Memiş, A. (2025). Türkçe Görüntü Altyazılama: Mevcut Çalışmalar, Uygulamalar, Veri Setleri, Metrikler ve Gelecekteki Olası Eğilimler Üzerine Bir İnceleme. EMO Bilimsel Dergi, 15(2), 81-97.

Derleme Makalesi
Makale Gönderilme Tarihi: 03.03.2025
Kabul Tarihi: 28.05.2025

Türkçe Görüntü Altyazılama: Mevcut Çalışmalar, Uygulamalar, Veri Setleri, Metrikler ve Gelecekteki Olası Eğilimler Üzerine Bir İnceleme

Turkish Image Captioning: A Review of Current Studies, Applications, Datasets, Metrics and Potential Future Trends

Abbas Memiş
0000-0003-2645-8071
Bilgisayar ve Bilişim Teknolojileri Fakültesi, Bilgisayar Mühendisliği Bölümü
İstanbul Üniversitesi
abbas.memis@istanbul.edu.tr

Özet

Son on yılda bilgisayarla görme ve makine öğrenmesinde kaydedilen sarsıcı gelişmeler, akıllı sistemlerin hızla yaygınlaştığı, çeşitlendiği ve insan hayatına daha çok etki ettiği bir dönemi başlatmıştır. Bu gelişmelerin ivmelendirdiği ve yön verdiği önemli araştırma alanlarından birisi de otomatik görüntü altyazılamadır. Son yıllarda otomatik görüntü altyazılama kapsamında çok farklı diller için kaydedeğer seviyede çalışmalar yapılmakla birlikte, Türkçe görüntü altyazılamada da önemli ve umut verici gelişmelerin yaşandığı görülmektedir. Sunulan makalede, Türkçe görüntü altyazılamadaki bu önemli ve umut verici gelişmelerin ışığında, Türkçe görüntü altyazılama özelindeki mevcut çalışmaları ve uygulamaları ele alan kapsamlı bir inceleme çalışması raporlanmıştır. Çalışma kapsamında, güncel literatürdeki çalışmalara ek olarak, uygulama alanları ve veri setleri de değerlendirilmiştir. Makalede ayrıca, görüntü altyazılama sistemlerinin altyazı oluşturma başarımlarını ölçmek amacıyla kullanılan genel ve standart metriklere de yer verilmiştir. Bununla birlikte, Türkçe görüntü altyazılama için gelecekteki olası eğilimler ve potansiyel gelişmeler makale yazarlarının perspektiflerinden değerlendirilmiş ve paylaşılmıştır.

Anahtar kelimeler: *Görüntü altyazılama, Türkçe görüntü altyazılama, görüntü altyazılama alanları, görüntü-altyazı veri setleri, görüntü altyazılama metrikleri.*

Abstract

Over the past decade, groundbreaking advances in computer vision and machine learning have led to an era in which intelligent systems have become more widespread, diverse and impactful on human life. One of the major research areas accelerated and driven by these advances has been automatic image captioning. In recent years, not only a remarkable amount of work on automatic image captioning has been proposed for a wide range of languages, but there have also been significant and promising advances in Turkish image captioning. In the context of these notable and promising achievements in Turkish image captioning, a comprehensive review of existing studies and applications in the field of Turkish image captioning is reported in this article. In addition to current studies in the literature, application domains and datasets are also covered. The paper also includes common and standard metrics used to measure the captioning performance of image captioning systems. Furthermore, possible future trends and potential developments for Turkish captioning are discussed from the authors' perspective and shared.

Keywords: *Image captioning, Turkish image captioning, image captioning domains, image-caption datasets, image captioning metrics.*

1. Giriş

Son birkaç on yılda, elektronik ve bilgisayar teknolojilerindeki ilerlemeler ve özellikle; yapay zekâ, bilgisayarla görme, görüntü işleme, makine öğrenmesi, derin öğrenme ve doğal dil işleme alanlarında yürütülen araştırmalar, akıllı sistemlerin gelişimine oldukça büyük bir katkı sağlamıştır. Bu gelişmelerin neticesinde ise, bugün akıllı sistemlerin seviye atlayarak hızla yaygınlaştığı, çeşitlendiği ve insan hayatına daha çok etki etmeye başladığı bir dönemden geçilmektedir. Akıllı sistemlerin son yıllardaki gelişiminde ve bu sistemlerin göstermekte olduğu başarıda, ImageNet 2012 Challenge'ta büyük bir etki yaratmış olan Evrişimsel Sinir Ağı (Convolutional Neural Network, CNN) AlexNet [1] ile başlayan derin öğrenme süreci çok önemli bir rol oynamıştır. Derin öğrenme sayesinde, geleneksel makine öğrenmesi yaklaşımları ile sağlanan başarımlar domine edilerek oldukça üst seviyelere taşınmış ve neredeyse uzman bir insan kadar etkili akıllı sistemlerin geliştirilebilmesinin önü açılabilmiştir. Son yıllarda, ulaşım [2] ve tıptan [3] finansa [4], eğlenceye [5], çeviriye [6], medyaya [7], pazarlamaya [8], robotiğe [9] ve siber güvenliğe [10] kadar geniş bir yelpazede derin öğrenme tabanlı yaklaşımlar önerilmektedir.

Derin öğrenmede önemli gelişmelerin yaşandığı araştırma alanlarından birisi de bilgisayarla görme ve görüntü işleme olmuştur [11, 12]. Son yıllarda geliştirilen derin modeller sayesinde, çeşitli görüntüleme sistemlerinden elde edilen görüntüler ve videolar üzerinde tespit [13], takip [14], bölütleme [15], tanıma [16], anlama [17] gibi başlıca bilgisayarla görme işlemleri geleneksel yöntemlere nazaran

daha başarılı gerçekleştirilebilir hâle gelmiştir. Bilgisayarla görme kapsamında üzerine araştırma yürütülen önemli ve zorlu araştırma konularından bir diğeri ise görüntü altyazılamadır [18]. Görüntü altyazılama, temel olarak görüntülerin bilgisayar sistemleri tarafından otomatik bir biçimde yorumlanması ve görüntülerin içeriğine dair görsel bilgilerin metinsel formda ifade edilmesi olarak tanımlanır [18]. Otomatik görüntü altyazılama [19-21] sayesinde, görüntü içerikleri akıllı sistemler tarafından anlaşılabilmekte [22], ve bu içerik bilgisi ilgili sistemlerin çeşitli dâhili ve harici mekanizmalarında kullanılmaktadır.

Her ne kadar güncel literatürde görüntü altyazılama üzerine kapsamlı derleme ve inceleme çalışmaları [18, 23-27] yapılmış olsa da, bu alandaki çalışmaların kapsamları güncel teknolojiler ve uygulama alanları bağlamında farklılıklar gösterebilmekte ve kaydedilen hızlı gelişmelere bağlı olarak önerilen yeni yöntemleri, ortaya çıkan yeni uygulama sahalarını ve mevcut durumu ele almada belirli ölçüde yetersiz kalabilmektedir. Bununla birlikte, bilindiği kadarıyla, güncel literatürde Türkçe görüntü altyazılama üzerine yapılmış genel bir inceleme ve derleme çalışması raporlanmamıştır. Ayrıca, Türkçe görüntü altyazılama geçmişinin çok uzun olmaması ve bu alanda yapılmış çalışmaların da kısıtlı olması, bu alanda çalışma, tez veya proje üretmek isteyen araştırmacılar için zorluk teşkil edebilmektedir. Son yıllarda otomatik görüntü altyazılama kapsamında çok farklı diller için kaydadeğer seviyede çalışmalar yapılmakla birlikte, Türkçe görüntü altyazılamada da önemli ve umut verici gelişmelerin yaşandığı görülmektedir. Sunulan makalede, Türkçe görüntü altyazılamadaki bu önemli ve umut verici gelişmelerin ışığında, Türkçe görüntü altyazılama özelindeki mevcut çalışmaları ve uygulamaları ele alan kapsamlı bir inceleme çalışması raporlanmıştır. Çalışma kapsamında, güncel literatürdeki çalışmalara ek olarak, uygulama alanları ve veri setleri de değerlendirilmiştir. Makalede ayrıca, görüntü altyazılama sistemlerinin altyazı oluşturma başarımlarını ölçmek amacıyla kullanılan genel metriklere de yer verilmiştir. Bununla birlikte yazarlar, Türkçe görüntü altyazılama için gelecekteki olası eğilimleri ve potansiyel gelişmeleri de kendi perspektiflerinden değerlendirmiş ve bu değerlendirmelerini paylaşmışlardır.

