# MSVD-Türkçe: Türkçe Video Altyazılama için Geniş Ölçekli Bir Veri Kümesi
# MSVD-Turkish: A Large-Scale Dataset for Video Captioning in Turkish

Begum Citamak, Menekse Kuyu, Aykut Erdem, Erkut Erdem
Bilgisayar Mühendisliği Bölümü, Hacettepe Üniversitesi, Ankara Türkiye
n16221821@cs.hacettepe.edu.tr, menekse.kuyu@hacettepe.edu.tr, {aykut,erkut}@cs.hacettepe.edu.tr

***Özetçe* —Video altyazılama olarak da adlandırılan, videoların doğal dilde açıklamalarının otomatik üretilmesi, yakın zaman önce çalışılmaya başlanan zorlu bir bütünleşik görme ve dil problemidir. Her ne kadar araştırmacılar İngilizce için çok sayıda çözüm önermiş olsalar da, Türkçe video altyazı modellerini eğitmek için uygun veri kümelerinin bulunmamasından dolayı, Türkçe üzerinde henüz bir çalışma ortaya konmuş değildir. Bu eksikliği gidermek için, bu çalışmada MSVD veri kümesindeki İngilizce açıklamaların Türkçe'ye özenle çevrilmesiyle geniş çaplı bir Türkçe denektaşı veri kümesi oluşturulmuştur. Buna ek olarak, zamansal dikkat mekanizmalarına sahip LSTM tabanlı diziden diziye mimarileri içeren çeşitli nöral modeller gerçekleştirimleri yapılmış olup bu güçlü temel yöntemlerin veri kümemiz üzerindeki başarımları gösterilmektedir. Veri kümemizin Türkçe video altyazılama üzerine yapılacak gelecek çalışmalar için iyi bir kaynak oluşturacağını düşünmekteyiz.**

***Anahtar Kelimeler*—*Video altyazılama, bilgisayarla görme, doğal dil işleme, yapay öğrenme.***

***Abstract*—Automatically generating natural language descriptions for videos, aka video captioning, has been recently introduced as a challenging integrated vision and language problem. Although researchers have demonstrated numerous solutions for English, to date there has been no study on Turkish language due to the lack of suitable datasets to train Turkish video captioning models. To tackle this, in this study we construct a large-scale Turkish benchmark dataset by carefully translating English descriptions from MSVD dataset to Turkish. Moreover, we implement several neural models, including LSTM-based sequence-to-sequence architectures with temporal attention mechanisms, and report the performances of these strong baselines on our dataset. We hope that our dataset will serve as a good resource for future efforts on Turkish video captioning.**

***Keywords*—*Video captioning, computer vision, natural language processing, machine learning.***

## I. GIRIŞ

Videoların doğal cümlelerle otomatik olarak tasvir edilmesi, bir başka deyişle video altyazılama problemi son yıllarda araştırılmaya başlanan bir bütünleşik görme ve dil problemidir. Bu problem üzerine giden yakın tarihli çalışmalar incelendiğinde bu çalışmaların büyük oranda İngilizce ile kısıtlı kaldıkları gözlemlenmektedir. Türkçe'nin sondan eklemeli yapısı sebebi ile farklı yaklaşımların uygulanmasını, bu problemin çözümü üzerine oluşturulmuş modellerin yöntemlerinin farklılaşmasını gerektirmektedir. Ancak literatürde herhangi bir Türkçe veri kümesinin olmayışı olası gerçekleşebilecek olan çalışmaların önüne geçmektedir.

