazı önemli kısımlarında nesnelere sahip olduğunu, kelimeleri de bu nesnelere göre oluşturmak için girdi görüntüdeki ilgili kısımlara dikkat edilmesi gerektiğini önermişlerdir [11]. Bu yöntem, insanların bir sahneyi tanımlamak için kullandığı yönteme oldukça benzerdir. Bu çalışmada, ImageNet veri kümesinde önceden eğitilmiş InceptionV3 [12] kodlayıcısı için iki farklı kod çözücü yapısı kullanılmıştır. Bu kod çözücüler tekrarlayan sinir ağları ve transformer modelleridir.

3.1. Tekrarlayan Sinir Ağları Kod Çözücüsü

Tekrarlayan sinir ağlarına sahip kodlayıcı-kod çözücü mimarisi, Sinirsel makine çevirisi (NMT) için etkili bir yaklaşım haline gelmiştir. Bu yaklaşımın temel faydası, tek bir uçtan uca modeli doğrudan kaynak ve hedef cümleler üzerinde eğitme yeteneği ve değişken uzunluktaki giriş ve çıkış dizilerini doğru bir şekilde işleme yeteneğidir. Bu nedenlerle görüntü altyazılama için oldukça kullanışlıdır.

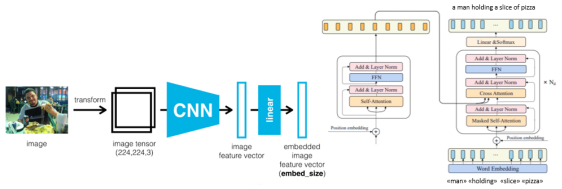


Şekil 3.1.1. Tekrarlayan sinir ağları kod çözücü kullanan sistem mimarisi.

InceptionV3 öznitelik çıkarıcısı, (8 × 8 × 2048) boyutunda alt evrişim katmanından öznitelikleri çıkarır. Bu öznitelikler (64 × 2048) olacak şekilde düzleştirilir. Bu vektör daha sonra tek bir tam bağlı katmandan oluşan evrişimli sinir ağı kodlayıcısından geçirilir. Daha sonra tekrarlayan sinir ağları, bir sonraki kelimeyi tahmin etmek için bu vektörün üzerinde çalıştırılır. Dikkat katmanı, evrişimli sinir ağı tarafından oluşturulan öznitelik haritasına bakar ve her adımda uzun-kısa süreli bellek kod çözücü için hangi kelimenin ilgili olduğuna karar verir. Kod çözücü, önceki adımdan, dikkat mekanizması tarafından oluşturulan vektörü ve önceki adım çıktısında oluşturulan bağlam vektöründeki gizli durumu (hidden state) alır. Yeni gizli durumu güncelleştirmek için bu iki vektörü birleştirir. Bu tekrarlayan sinir ağı kod çözücüsünü kullanan mimari Şekil 3.1.1.'de verilmiştir.

3.2. Transformer Kod Çözücüsü

Transformer tabanlı mimariler, makine çevirisi ve dil-bağlam bütünlüğü anlama gibi sıralı modelleme görevlerinde son teknolojiyi temsil eder. Ancak, resim altyazılama gibi çoklu bağlamlara uygulanabilirlikleri, yeni keşfedilmeye başlanmıştır [13, 14, 15]. Tekrarlayan sinir ağlarından daha hızlı olması, GPU üzerinde paralel olarak çalıştırılabiliyor olması ve daha uzun bağlamlardaki anlam bütünlüğünü koruyabilir olması ve basit konum kodlamasının görüntü altyazılama performansını artırması [16] nedeniyle görüntü altyazılama işlemi için kullanılmıştır.



Şekil 3.2.1. Transformer kod çözücü kullanan sistem mimarisi.

Girdiler ve çıktılar (hedef açıklamalar) ilk olarak n-boyutlu bir uzaya gömülür. Modelin önemli bir parçası, farklı kelimelerin konumsal kodlamasıdır. Dizilimdeki her kelimeye göreceli bir konum verilir. Bu konumlar, her kelimenin gömülü temsiline (n-boyutlu vektör) eklenir. Kod çözücü, maskeli bir çok başlı öz dikkat alt katmanı ve ardından çok başlı bir çapraz dikkat alt katmanı ve sırayla konumsal bir ileri besleme alt katmanı içeren yığılmış özdeş katmanlardan oluşur. Son kod çözücü katmanının çıktı özelliği, çıktı boyutu kelime boyutuna eşit olan doğrusal bir katman aracılığıyla bir sonraki sözcüğü tahmin etmek için kullanılır. Bu transformer kod çözücüsünü kullanan mimari Şekil 3.2.1'de verilmiştir.