Makalenin ilerleyen bölümleri şu şekilde organize edilmiştir: Makalenin 2. bölümünde, otomatik görüntü altyazılama kapsamında; genel görüntü altyazılama yaklaşımlarına, görüntü altyazılama veri setlerine, görüntü altyazılama metriklerine ve görüntü altyazılamanın genel uygulama alanlarına yer verilmiştir. Bölüm 3'te ise, özel olarak otomatik Türkçe görüntü altyazılama ele alınmıştır. Bu bağlamda, Türkçe görüntü altyazılamada genel durum, mevcut çalışmalar, bu çalışmaların dâhil olduğu uygulama alanları, Türkçe görüntü-altyazı veri setleri ve Türkçe görüntü altyazılamanın geleceği ile olası eğilimler tartışılmıştır. Makalenin son bölümü olan Bölüm 4'te ise, sonuçlar ve genel değerlendirmeler paylaşılmıştır.

2. Otomatik Görüntü Altyazılama

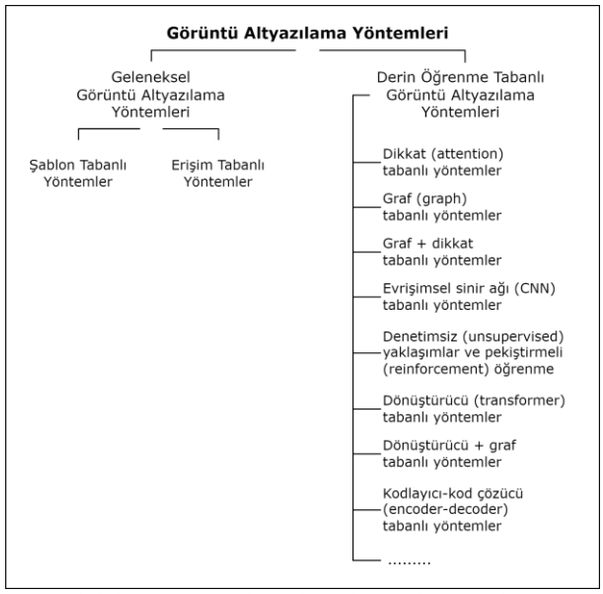
Görüntü altyazılama, en sade biçimiyle, bir resmin içeriğinin kelimelerle tanımlanması, tasvir edilmesi olarak ifade edilebilir [18]. Bir görüntü altyazılama sistemi genel yapısı itibariyle, girdi olarak aldığı her bir görüntü için, referans olarak kullandığı bir görüntü-altyazı seti üzerinden bir altyazı üretir. Üretilen bu altyazıların hem biçimsel (syntactic) hem de anlamsal (semantic) olarak doğru olması beklenir [26]. Bu işlem temel olarak bir bilgisayarla görme görevi olmasına rağmen, metin oluşturma ve işleme de içerdiğinden belirli bir ölçüde doğal dil işlemenin de kapsamına girer. Dolayısıyla otomatik görüntü altyazılama, her iki araştırma alanının da kesişiminde yer alan spesifik araştırma konularından birisidir. Görüntü altyazılama, öne çıkan ve popüler bazı uygulama alanlarında (uzaktan algılama [28-30], sosyal medya [31-33], haber [34-36], tıp [37-39], turizm [40-42], sanat ve kültür [43-45], kimya ve biyoloji [46-48]) kullanılabiliyor olmasının yanında, doğası gereği birçok alanda ve çok farklı amaçlara yönelik olarak uygulanabilme potansiyeline de sahiptir [49]. Görüntü altyazılama probleminin çözümü için güncel literatürde, farklı kategorilerde sınıflandırılabilecek birçok yöntem önerilmiştir. Bununla birlikte, bu alandaki araştırma çalışmalarına destek vermek amacıyla çeşitli araştırma grupları tarafından farklı veri setleri de oluşturulmuş ve paylaşılmıştır. Aşağıdaki alt başlıklarda, görüntü altyazılamada kullanılan genel yaklaşımlar ile öne çıkan veri setleri ele alınmıştır. Ayrıca, görüntü altyazılama modellerinin altyazı üretimindeki başarımını değerlendirmek amacıyla kullanılan çeşitli metriklere de yer verilmiştir. Son kısımda ise görüntü altyazılamanın uygulama alanlarından bahsedilmiş ve bu alanlarda öne çıkan bazı spesifik örnekler tartışılmıştır.

2.1. Görüntü Altyazılamada Genel Yaklaşımlar

Görüntü altyazılama için geçmişten günümüze doğru önerilen yöntem ve yaklaşımları: 1) Geleneksel görüntü altyazılama yöntemleri ve 2) Derin öğrenme tabanlı görüntü altyazılama yöntemleri olmak üzere kabaca iki ana başlık altında kategorize etmek mümkündür [25]. Derin öğrenme öncesi dönemde önerilen yöntemler daha çok geleneksel yöntemler olarak sınıflandırılmakta olup, günümüz görüntü altyazılama sistemlerine nazaran daha basit ve ilkel bir yapıdadır. Bu kategoride yer alan yöntemler ise genel olarak şablon tabanlı (template-based) ve erişim tabanlı (retrieval-based) yöntemlerdir. Geleneksel yöntemler her ne kadar performans açısından belirli bir seviyede kabul görmüş olsa da açık problemlere sahiptir [50] ve otomatik görüntü altyazılamada yetersiz kalabilmektedir. Geleneksel yöntemlerle ilgili bu bağlamdaki hususlar ilerleyen satırlarda daha detaylı olarak ele alınmıştır. Derin öğrenmenin doğuşu ve son on yıldaki hızlı yükselişi birçok konuda ve araştırma alanında olduğu gibi görüntü altyazılamada da önemli gelişmeleri beraberinde getirmiştir. Bu sayede, geleneksel yöntemlerdeki bazı önemli eksikliklerin giderilebilmesinin önü açılmış, görüntü altyazılama için domine edici başarımlar gözlemlenmeye ve geleneksel yöntemler yerini derin modellere bırakmaya başlamıştır [50]. Derin öğrenme tabanlı yöntemler dâhilinde ise; CNN tabanlı, graf tabanlı, dönüştürücü (transformer) tabanlı, dikkat (attention) mekanizması tabanlı vb. olmak üzere çok çeşitli yaklaşımlar önerilmektedir. Şekil 1'de, görüntü altyazılama için kullanılan yöntem ve yaklaşımların kategorizasyonunu gösteren genel bir şema sunulmuştur.