Video altyazılama modellerinin öğrenilmesi, ve önerilmiş olan modellerin başarım analizleri belirli sayıda video klibi ve onlara ait bir veya daha fazla cümleden oluşan metinsel açıklamalar içeren çok biçimli veri kümeleri üzerinde gerçekleştirilmektedir. Videoları doğal cümleler ile açıklama işlemi, insan robot ilişkilerinden, görme engelliler için film açıklaması üretimine kadar birçok farklı alanda kullanılmaya başlanmıştır. Derin öğrenme alanındaki gelişmeler ile birlikte, birçok farklı video altyazılama yaklaşımları geliştirilmiştir. Bu yaklaşımların bir çoğu, görüntü dizilerini doğal cümleye çevirmek için yinelemeli sinir ağlarını kullanmaktadır. Video altyazılama problemi, çoğu zaman yapay makine çevrimi problemine benzetilebilir. Video altyazılama otomatik çeviri problemi olarak ele alındığında girdi cümlesine karşılık gelen yapı görüntü dizisi olarak düşünülebilir. Bu sebeple video altyazılama yaklaşımlarında çözücü-kodlayıcı (*encoder-decoder*) yapısına sahip diziden diziye (*sequence-to-sequence*) modeller yaygın olarak kullanılmaktadır. Burada çoğu zaman, evrişimsel (*convolutional*) ve yinelemeli sinir ağları (*recurrent neural network*) görüntüleri kodlamak için birlikte kullanılmakta ve açıklama üretme kısmında ise ayrı bir yinelemeli sinir ağı devreye girerek; doğal dilden bir cümle oluşturulmaktadır. Video altyazılama problemi, hem görüntüyü, hem de cümle yapısında yer alan açıklamaları iyi ve etkin anlamayı gerektirmektedir. Bu yüzden görsel içerik ile doğal cümleler arasındaki bağlantıyı kurmak kritik önem taşımaktadır.

Bu çalışma kapsamında, öncelikli olarak videoların Türkçe doğal cümlelerle otomatik olarak betimlenmesinde denektaşı olarak kullanılabilecek büyük hacimli ve çok kipli bir veri kümesi oluşturulması üzerine gidilmiş ve bu Türkçeye özel toplanılan veri kümesi üzerinde bazı temel video altyazılama modellerinin başarımları araştırılmıştır. Bu amaçla özellikle video altyazılama çalışmalarında çok fazla kullanılmakta olun MSVD (Microsoft Research Video Description Corpus) [1] veri kümesine paralel bir Türkçe veri kümesi oluşturulması amaçlanmıştır. Özellikle belirtmek gerekirse; çalışmamızda, nöral makine çevrimi (*neural machine translation*) sistem-

978-1-5090-6494-6/17/$31.00 ~©~2017 IEEE

lerinden ilhamla zamansal dikkat mekanizması ile donatılan uçtan uca video altyazılama modelleri geliştirilmiştir. Bu modeller, çıktı cümlesi üretilirken, girdi videosunun hangi sahnelerine odaklanılması gerektiğine karar verebilmektedir. Bu tarz nöral dikkat mekanizmaları sayesinde standart video altyazılama sistemlerinin başarılarının gözle görülür bir şekilde arttığı gözlemlenmiştir.

## II. İLGILI ÇALIŞMALAR

Video altyazılama problemi için önerilmiş olan veri kümeleri veri toplama yöntemi bakımından çoğunlukla kitle kaynak yaklaşımları kullanılarak toplanmıştır. Bu veri kümeleri metinsel açıklamaların birkaç saniye süren video klipler için oluşturulmuştur. Oluşturulan veri kümeleri alana özel ve serbest alan olarak iki grupta incelenebilir. İlk grup olan alana özel veri kümeleri, YouCook [2], TACoS [3], TACoS Multi-Level [4] gibi tek bir alandan ve örneğin yemek pişirme gibi tek bir temayla ilgili örnek eylemlerin gerçekleştirildiği YouTube gibi video paylaşım sitelerinden toplanan verilerden oluşmaktadır. MSVD [1], TGIF [5], MSR-VTT [6], M-VAD [7], MPII-Movie Description (MPII-MD) [8], LSMDC [9] gibi ikinci grup veri kümeleri incelendiğinde ise, onların serbestçe toplanmış, herhangi genel bir konu içermeyen veri kümeleri olduğu gözlemlenmektedir. Bu veri kümelerinden M-VAD [7], MPII-M [8], LSMDC [9] ise diğerlerinden profesyonel uzmanlarca oluşturulmuş açıklamalar içermeleri bakımından farklılaşmaktadırlar.