4. DENEYSEL SONUÇLAR

Eğitilen iki farklı modelin görüntü altyazılarının başarısını değerlendirmek için çeşitli değerlendirme ölçütleri kullanılmıştır. Temelde iki farklı değerlendirme kategorisi vardır: insan temelli öznel değerlendirme ve otomatik değerlendirme. Otomatik değerlendirme için BLUE, ROUGE_L, METEOR ve CIDEr ölçütleri kullanılmıştır. Tüm bu ölçütler 0 ile 1 arasında ölçülür ve 1'e yakın bir skor elde etmek insan eliyle yapılan altyazılama açıklamalarına yakın bir açıklama olduğu anlamına gelir. Yapılan çalışmaların çoğunda modellerin başarısını objektif bir şekilde değerlendirebilmek için bu ölçütler tercih edilmiştir. Yapılan çalışmada modellerin başarılarını otomatik değerlendirme ölçütleri ile değerlendirmek için 1.200 adet görüntü ve 6.000 adet altyazı/açıklama içeren test veri kümesi oluşturulmuştur. Bu test veri kümesi üzerinde oluşturulan her iki modelin başarısı otomatik değerlendirme ölçütleri kullanılarak değerlendirilmiştir. Tablo 4.1 üzerinde modellerin otomatik değerlendirme ölçütleri için sonuçları verilmiştir. Tablo 4.1'de görüldüğü üzere kod çözücü için transformer kullanan modelin daha başarılı olduğu

belirtilmiştir. Ayrıca bu mimari her bir ölçüt için en iyi sonucu vermiştir. Oluşturulan görüntü tanımlarının kalitesini değerlendiren CIDEr metriği için ise çok yüksek bir farkla transformer kullanan mimarinin daha iyi olduğu görülmektedir. Ölçütlerin sonuçları için BLEU ve CIDEr ölçütleri METEOR ve ROUGE_L ölçütlerine göre daha yüksektir. Bunun nedeni otomatik çeviri ile oluşturulmuş bir veri kümesi kullanıldığından cümlelerin daha uzun bölümlerini inceleyen metriklerin skorlarının daha düşük kalmasıdır.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Transformer	0.171	0.241	0.325	0.355	0.052	0.169	0.176
RNN	0.150	0.024	0.055	0.013	0.012	0.062	0.079

Tablo 4.1. Tekrarlayan sinir ağı modeli ve transformer modeli için otomatik değerlendirme sonuçları.

Şekil 4.2. Tekrarlayan sinir ağı kod çözücü mimarisi için görsel altyazılama sonuçları.

Şekil 4.3. Transformer kod çözücü mimarisi için görsel altyazılama sonuçları.

Şekil 4.2 ve 4.3 üzerinde her iki mimarinin görsel altyazılama sonuçları verilmiştir. Tekrarlayan sinir ağlarını kullanan modelin, görüntülerin içeriğini tanımlamakta yeterli olmadığı, dilbilgisi açısından doğru ve anlamlı cümleler kurmada çoğu zaman yetersiz kaldığı görülmüştür. Görece kısa açıklamalar/altyazılar üretmek için uzun-kısa süreli belleğin karmaşık olabileceği sonucuna varılmıştır. Öte yandan transformer kullanan modelin, görüntülerin içeriğini tanımlamada oldukça iyi sonuçlar verdiği, dilbilgisi açısından görece iyi ve anlamlı cümleler kurduğu, çoğu zaman gerçek referans altyazıdan daha iyi açıklamalar ürettiği görülmüştür.

5. VARGILAR

Bu çalışmada Türkçe görüntü altyazılama problemi için iki farklı mimari incelenmiştir. MS-COCO otomatik çeviri veri kümesi üzerinde tekrarlayan sinir ağlarını ve transformer mimarisini kod çözücü olarak kullanan iki farklı kodlayıcı-kod çözücü modeli değerlendirilmiştir. Deneysel sonuçlar incelendiğinde, transformer kullanan modelin tekrarlayan sinir ağlarını kullanan modele göre ölçülen tüm metriklerde daha iyi sonuçlar verdiği gözlenmiştir. Transformer kullanan mimarinin Türkçe dilbilgisi kurallarına daha uygun, anlamlı ve çoğu durumda görece iyi altyazılar ürettiği gösterilmiştir.

Gelecek çalışmalarda, transformer mimarisinin konumsal kodlayıcısının bölgesel tabanlı evrişimsel sinir ağları ile birleştirilerek hem görsel kodlayıcı hem de altyazılama için anlamsal kod çözücü olarak kullanıldığı uçtan uca bir sinir ağı tasarlanması planlanmaktadır.

6. REFERANSLAR

[1] Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).

[2] Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).

[3] Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., ... & Sienkiewicz, C. (2016). Rich image captioning in the wild. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 49-56).

[4] You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4651-4659).

[5] Yao, T., Pan, Y., Li, Y., Qiu, Z., & Mei, T. (2017). Boosting image captioning with attributes. In Proceedings of the IEEE international conference on computer vision (pp. 4894-4902).

[6] Unal, M. E., Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N. I., & Cakici, R. (2016). Tasviret: Görüntülerden

otomatik türkçe açıklama olusturma Için bir denektaçı veri kümesi (TasvirEt: A benchmark dataset for automatic Turkish description generation from images). IEEE Sinyal Isleme ve Iletisim Uygulamaları Kurultayı (SIU 2016).

[7] Samet, N., Hiçsönmez, S., Duygulu, P., & Akbas, E. (2017). Görüntü Altyazılama için Otomatik Tercümeyle Egitim Kümesi Olusturulabilir mi? Could We Create A Training Set For Image Captioning Using Automatic Translation?. In 25th Signal Processing and Communications Applications Conference (SIU), Antalya-TR.

[8] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[9] Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., & Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision (pp. 2641-2649).

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[11] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.

[12] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).

[13] He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N. (2020). Image captioning through image transformer. In Proceedings of the Asian Conference on Computer Vision.

[14] Messina, N., Falchi, F., Esuli, A., & Amato, G. (2021, January). Transformer reasoning network for image-text matching and retrieval. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 5222-5229). IEEE.

[15] Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10578-10587).

[16] Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 8928-8937).