2.1.1. Geleneksel Görüntü Altyazılama Yöntemleri

Şekil 1'de de görsel olarak ifade edildiği üzere, geleneksel görüntü altyazılama yöntemleri: 1) Şablon tabanlı ve 2) Erişim tabanlı yöntemler olmak üzere iki ana gruba ayrılır [51]. Şablon tabanlı yaklaşımlarda genel olarak önceden birtakım



Şekil 1: Görüntü altyazılamada genel yöntemler.

biçimsel kurallar tanımlanır ve buna bağlı olarak çeşitli altyazı şablonları belirlenir [52]. Görüntü altyazıları da bu şablonlara göre oluşturulur. Şablon tabanlı yaklaşımlarda öncelikle bir görüntüdeki içeriğe dair görsel bileşenlerin/kavramların (nesneler, bu nesnelerin özellikleri, edatlar ve görüntüdeki aksiyonlar yani fiiller vb. gibi) tespit edilmesi gerekir. Sonrasında ise bu bilgiler, altyazı şablonlarındaki önceden tanımlanmış boşluklara yerleştirilir ve bu sayede görüntü altyazıları oluşturulur. Şablon tabanlı yöntemler ile başarılı bir görüntü altyazılama gerçekleştirebilmek için bazı faktörlerin dikkate alınması gerekir. Bu faktörler sırasıyla; şablon oluşturma, görsel öğelerin ve kavramların tespiti ve bunlara göre anlamlı altyazı metinlerinin oluşturulması ile ilgilidir. Sistem tarafından oluşturulacak olan altyazının anlamsal bağlamdaki başarısında ise nesnelerin, nesne özelliklerinin ve görüntüdeki aksiyonların tespiti ayrıca önemlidir. Bu bakımdan, şablon tabanlı bir görüntü altyazılama modelinde nesnelerin, nesne özelliklerinin ve aksiyonların yüksek doğrulukla tespiti ve nesneler arası edatsal bilgilerin yüksek doğrulukla çıkarılması gerekir. Görüntülerdeki aksiyonların belirlenmesi noktasında, belirli nesne veya nesne grupları ile ilişkileri içeren ön-tanımlı bir nesne-aksiyon sözlüğü de tercih edilebilmektedir [52]. Ancak, şablon tabanlı yöntemler ile her ne kadar biçimsel açıdan hatasız görüntü altyazıları oluşturulabilse de, anlamsal açıdan her zaman başarılı altyazılar oluşturulamayabilir. Şekil 2'de, genel bir şablon tabanlı görüntü altyazılama modeli resmedilmiştir.

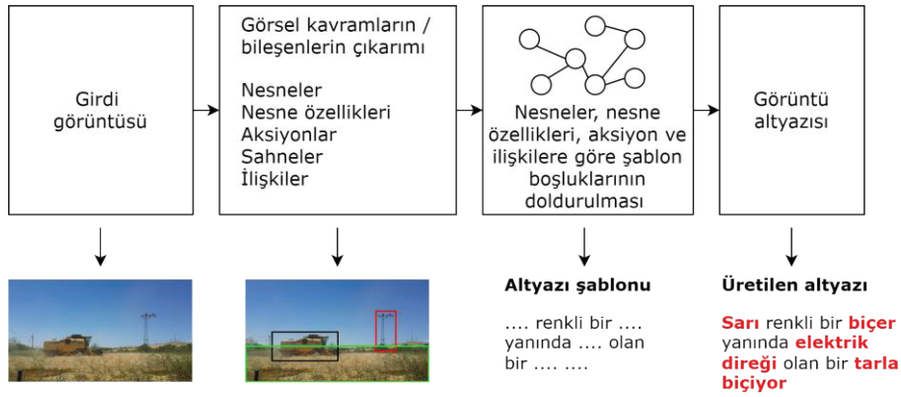
Geleneksel görüntü altyazılama teknikleri dâhilindeki ikinci yaklaşım ise erişim tabanlı görüntü altyazılama yaklaşımıdır. Bu yaklaşım, bilgisayarla görmede ele alınan görüntü erişimi (image retrieval) görevi ile oldukça benzer bir işlem mekanizmasına sahiptir. Erişim tabanlı görüntü altyazılama yöntemlerinde genel işleyiş temel olarak şu şekildedir: Altyazısı üretilecek bir görüntü (ki bu görüntü genelde sorgu görüntüsü olarak adlandırılır) mevcut bir görüntü veri setinde sorgulanır ve bu sorgu görüntüsüne içerik olarak en çok benzeyen görüntüler (aday görüntüler) belirlenir [51, 53]. Devamında ise, aday görüntülere ait önceden tanımlanmış görüntü altyazıları referans alınarak sorgu görüntüsü için bir altyazı üretilir [53]. Üretilen görüntü altyazısı, doğrudan altyazı setindeki bir cümle olabileceği gibi, birden fazla altyazının birleşiminden elde edilen bir metin de olabilir [51]. Şekil 3'te, erişim tabanlı görüntü altyazılama modelleri için genel bir sistem yapısı görsel olarak ifade edilmiştir. Erişim tabanlı görüntü altyazılama yöntemlerinin, belirli bir seviyede başarılı altyazılar üretmesine rağmen, şablon tabanlı yöntemlerde olduğu gibi bazı kısıtları da bulunmaktadır. Bu tarz metotlar, biçimsel olarak doğru altyazılar üretebilmesine rağmen anlamsal olarak doğru görüntü altyazıları oluşturmada yetersiz kalabilirler [51]. Bununla birlikte, üretilecek olan görüntü altyazısının biçimsel ve anlamsal doğruluğu korpus olarak kullanılan görüntü altyazısı setine de doğrudan bağlıdır. Örneğin, eğitim setinin (korpus altyazı setinin) çok küçük ve anlamsal varyansının düşük olması, üretilecek olan altyazıların hem biçimsel hem de anlamsal olarak birbirine benzerliğini arttırır. İlaveten, yine aynı kısıta bağlı olarak, veritabanında sorgu görüntüsüne benzer görüntülere erişim olasılığı azalacağından, sorgu görüntüsü için çok anlamlı olmayan aday görüntüler seçilebilir. Bu durum ise, sorgu görüntüsü için oluşturulacak olan altyazının başarısına doğrudan etki eder ve görüntü ile uyum seviyesi düşük altyazıların üretilebilmesine sebep olur. Bahsi geçen hususlar sebebiyle, erişim tabanlı görüntü altyazılama için veri setinin büyüklüğü ile altyazı korpusunun anlamsal çeşitliliği oldukça önem arz eder.

2.1.2. Derin Öğrenme Tabanlı Görüntü Altyazılama Yöntemleri

Derin öğrenme birçok bilgisayarla görme uygulamasında olduğu gibi görüntü altyazılamada da önemli gelişmeleri tetiklemiştir [27, 50]. Derin öğrenme tabanlı yaklaşımlar sayesinde, sistemlerin görüntü altyazılamadaki başarımlarının artmasına ve daha anlamlı görüntü altyazıları üretilmesine ek olarak, geleneksel görüntü altyazılama yöntemlerindeki önemli eksikliklerin de üstesinden gelinebilmesinin önü açılmıştır. Derin öğrenmenin çok uzun bir geçmişi olmamasına rağmen, görüntü altyazılama konusunda derin öğrenme teknikleri kullanan, özellikle son yıllarda, kaydadeğer miktarda çalışma yapılmaktadır [54] ve bu bağlamda derin öğrenme tabanlı çok çeşitli görüntü altyazılama mimarileri [55-68] sunulmaktadır.

Şekil 1'de, görüntü altyazılama için kullanılan yöntem ve yaklaşımların kategorizasyonunu gösteren genel bir şemaya yer verilmişti. İlgili şekil üzerinden de takip edilebileceği üzere, derin öğrenme tabanlı görüntü altyazılama kapsamında; dikkat (attention) mekanizması tabanlı [57], graf tabanlı [60], CNN tabanlı [55], denetimsiz (unsupervised) öğrenme tabanlı [66], pekiştirmeli (reinforcement) öğrenme tabanlı [67], dönüştürücü (transformer) tabanlı [69], kodlayıcı-kod çözücü (encoder-decoder) tabanlı [62] vb. olmak üzere çok farklı yaklaşımlar önerilmiştir. Bununla birlikte, Yinelemeli Sinir Ağı (Recurrent Neural Network, RNN) [70, 71], Uzun Kısa-Süreli Bellek (Long Short-Term Memory, LSTM [72, 73] ve Geçitli Tekrarlayan Birim (Gated Recurrent Unit, GRU) [74] gibi yapıların kullanımına da yaygın olarak rastlanmaktadır.

Derin öğrenme tabanlı yaklaşımların geleneksel görüntü altyazılamaya nazaran daha başarılı olmasındaki temel faktör, normale göre çok daha fazla, çeşitli ve ayırt edici özelliği çıkarmaya imkân verebilmesidir. Yakın geçmişte CNN tabanlı



Şekil 2: Şablon tabanlı görüntü altyazılama sistemlerinde genel yapı.