Türkçe özelinde gerçekleştirlen çalışmalar incelenecek olursa, bu alanda video altyazılama için herhangi bir veri kümesi olmamasına karşın benzer bir problem olan Türkçe görüntü altyazılama problemi için Unal vd. [10] tarafından gerçekleştirilen bir çalışmada kitle kaynak yaklaşımı ile TasvirEt adlı bir veri kümesi önerilmiştir. Daha sonrasında Kuyu vd.[11]'nin bu veri kümesini kullanarak gerçekleştirdiği çalışmada altsözcük öğeleri kullanılarak başarılı sonuçlar elde edilebileceği gösterilmiştir. Türkçe görüntü altyazılama üzerine gerçekleştirilen bir diğer çalışma da Samet vd. [12] tarafından gerçekleştirilmiştir. Yazarlar, bu çalışma kapsamında MS-COCO [13] veri kümesindeki İngilizce açıklamaları otomatik tercüme aracı kullanarak Türkçe'ye çevirmişlerdir ve işte bu nedenle bu veri kümesi gürültülü açıklamalar içermektedir.

## III. TÜRKÇE VIDEO ALYAZILAMA İÇIN VERI KÜMESI

MSVD veri kümesindeki [1] videoların İngilizce açıklamalarının karşılıkları olacak şekilde planlanan veri kümemizin oluşturulma aşamasında öncelikli olarak Google Translate uygulamasının sunduğu ücretsiz API hizmetinden yararlanarak videolara ait altyazıların İngilizce'den Türkçe'ye çevrimi sağlanmıştır. Ancak gözlemlerimize göre bu çevirme işlemi sonuçlarının çok da doğru olmadığı, bazı eklerin eksikliği, yanlışlığı ya da bazı İngilizce çok anlamlı kelimelerin Türkçeye çevriminde yanlışlıklar olabildiği gözlemlenmiştir. Çok anlamlı kelimelerin çevriminde yaşanan belirsizliklere bir örnek vermek gerekirse "old" kelimesinin İngilizce'den Türkçeye çevrildiğinde bağlama göre "yaşlı" veya "eski" anlamında kullanılabilmektedir. Bu sebeple otomatik olarak çevrilen cümleler daha sonra gönüllü çevirmenler tarafından yeniden gözden geçirilerek cümlelerin anlamları bozulmadan gerekli düzeltmeler gerçekleştirilmiştir.

Tablo I: MSVD-Türkçe veri kümemizin esas alınan MSVD veri kümesine kıyaslamalı olarak sunulan bazı istatistikleri.

| Veri Kümesi | Video Sayısı | Ortalama Klip Uzunluğu | Cümle Sayısı | Kelime Sayısı | Biricik Kelime |
|---|---|---|---|---|---|
| MSVD | 1970 | 10 sn | 70028 | 607339 | 13010 |
| MSVD-Türkçe | 600 | 9,2 sn | 24135 | 153571 | 5724 |

Esas aldığımız MSVD veri kümesi, YouTube'dan toplanmış ve uzunlukları yaklaşık 10 sn civarında olan toplam 1970 video klip barındırmaktadır ve ortalama olarak klip başına 41 açıklama vardır. Gerçekleştirdiğimiz çalışmalar sonucu oluşturduğumuz veri kümesinde şu ana kadar 600 videonun altyazılarının çevrimi sağlanmıştır. İlgili bazı istatistikler Tablo I'de yer verilmektedir.

## IV. TÜRKÇE VIDEO ALTYAZILAMA MODELLERI

### A. Tek Yönlü Model

Bu çalışmada, Venugopalan vd. [14] tarafından önerilen video altyazılama modeli temel alınmıştır. Bu modeldeki ana fikir, girdi görüntü dizisi ile çıktı kelime dizisi arasındaki bağlantının nöral makine çevrimi problemlerinde sıklıkla kullanılan kodlayıcı-kod çözücü tabanlı bir yapı üzerinden öğrenilmesidir. Kodlama aşamasında evrişimsel sinir ağları kullanılarak girdi video karelerinin özniteliklerinin çıkarılmakta, ardından çıkarılan öznitelikler uzun-kısa süreli bellek (*long short term memory - LSTM*) hücresine beslenmektedir. Çıktı açıklamayı oluşturan sözcük dizileri ise kod çözücü tarafında bulunan bir başka LSTM kullanılarak oluşturulmaktadır. Bu modelin basit bir gösterimi Şekil 1(a)'da verilmiştir.