Şekil 3: Erişim tabanlı görüntü altyazılama sistemlerinde genel yapı.

ve kodlayıcı-kod çözücü tabanlı yöntemler ile görece yüksek başarımlara ulaşılmıştır [75]. Bu kategorilerdeki yöntemler, görüntülerdeki birincil ilgi bölgelerine ve esas aksiyonlara odaklanmadan genel olarak görüntüleri bir bütün olarak ele alıp özellik çıkarımı gerçekleştirmektedir. Bu nedenle, görüntüleri tasvir edecek olan önemli özellikler ile, görüntüyle ilintili olan ancak altyazılama için çok da değerli olmayan özellikler birlikte değerlendirilmektedir. Bu durumun altyazılama başarımı açısından ortaya çıkaracağı zayıflıkları gidermek amacıyla, görüntülerdeki önemli ve öne çıkan ilgi bölgelerini, yani görüntüyü daha başarılı tasvir etmeyi sağlayacak olan önemli özellikleri ön plana çıkarmayı amaçlayan dikkat mekanizmaları önerilmiştir [18]. Dikkat mekanizmalarının kodlayıcı-kod çözücü modellerdeki başarısı, metinlerdeki bağlamı ve metnin farklı bölümleri arasındaki ilişkiyi öğrenmeye imkân veren dönüştürücü modellerin de geliştirilmesinin önünü açmıştır. Dönüştürücü modellerdekine benzer bir yaklaşım, görüntüleri çeşitli bölgelere ayıran ve bu bölgeler arasındaki bağlamı ve ilişkiyi öğrenmek için dikkat mekanizmasını kullanan görü dönüştürücülerde de (vision transformer) [76] kullanılmaktadır. Dönüştürücü tabanlı modeller, bahsedilen karakteristik özellikleri sayesinde, LSTM tabanlı klasik kodlayıcı-kod çözücü modellere nazaran daha başarılı görüntü altyazılama gerçekleştirebilmektedir. Derin öğrenme tabanlı bu yöntem ve yaklaşımlara ek olarak, graf tabanlı mekanizmalar ile varlıklar arasındaki etkileşimler de tanımlanabilmekte ve görüntülerdeki zengin anlamsal bilgiler çıkarılabilmektedir [77, 78].

Son yıllarda, derin öğrenme tabanlı yöntem ve yaklaşımların görüntü altyazılama görevindeki kullanım ve uygulamalarına dair çeşitli spesifik inceleme raporları da yayınlanmaktadır [24, 25, 27, 18, 79, 75]. Oldukça kapsamlı olarak hazırlanmış olan bu raporlarda, derin öğrenme tabanlı yöntem ve yaklaşımlar etraflıca ele alınmaktadır. Derin öğrenme tabanlı görüntü altyazılama yöntemlerine dair daha detaylı bilgiler, ilgili derleme ve inceleme raporlarından ve özel araştırma çalışmalarına ait yukarıda paylaşılan kaynaklardan takip edilebilir.

2.2. Görüntü Altyazılama Veri Setleri

Genel-amaçlı görüntü altyazılama konusunda yapılan çalışmalarda gerek sistemlerin performanslarını sınamak, gerekse de önerilen sistemleri mevcut çalışmalarla kıyaslamak amacıyla genellikle iki temel görüntü-altyazı veri setinin kullanımı ön plâna çıkmaktadır: MS COCO [80] ve Flickr [81, 82]. Literatürdeki çalışmalarda sıklıkla tercih edilen bu iki veri seti, görüntü altyazılamada en çok kullanılan denektaşı (benchmark) veri setleri arasında yer almaktadır. İlgili veri setlerinden türetilmiş alt veri kümeleri de bulunmakla birlikte, bu alt veri kümelerinin de yaygın kullanımı göze çarpmaktadır. Örneğin, Flickr görüntü-altyazı veri seti için Flickr8K [81] ve Flickr30K [82] olarak

Tablo 1: Genel-amaçlı görüntü altyazılama için önerilmiş ve kullanıma açık bazı veri setleri.

Veri seti	Kaynak	Yıl	Görüntü sayısı	Görüntü başına altyazı	Altyazı dili
IAPR TC-12	[84]	2006	20,000	1-5	İngilizce, Almanca, İspanyolca
PASCAL1K	[85]	2010	1,000	5	İngilizce
UIUC Pascal	[86]	2010	1,000	5	İngilizce
SBU dataset	[87]	2011	1,000,000	1	İngilizce
Flickr8K	[81]	2013	8,092	5	İngilizce
Flickr30K	[82]	2014	31,783	5	İngilizce
MS COCO	[80]	2014	328,000	5	İngilizce
YFCC100M	[88]	2016	100,000,000	>=1	İngilizce, İspanyolca, Çince, Fransızca vd. [89]
Google Refexp	[90]	2016	26,711	>=1	İngilizce
InstaPIC	[31]	2017	721,176	1	İngilizce
FlickrStyle10K	[91]	2017	10,000	1-5	İngilizce
Google's Conceptual Captions	[92]	2018	≈3,300,000	1	İngilizce
Nocaps	[93]	2019	15,100	11	İngilizce
TextCaps	[94]	2020	28,408	5	İngilizce
Conceptual 12M	[95]	2021	12,000,000	1	İngilizce

isimlendirilmiş ve farklı sayılarda görüntüler içeren alt kümeler tanımlanmıştır. Bununla birlikte, "Karpathy's split" [83] olarak bilinen ve MS COCO, Flickr8K ve Flickr30K veri setlerinden oluşturulmuş bir veri seti de mevcuttur. Andrej Karpathy tarafından tanımlanan bu veri kümesi de oldukça popülerdir. MS COCO ve Flickr ile birlikte, genel-amaçlı görüntü altyazılama için önerilmiş ve genel kullanıma açık diğer bazı veri setlerine ait bilgilere Tablo 1'de yer verilmiştir. Tablo 1'de, bu veri setlerindeki toplam görüntü sayısı, görüntü başına tanımlı altyazı sayısı ve altyazıların yazıldığı dil bilgisi de paylaşılmıştır.

Tablo 1'de verilen ve ekseriyeti İngilizce olan genel görüntü-altyazı veri setleri haricinde; Japonca [96], Arapça [97], Bengalce [98] ve Hintçe [99] gibi diller esas alınarak oluşturulmuş daha özel ve yerel görüntü altyazılama veri seri setleri de mevcuttur. Ancak bu tarz veri setlerinin bir bölümünün (örneğin [97, 99]) Flickr gibi bilinen veri setlerinden otomatik çeviri yoluyla elde edildiği de göz önünde bulundurulmalıdır. Tablo 1'deki genel görüntü-altyazı veri setleri dışında, görüntü altyazılama için önerilmiş SentiCap [100] (duygu tabanlı görüntü altyazılama tabanlı) ve Visual Genome [101] (bölgesel görüntü açıklamaları tabanlı) gibi spesifik veri setleri de mevcuttur. Buna ek olarak, uzaktan algılama görüntüsü altyazılama [102] ve tıbbi görüntü altyazılama [103] gibi alana-özel (domain-specific) görüntü altyazılama çalışmalarında kullanılan veri setleri de mevcut olmakla birlikte, bu tarz veri setleri bu çalışma kapsamında özel olarak ele alınmamıştır. Ayrıca, video görüntü altyazılama [104-106] ve bu amaçla önerilmiş veri setleri de (ör. MSVD [107], MSR-VTT [108], MSVD-S [109] vb.) çalışma kapsamında incelenmemiştir.

2.3. Görüntü Altyazılama Metrikleri

Otomatik görüntü altyazılama sistemleri için başarım ölçümünün, diğer bilgisayarla görme ve görüntü işleme görevlerine nazaran biraz daha zorlu olduğu söylenebilir. Bu durum daha çok, sistemlerin oluşturmakta olduğu altyazıların hem biçimsel hem de anlamsal açıdan doğru olması gerekliliğinden kaynaklanmaktadır. Bu bağlamda, etkin bir görüntü altyazılama sisteminin veya yönteminin, sadece yazım açısından doğru ifadeler üretebilmesi değil, aynı zamanda görüntünün içeriğindeki anlamı da doğru yansıtabilmesi (nesneler arası ilişkiler ve aksiyonlar/fiiler) gerektiği söylenebilir.

Görüntü altyazılama sistemlerinde girdi/sorgu görüntüleri için oluşturulan altyazıların doğruluğunu analiz etmek için niteliksel (qualitative) ve niceliksel (quantitative) olmak üzere iki temel değerlendirme yöntemi izlenebilir. Niteliksel yöntemlerde, görüntüler için tahmin edilen altyazılar bir veya daha fazla değerlendirici tarafından değerlendirilip skorlanır. Bu değerlendirme yöntemi, kişiye bağlı olması sebebiyle ve kişilerin farklı zamanlarda farklı değerlendirmelerde bulunabileceği göz önüne alındığında niceliksel yöntemler kadar güvenilir sonuçlar üretme potansiyeline sahip değildir ve oldukça maliyetlidir. Otomatik görüntü altyazılama sistemlerinde altyazı başarımını ölçmek amacıyla güncel literatürde çok sayıda niceliksel ölçüm metriği önerilmiştir. Bu metriklerin ekseriyeti de bu alanda çalışan araştırma toplulukları tarafından genel olarak kabul görmüştür ve çalışmalarda önerilen sistemlerin performansını analiz etmek için tercih edilmektedir. Görüntü altyazılamada performans ölçümü amacıyla kullanılan bu metriklerden öne çıkan bazıları şunlardır:

- BiLingual Evaluation Understudy (BLEU) [110]
- Metric for Evaluation of Translation with Explicit ORdering (METEOR) [111]
- Recall Oriented Understudy for Gisting Evaluation (ROUGE-L) [112]
- Consensus-based Image Description Evaluation (CIDEr) [113]
- Semantic Propositional Image Caption Evaluation (SPICE) [114]

Görüntü altyazılama başarımının ölçümü için en sık kullanılan metriklerden birisi BLEU'dur. BLEU metriğinin hesaplanmasında öncelikle test korpusunun modifiye edilmiş keskinlik (precision) skorlarının geometrik ortalaması alınır ve sonrasında ise bu değer, bir üstel kısalık ceza faktörü (exponential brevity penalty factor) ile çarpılır. BLEU skoru, tahmin edilen ve referans görüntülerdeki farklı sayıdaki kelime grubu eşleşmeleri (n-gram) için de hesaplanabilmekle

(BLEU-n) birlikte, bu kelime eşleşmeleri genellikle ayrı ayrı olmak üzere 4-gram'a kadar gerçekleştirilerek BLEU-1, BLEU-2, BLEU-3, BLEU-4 ölçümleri yapılır. BLEU metriği görüntü altyazılarının başarımını biçimsel bağlamda ele aldığından küçük n-gram eşleşmelerinde elde edilen yüksek skorlar her zaman anlamsal doğruluğu yansıtmayabilir. Bunun için, BLEU-4 gibi büyük n-gram eşleşmelerine dayalı ölçümleri de dikkate almak gereklidir. BLEU metriği, referans ve tahmin edilen görüntü altyazıları arasındaki benzerliğin ölçümünde duyarlılığı (recall) yani referans altyazıdaki toplam n-gram sayısı içinde eşleşen n-gramların oranını, ki bu bilgi oldukça önemli olmasına rağmen, hesaba katmaz. Bununla birlikte, BLEU metriğinin diğer bazı çalışmalarda [111] değinilen başka zayıf yönleri de bulunmaktadır. Bu zayıf yönler ilerleyen dönemde Banerjee ve Lavie tarafından ele alınmış, giderilmeye çalışılmış ve yapılan iyileştirmeler neticesinde yeni bir metrik olarak METEOR önerilmiştir. METEOR da, görüntü altyazılama modelinin ürettiği altyazı ile gerçek-referans görüntü altyazısı arasındaki kelime (word-to-word) eşleşmelerine dayalı olarak bir puan hesaplar ve tahmin edilen görüntü altyazısının doğruluğu bu puan üzerinden değerlendirilir. METEOR'un açıklamalı matematiksel notasyonu ve detayları için [111] numaralı referans kaynak incelenebilir.

ROUGE-L, CIDEr ve SPICE de görüntü altyazılamada öne çıkan diğer sayısal ölçüm metrikleridir. ROUGE-L, duyarlılık bilgisini göz önünde bulundurarak, LCS (Longest Common Subsequence) tabanlı bir altyazı doğruluk ölçümü gerçekleştirir. Bu işlemde, referans bir görüntü altyazısı ile tahmin edilen altyazı arasındaki maksimum uzunluğa sahip ortak alt dizi belirlenir ve bu en uzun ortak alt dizi sayesinde iki altyazı metni arasındaki benzerlik ortaya konulur. Bir diğer n-gram tabanlı ölçüm yöntemi olan CIDEr ise, TF-IDF (Term Frequency-Inverse Document Frequency) yöntemini kullanarak n-gramları ağırlıklandırır ve n-gram ağırlıklarına göre, aslında n-gramların taşıdıkları bilgi değerini ortaya koymaya çalışarak (veri setinde yaygın olarak görülen n-gramlara daha az bilgi verici olmaları muhtemel olduğundan daha düşük ağırlık verilmesi), daha anlamlı bir ölçüm yapmayı amaçlar. Görüntü altyazılama sistemlerince üretilen altyazıların başarımlarının değerlendirilmesinde BLEU, ROUGE-L, CIDEr veya METEOR gibi metrikleri kullanmanın problem oluşturabilecek temel noktalarından biri, bu metriklerin n-gram örtüşmesine duyarlı olmasıdır. Tahmin edilen ve referans görüntü altyazıları arasındaki n-gram örtüşmesi, iki altyazının aynı anlamı taşıdığının gösterilmesi açısından yeterli değildir [115, 114]. Bu temel noktadan hareketle Anderson ve ark. tarafından önerilen SPICE metriği, sahne graflarını da (scene graphs) kullanarak görüntü altyazıları arasındaki benzer anlamsal yapıları yakalamayı hedefler. Anlamsal içeriklerin benzerliği üzerinden bir keskinlik ve duyarlılık bilgisi de kullanılarak modellerin ürettiği altyazılar için SPICE skorları hesaplanır. SPICE metriğinin, insan değerlendirmeleriyle uyum açısından mevcut n-gram metriklerinden daha iyi performans gösterdiği de belirtilmektedir [114].

2.4. Görüntü Altyazılamanın Uygulama Alanları

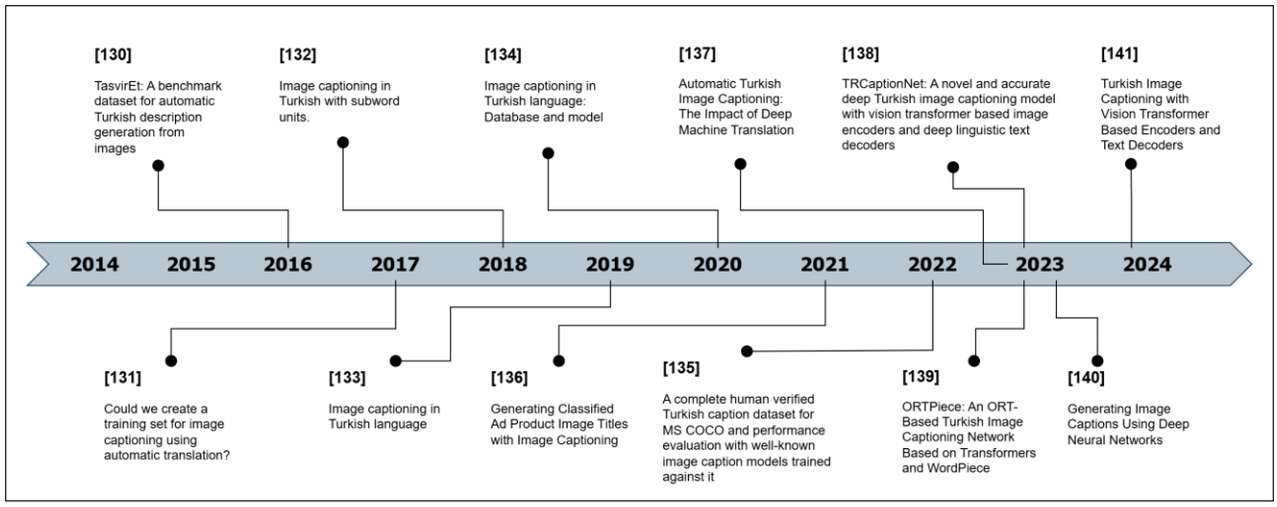
Otomatik görüntü altyazılama üzerine yapılan çalışmalar her ne kadar yoğunlukla genel-amaçlı görüntü analizi üzerine odaklansa da, bu bağlamda önerilen yöntemler tabii olarak birçok görü/kamera tabanlı akıllı sistem uygulamasında kullanılabilme potansiyeli taşır. Nitekim güncel bilimsel literatür incelendiğinde, bu tarz uygulamaların; trafik sahnesi analizi [22, 116, 117], uzaktan algılama [28-30, 118], sosyal medya [31-33], haber [34-36], tıp [37-39], turizm [40-42], sanat ve kültür [43-45], kimya ve biyoloji [46-48], güvenlik [119-121], otonom sürüş ve akıllı ulaşım/taşıma [122-124] ve askeri alan [125-127] gibi oldukça geniş bir alan (domain) yelpazesine yayıldığı da görülebilir.