### B. Nöral Dikkat Mekanizması

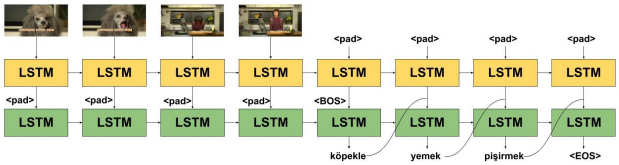
Nöral makine çevrimi modellerinde son yıllarda kullanılmaya başlanan dikkat (*attention*) mekanizması, girdi ve çıktı cümleleri arasındaki özellikle uzak mesafeli bağlantıların kurulmasını kolaylaştırmakta ve bu sayede bir modelin kelime üretme aşamasında girdi cümlesinin ilgili alanlarına odaklanması mümkün olmaktadır. Video altyazılama, bir diziden dizi öğrenme problemi olarak ele alındığında nöral makine çevrimi problemine benzemektedir. Buradan hareketle, bu çalışmada incelediğimiz noktalardan biri yukarıda anlatılan nöral video altyazılama modeline bir kelimeyi üretilirken girdi videosunda en uygun bölümü otomatik olarak seçen bir dikkat mekanizması eklenmesi olmuştur.

Kullanılan bu yapıyı matematiksel olarak daha detaylı açıklamak gerekirse, cümle oluşturma aşamasında kullanılan kod çözücü LSTM mimarisi şu şekilde ifade edilebilir:

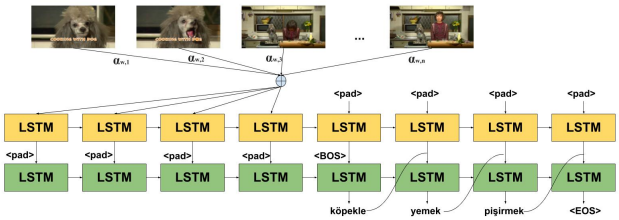
$$h_t = LSTM(w_{t-1}; h_{t-1}; c_{t-1}) \quad (1)$$

Burada $h_t$ bir önceki gizli durumu, $w_{t-1}$ bir önceki kelimeyi ve $c_t$ $t$ zamandaki bağlam vektörünü göstermektedir. Bağlam vektörü $c_t$ aşağıdali denklemdeki formüle göre hesaplanmaktadır:

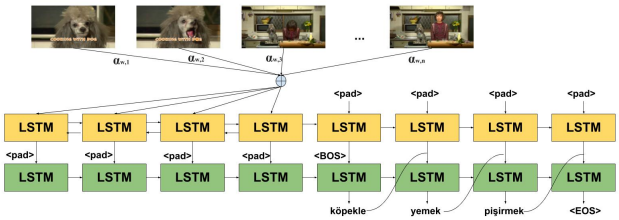
$$c_t = \sum_{i=1}^{N} \alpha_{t,i} h_i \quad (2)$$




(a)




(b)




(c)

Şekil 1: Gerçekleştirimi yapılan LSTM tabanlı nöral video altyazılama modelleri. (a) Tek yönlü model, (b) Tek yönlü nöral dikkat mekanizması ile donatılmış model, (c) Çift yönlü nöral dikkat mekanizmasına sahip model.

$$_{t;i} = align(h_t; h_s) = \frac{e^{a(h_t;h_s)}}{\sum_{s'} e^{a(h_t;h_{s'})}} \quad (3)$$

Denklem 3'te görülebileceği üzere; bağlam vektöründen kastedilen kodlayıcının gizli durumlarının ağırlıklı ortalamasıdır. Burada $_{t;i}$ dikkat ağırlığını temsil etmekte ve $a$ ise gizli durumlar arasındaki ilişkiyi kuran hizalama (*alignment*) skorlarını hesaplayan fonksiyona karşılık gelmektedir. Nöral dikkat mekanızması ile genişletilmiş bu model Şekil 1(b)'de gösterilmektedir.