Li ve ark. tarafından sunulan bir makalede [22], görüntü altyazılamaya dayalı olarak trafik sahnelerinin anlaşılması ve tahmin edilmesi üzerine bir çalışmaya yer verilmiştir. Cheng ve ark. ise, uzaktan algılama görüntülerinin betimlenmesinde otomatik görüntü altyazılamadan faydalanmışlardır [118]. [31] numaralı kaynakta refere edilen çalışmada, Chunseong Park ve ark. fotoğraf paylaşım sosyal ağlarından biri olan Instagram'dan resim gönderileri toplamışlar ve bu sosyal medya görüntüleri üzerinde görüntü altyazılama gerçekleştirmişlerdir. Tran ve ark. ise, haber makaleleri ile birlikte sunulan resimler için altyazılar üreten uçtan uca bir model önermişlerdir [34]. Akıllı turizm amaçlı yapılan alan-spesifik bir başka çalışmada [40], Fudholi ve ark., ileride çeşitli kullanıcılara yardımcı olacak yapay zekâ destekli sistemlerin geliştirilmesini de destekleyecek, yerel turizme özgü bir görüntü altyazılama modeli tasarlamışlardır. Nivedita ve ark. ise, akıllı ve güvenli video gözetim sistemleri için görüntü altyazılamadan faydalanmayı tasarlamışlardır [119]. Bu amaçla, görüntülerdeki nesneler hakkında bağlamsal açıklamalar üreten ve daha sonra bu açıklamaları, sahne hakkında bilgi sağlamak için kullanan görsel sistemler geliştirmek için bir yapı önermişlerdir. Mori ve ark. tarafından sunulan bir bildiride, görüntü altyazılamanın otonom sürüş ve akıllı ulaşım/taşıma alanında kullanımına dair bir çalışma sunulmuştur [122]. Kazaları önlemek ve yolcuları uyarmak adına, yakın gelecekte gerçekleşecek herhangi bir olay için görüntü altyazısı oluşturmanın önemli olduğundan yola çıkarak araç içi kamera görüntüleri için uygun olan bir görüntü altyazılama yöntemi önermişlerdir.

Otomatik görüntü altyazılamanın son yıllarda öne çıkan önemli uygulama alanlarından bir diğeri ise tıp ve sağlıktır. Bu alanda yapılan çalışmalar özellikle tıbbi görüntülerin otomatik olarak yorumlanması için oldukça ümit vericidir. Selivanov ve ark. tarafından yakın geçmişte yayınlanan bir makalede [37], otomatik klinik görüntü altyazısı oluşturmak için, radyolojik taramaların analizini metinsel kayıtlardan gelen yapılandırılmış hasta bilgileriyle birleştiren bir model önerilmiştir. Çalışmanın sonuçları, ilgili araştırmacıların önermiş olduğu modelin, göğüs röntgeni görüntülerinin altyazılanmasında etkin bir şekilde uygulanabilirliğini göstermiştir.

Otomatik görüntü altyazılamanın sanat ve kimya gibi çok daha spesifik konularda da uygulandığı dikkat çekmektedir. Oldukça yakın tarihli bir makalede [43], Zheng ve ark. kültürel sanat eserleri için görüntü altyazılamayı ele almış ve seramikler üzerine bir vaka çalışması gerçekleştirmişlerdir. Bu çalışmalardan daha spesifik bir konuda ise, Yi ve ark. önceden eğitilmiş bir kodlayıcı-kod çözücü mimarisi ile moleküler görüntü altyazılama gerçekleştirmişlerdir [46]. MICER olarak adlandırılan bu yapının [46], kapsamlı ve doğru otomatik moleküler yapı tanımlama araçları geliştirmek için pratik bir çerçeve sunduğu da belirtilmiştir.

Otomatik görüntü altyazılamanın askeri amaçlara yönelik olarak da uygulandığı bilinmekle birlikte, bu alandaki



Şekil 4: Türkçe görüntü altyazılama kapsamında son on yıllık süreçte yapılan çalışmalar ve bu çalışmaların kronolojisi.

çalışmalar kısıtlıdır. Literatürdeki bu kısıtlı çalışmalar da, askeri karar verme mekanizmalarını desteklemek [125] ve doğrudan görüntü alyazıları oluşturmak [126, 127] için önerilmiştir. Pan ve ark., gerçekleştirdikleri güncel bir çalışmada [127] askeri görüntü altyazılama problemini ele almışlardır. Bu çalışmada, insansız hava ve kara araçlarına ait görüntülerden ve bu görüntülere karşılık gelen altyazılardan oluşan MOCO (Military Objects in Real Combat) isimli bir açık erişimli denektaşı veri setini sunmuşlardır. Yine ilgili çalışmada [127], MAE-MilitIC olarak adlandırılan yeni bir görüntü altyazılama mimarisi de önerilmiştir.

Otomatik görüntü altyazılama, yukarıda bahsi geçen uygulama alanları haricinde, görüntü analizine açık çok daha farklı ve özel araştırma alanlarında da uygulanabilme potansiyeli taşımaktadır.

3. Türkçe Görüntü Altyazılama

Bu bölümde, Türkçe görüntü altyazılama özelinde olmak üzere, öncelikle mevcut durumun genel bir değerlendirmesi yapılmış ve literatürdeki öne çıkan çalışmalar ele alınarak bu çalışmalara dair bilgiler verilmiştir. Bununla birlikte, otomatik Türkçe görüntü altyazılama için oluşturulmuş veri setlerine dair bilgiler de paylaşılmıştır. Ayrıca yazarlar, Türkçe görüntü altyazılamanın geleceğine ve bazı potansiyel araştırma konularına dair öngörülerini paylaşmışlardır.

3.1. Türkçe Görüntü Altyazılamada Genel Durum

Güncel görüntü altyazılama literatürü İngilizce gibi yaygın ve geniş kitlelerce konuşulan diller için oldukça zengin olmasına rağmen, ne yazık ki Türkçemiz için görüntü altyazılama kapsamında yapılan çalışmalar oldukça kısıtlıdır. Bu bağlamda, Türkçe görüntü altyazılama literatürünün henüz doygun bir seviyede olmadığı ve araştırmaya oldukça açık olduğu söylenebilir. Güncel literatür göz önüne alındığında, Türkçe özelinde yapılan otomatik görüntü altyazılama çalışmalarının neredeyse on yıllık bir geçmişi olduğu görülebilir. Bu alanda yapılan ilk çalışmalara ise 2010'lu yılların ortalarında rastlanmaktadır. Bununla birlikte, son yıllarda, doğal dil işleme ve geniş dil modellerindeki (Large Language Models) [128, 129] gelişmelere de parelel olarak, Türkçe görüntü altyazılama çalışmalarında önceki yıllara nazaran umut verici bir yükseliş dikkat çekmektedir. Otomatik Türkçe görüntü altyazılama kapsamında yapılan çalışmaların detaylarına bir sonraki bölümde etraflıca yer verilmiş olup, bu çalışmaların son on yıllık süreçteki gelişimi ise Şekil 4'te yer alan zaman çizelgesi (timeline) üzerinde görsel olarak sunulmuştur.

3.2. Türkçe Görüntü Altyazılamada Mevcut Çalışmalar ve Uygulama Alanları

Otomatik Türkçe görüntü altyazılama kapsamında yapılan çalışmalar, Şekil 4 üzerinden de açıkça görüleceği üzere, yaklaşık 10 yıllık bir geçmişi kapsamaktadır. 2016 yılında Unal ve ark. tarafından sunulan görüntü betimleme çalışması [130], Türkçe görüntü altyazılamada yapılan ilk çalışmalar arasında yer almaktadır. Bu çalışmada temel olarak, literatürde ilk kez, görüntülerden Türkçe altyazılar üretmeye imkân veren ve bu amaçla denektaşı olarak kullanılabilecek yeni bir veri seti sunulmuştur. TasvirEt olarak adlandırılan bu veri seti, Flickr8K veri kümesinde yer alan görüntüler için Türkçe altyazıların toplanması ile oluşturulmuştur. Unal ve ark. tarafından yapılan bu çalışmada, aynı zamanda, otomatik Türkçe görüntü altyazılama için ilk yöntemler de önerilmiştir. Bu bağlamda ilgili çalışma [130], otomatik Türkçe görüntü altyazılama literatüründe oldukça önemli bir yere sahiptir.

2017 yılında gerçekleştirilen bir çalışmada ise [131], Samet ve ark. "Görüntü altyazılama için otomatik tercümeyle eğitim kümesi oluşturulabilir mi?" sorusuna yanıt aramışlardır. Bu amaçla, MS COCO ve Flickr30K veri setlerindeki İngilizce görüntü altyazılarını Google Translate API kullanarak Türkçe'ye çevirmişler ve çeviri Türkçe altyazılar üzerinden bir görüntü altyazılama modeli eğitmişlerdir. Samet ve ark., yaptıkları bu çalışmada oldukça yüksek başarımlar gözlemlemişlerdir. Sonuç olarak, Türkçe bir görüntü altyazısı seti haricinde, yabancı bir dilden otomatik tercüme araçları ile Türkçe'ye çevrilen görüntü altyazıları kullanarak da bu işlemin efektif olarak gerçekleştirilebileceğini göstermişlerdir. Çeviri tabanlı bu tarz bir otomatik görüntü altyazılama yaklaşımında, ilgili çalışmada da açıkça belirtildiği gibi [131], İngilizce'den Türkçe'ye çevrilmiş altyazılarla eğitilmiş bir model kullanmak yerine, doğrudan İngilizce orijinal altyazılarla eğitilmiş bir modelin üreteceği altyazıyı Türkçe'ye çevirmek

de tercih edilebilir. Hatta, İngilizce altyazılarla eğitilmiş bir modelin üreteceği altyazıyı Türkçe'ye çevirmenin daha pratik ve iyi olacağı da akla gelebilir. Samet ve ark. bu tarz bir olası durumu da ayrıca ele almışlar ve Türkçe'ye çevrilmiş görüntü altyazıları ile eğitilen modelin, İngilizce görüntü altyazıları ile eğitilip ürettiği çıktı altyazısını Türkçe'ye çeviren modelden daha başarılı olduğunu nümerik sonuçlarla da destekleyerek raporlamışlardır.

Bilindiği üzere Türkçe, dil ailesi ve dil yapısı olarak İngilizce'den oldukça farklıdır. Sondan eklemeli bir dil yapısına sahip olan Türkçe'de, sözcüklerin sonuna eklenen ekler sözcüğün anlamını değiştirebilmektedir. Bu sebeple, Türkçe için geliştirilecek olan görüntü altyazılama yöntemlerinde, yaklaşımlarında ve modellerinde, dilin yapısına ait bu tarz özelliklerin de göz önünde bulundurulması gerekebilir. 2018 yılında Kuyu ve ark. tarafından sunulan özgün bir bildiride [132], Türkçe'nin sondan eklemeli dil özelliğini dikkate alan ve bu durumun sebebiyet verebileceği performans noksanlığını tolere etmek adına, kelimeler yerine altsözcük öğelerini kullanan bir altyazılama modeli ele alınmıştır. İlgili çalışmada, önerilen bu altsözcük tabanlı modelin sözcük tabanlı normal modellere göre daha başarılı sonuçlar verdiği, ve bu sayede daha anlamlı ve Türkçe'nin dil bilgisi kurallarına daha uygun görüntü altyazıları üretildiği ifade edilmiştir. 2019 yılında yapılan bir çalışmada ise [133], Yılmaz ve ark. otomatik Türkçe görüntü altyazılama için CNN tabanlı bir görüntü kodlayıcı ile RNN'nin özel bir türü olan LSTM kullanmışlar ve başarılı bir görüntü altyazılama gerçekleştirmişlerdir. İlgili çalışmada ayrıca, MS COCO altyazı veri tabanının makine çevirisi kullanılarak Türkçe diline kazandırılmaya başlandığı da (Türkçe MS COCO) belirtilmiştir. [133]'daki ile benzer bir çalışmada da [134], Yıldız ve ark. tarafından, CNN tabanlı bir görüntü kodlayıcı ve LSTM ile bir görüntü altyazılama gerçekleştirilmiştir. Ayrıca, [133] ile refere edilen çalışmada bahsi geçen Türkçe MS COCO görüntü-altyazı veritabanı [135]'deki çalışmada detaylı olarak tanıtılmıştır. Atıcı ve Omurca tarafından gerçekleştirilmiş olan 2021 tarihli bir çalışmada [136], Türkçe görüntü altyazılama işlemi görece farklı bir problemin çözümünde ele alınmıştır. Bu bağlamda Atıcı ve Omurca, sınıflandırılmış reklamların alışveriş kategorisindeki ürün resimleri için en uygun başlığı oluşturmayı amaçlamışlardır. Derin öğrenme tabanlı görüntü altyazılama modelleri kullanılarak gerçekleştirilen bu çalışma ile ürün görsellerinin başarılı bir biçimde tanımlanabileceği ortaya konulmuştur. Bu çalışma [136] literatürde, sınıflandırılmış reklam ürün görsellerine Türkçe başlık/açıklama üreten ilk çalışma olma özelliği ile de öne çıkmaktadır.