Deneylerde iki farklı dikkat mekanizması test edilmiştir. Bu dikkat mekanizmaları sadece Denklem 2'deki durum vektörünün hesaplanması açısından farklılık göstermektedir. Bahdanau vd. tarafından önerilen çalışmada [16] bir sonraki adımın gizli durumu ($h_t$), $h_{t\ 1}$'deki gizli durumun, bağlam vektörü $c_t$ ile birleştirilmesi ile şu şekilde oluşturulmaktadır:

$$h_t = LSTM(h_{t\ 1}; [c_t; h_t]) \quad (4)$$

Luong vd. ise alternatif bir formülasyon olarak bağlam vektörü $c_t$'yi $t$ anındaki gizli durum vektörü $h_t$ hesaplandıktan sonra işlemlere katmayı önermiş ve böylece dikkat ağırlıklarını daha sonraki bir aşamada nihai gizli durum vektörünü belirlerken kullanmaktadır [17]:

$$\tilde{h}_t = tanh(W_c[c_t; h_t]) \quad (5)$$


Şekil 2: Zamansal dikkat mekanizması. Farklı renkler ilgili kelime üretilirken bazı video kareleri için hesaplanan dikkat ağırlıklarını göstermektedir.

Bu noktada ele aldığımız problem olan video altyazılamaya geri dönecek olursak bu tarz bir dikkat mekanizması dahil edildiğince ilk olarak video yani görüntü dizisi, kodlayıcıya gönderilmekte ve dikkat ağırlıkları hesaplanmaktadır. Daha sonra, hesaplanan bu ağırlıklar kod çözücüye beslenmekte ve kod çözücü, sözcükleri adım adım oluştururken bu ağırlıkları kullanarak videonun ilgili kısmına odaklanmaktadır. Bu yöntem ile birlikte, görüntü dizisi ve cümlenin arasındaki ilişki daha iyi modellenebilmektedir. Şekil 2'de oluşturulan örnek bir örnek altyazı için video karelerinin belirli kelimeler üretilirken hangi oranda etki ettikleri gösterilmektedir.

### C. Çift Yönlü Model

Geliştirilen dikkat tabanlı uçtan uca video altyazılama modeline yanında gerçekleştirilen bir diğer yaklaşım da kodlayıcı tarafında çift yönlü (*bi-directional*) LSTM [15] kullanılması olmuştur. Çift yönlü LSTM, adını gizli durumların video karelerinin öncelikli olarak zaman ekseninde ileriye doğru, daha sonra da geriye doğru aktarılmasından almaktadır. Bu çift yönlü yapı sayesinde video karelerindeki farklı zamanlar arasındaki gizli durum bilgisi hem ileri yönlü hem de geri yönlü bağlantılar kurarak sahne anlamının sadece bir önceki sahneye değil aynı zamanda bir sonraki sahneye de bağlı olması sağlanarak ilgili bağlam bilgisinin daha etkili bir şekilde kodlanmasını sağlamaktadır. Son olarak uyguladığımız bu çift yönlü LSTM modeli Şekil 1(c)'de gösterilmiştir.

## V. Deneysel Sonuçlar

Model bölümünde anlatılan yöntemlerin oluşturulan veri kümesi üzerindeki başarısını ölçmek için BLEU [18], CIDEr [19], Rouge-L [20] ve METEOR [21] metrikleri kullanılmıştır. Bu modellerin eğitiminde MSVD-Türkçe veri kümesinden rastgele seçilen toplam 480 video (19228 altyazı) öğrenme, 60 video (2266 altyazı) doğrulama amacıyla kullanılmıştır. Yöntemlerin başarıları ise geri kalan 60 video (2641 altyazı) üzerinden gerçekleştirilmiştir. Deneysel sonuçlar Tablo II'de gösterilmektedir. Bu sonuçlara göre, kullanılan iki farklı dikkat mekanizmasının da basit LSTM modelinden daha başarılı olduğu gözlemlenmiştir. Dikkat mekanizması kullanılarak yapılan deneylerde ise Bahdanau dikkat mekanizması Bleu metriğine göre daha başarılı iken Luong dikkat mekanizmasını diğer başarım metriklerinde daha iyi sonuçlar vermiştir. Çift yönlü LSTM'e Luong vd. tarafından önerilmiş dikkat mekanizmasının daha iyi sonuçlar vermesi sebebiyle bu dikkat mekanizması uygulanarak alınan sonuçlarda ise CIDEr metriğinde iyileşme gözlemlenirken diğer metrikler için bir başarım kaybı söz konusudur.