Yukarıda değinilen çalışmalara ek olarak, özellikle son yıllarda otomatik Türkçe görüntü altyazılama üzerine yapılan çalışmalar da [137-141] oldukça dikkat çekicidir. Bilindiği üzere, Türkçe görüntü altyazı setlerinin azlığı ve kısıtlılığı sebebiyle, Türkçe görüntü altyazılama modellerinin eğitiminde kullanılacak olan görüntü altyazılarını oluşturmak için Google Translate gibi web tabanlı gelişmiş ve başarılı otomatik dil çeviri sistemlerinin kullanımı oldukça yaygındır. Yıldız ve ark. tarafından yakın geçmişte sunulan bir bildiride [137], bu yaklaşımın biraz dışına çıkılarak, otomatik Türkçe görüntü altyazılamada derin makine çevirisinin etkisine yönelik bir araştırma çalışması raporlanmıştır. İlgili çalışmada, genel-amaçlı ve birçok dili destekleyen derin bir dil modeli kullanılarak İngilizce görüntü altyazıları otomatik olarak Türkçe'ye çevrilmiş ve derin öğrenme tabanlı bir Türkçe görüntü altyazılama modeli, makine çevirisi Türkçe görüntü altyazıları kullanılarak eğitilmiştir. Yıldız ve ark. tarafından gerçekleştirilen başka bir güncel çalışmada ise [138], TRCaptionNet olarak isimlendirilen ve görü dönüştürücü tabanlı görüntü kodlayıcılar ile derin dilsel metin kod çözücüler üzerine inşa edilmiş yeni ve başarılı bir derin Türkçe görüntü altyazılama modeli önerilmiştir. İlgili çalışmada farklı veri setleri için raporlanan sonuçlar ile bu modelin Türkçe görüntü altyazılamadaki ümit verici performansı da desteklenmiştir. TRCaptionNet Türkçe görüntü altyazılama modeli, yine Yıldız ve ark. tarafından yapılmış başka bir çalışma dâhilinde [141] özel olarak TasvirEt veri seti [130] için de analiz edilmiştir. Bu çalışmada [141] gerçekleştirilen görüntü altyazılama analizlerinde, TRCaptionNet ile TasvirEt veri seti üzerinde literatürdeki en iyi başarımlara ulaşıldığı raporlanmıştır. Yine TasvirEt veri seti üzerinde gerçekleştirilen, ve Ertuğrul ve Omurca tarafından sunulan 2023 tarihli bir çalışmada [140], VGG-16, DenseNet ve LSTM kullanılarak Türkçe altyazı oluşturmaya yönelik bir analiz yapılmıştır ve başarılı addedilebilecek sonuçlar gözlemlenmiştir. TRCaptionNet'e ek olarak, otomatik Türkçe görüntü altyazılama için son yıllarda önerilmiş görü dönüştürücü tabanlı bir diğer spesifik ve yeni derin model ise ORTPiece'dir [139]. Ersoy ve ark. tarafından önerilen bu model, temel olarak girdi görüntüleri üzerinden görünüm (appearance) ve geometri özelliklerini çıkarır ve bunları kelime parçaları (wordpiece) gösterimleri ile birleştirerek Türkçe görüntü altyazılarını oluşturur. Kodlayıcı-kod çözücü bir mimariye sahip olan ORTPiece, Nesne İlişkisi Dönüştürücü (Object Relation Transformer, ORT) yöntemini de kullanır ve wordpiece tokenizasyonu ile altsözcük öğelerini analiz eder. Bu sayede, sondan eklemeli bir dil yapısına sahip olan Türkçe için daha başarılı görüntü altyazıları oluşturmayı amaçlar.

Yukarıdaki paragraflarda ana hatlarıyla ele alınmış olan otomatik Türkçe görüntü altyazılama çalışmalarına ait yöntemsel detaylar, kullanılan veri seti bilgileri ve ölçülen başarım oranları gibi daha detay bilgiler ise Tablo 2'de verilmiştir.

Şekil 4 ve Tablo 2 üzerinden de açıkça görüleceği üzere, bilimsel literatürdeki otomatik Türkçe görüntü altyazılama çalışmaları uzun bir geçmişe sahip olmamakla birlikte nicelik bakımından da henüz yüksek bir seviyede değildir. Bu alanda yapılan çalışmaların neredeyse tamamı tek bir konuya, bilinen ve genel kullanıma açık görüntü-altyazı veri setleri üzerinden genel-amaçlı görüntü altyazılamaya yöneliktir. İlgili çalışmalar arasında, yalnızca, Atıcı ve Omurca tarafından gerçekleştirilmiş olan 2021 tarihli çalışmada [136] genel-amaçlı görüntü altyazılamadan farklı bir uygulama alanına meyledildiği ve Türkçe görüntü altyazılama işleminin görece farklı bir problemin çözümüne uygulandığı (sınıflandırılmış reklamların alışveriş kategorisindeki ürün resimleri için başlık oluşturma) görülmektedir. Görüntü altyazılamanın genel olarak; tıbbi görüntü analizi, uzaktan algılama, güvenlik, görsel yardım ve çok modlu arama motorları gibi birçok farklı alanda uygulanabilirliği göz önüne alındığında, Türkçe görüntü altyazılamanın birçok uygulama alanında henüz aktif olarak yer almadığı ve bu alanların araştırmaya ve çalışmaya oldukça açık olduğu ifade edilebilir.

Tablo 2: Güncel literatürdeki otomatik Türkçe görüntü altyazılama çalışmalarında yöntemsel detaylar, kullanılan veri seti bilgileri ve ölçülen başarım oranları.

Kaynak	Yıl	Yöntem	Veri seti	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[130]	2016	Benzer görüntü altyazıları üzerinden altyazılama	TasvirEt (eğitim ve test)	0.2110	0.0720	0.0200	-	-	-	-	-
[130]	2016	Adaptif komşuluk temelli konsensus altyazılama	TasvirEt (eğitim ve test)	0.2600	0.1020	0.0340	-	-	-	-	-
[131]	2017	Derin öğrenme (CNN + kök alma)	MS COCO (eğitim ve test)	0.3420	0.1810	0.0850	0.0390	0.1570	0.3220	0.4610	-
[131]	2017	Derin öğrenme (CNN + kök alma)	Flickr30K (eğitim) ve TasvirEt (test)	0.3350	0.1720	0.0820	0.0350	0.1230	0.2680	0.2360	-
[132]	2018	Derin ögrenme (Altsözcük tabanlı)	MS COCO (eğitim ve test)	0.2930	0.1650	0.0880	0.0530	0.1470	0.3020	0.5670	-
[132]	2018	Derin ögrenme (Altsözcük tabanlı)	MS COCO (eğitim) ve TasvirEt (test)	0.2630	0.1630	0.0970	0.0550	0.1030	0.2670	0.3200	-
[132]	2018	Derin ögrenme (Altsözcük tabanlı)	Flickr30K (eğitim ve test)	0.2970	0.1650	0.0790	0.0380	0.1120	0.2650	0.2140	-
[133]	2019	Derin öğrenme (CNN + LSTM)	MS COCO (eğitim ve test)	0.2880	0.1550	0.0710	0.0300	0.1250	0.2660	0.4790	-
[134]	2020	Derin öğrenme (CNN + RNN)	MS COCO (eğitim ve test)	0.2880	0.1550	0.0710	0.0300	0.1250	0.2660	0.4790	-
[134]	2020	Derin öğrenme (CNN + LSTM)	MS COCO (eğitim ve test)	0.2970	0.1640	0.0760	0.0350	0.1290	0.2720	0.5280	-
[135]	2022	Derin öğrenme (Resnet101 Batch60 English)	MS COCO (eğitim ve test)	0.8330	0.6450	0.4920	0.4150	0.2330	0.4690	1.3620	-
[135]	2022	Derin öğrenme (Resnet101 Batch60 Turkish)	MS COCO (eğitim ve test)	0.4990	0.4240	0.3130	0.2370	0.2500	0.4630	1.2570	-
[135]	2022	Derin öğrenme (Resnet152 Batch60 Turkish)	MS COCO (eğitim ve test)	0.5620	0.4010	0.2990	0.2270	0.1930	0.4480	0.9800	-
[135]	2022	Derin öğrenme (Resnet101 Batch32 Turkish)	MS COCO (eğitim ve test)	0.4370	0.3540	0.2750	0.2140	0.1700	0.4050	0.8990	-
[135]	2022	Derin öğrenme (Meshed-Memory Transformers Turkish)	MS COCO (eğitim ve test)	0.7230	0.5620	0.4130	0.2960	0.2550	0.5270	0.9350	-
[137]	2023	Derin öğrenme (ResNet18 + LSTM)	MS COCO (eğitim ve test)	0.3122	0.1054	0.0441	0.0217	0.1111	0.2636	0.0465	-
[137]	2023	Derin öğrenme (ResNet34 + LSTM)	MS COCO (eğitim ve test)	0.3116	0.1068	0.0454	0.0206	0.1111	0.2642	0.0466	-
[137]	2023	Derin öğrenme (ResNet50 + LSTM)	MS COCO (eğitim ve test)	0.3115	0.1060	0.0446	0.0213	0.1110	0.2636	0.0447	-
[137]	2023	Derin öğrenme (ResNet101 + LSTM)	MS COCO (eğitim ve test)	0.3079	0.1024	0.0434	0.0191	0.1102	0.2623	0.0435	-
[137]	2023	Derin öğrenme (ResNet152 + LSTM)	MS COCO (eğitim ve test)	0.2989	0.1028	0.0432	0.0198	0.1070	0.2539	0.0440	-
[137]	2023	Derin öğrenme (ResNext50 (32x4d) + LSTM)	MS COCO (eğitim ve test)	0.3086	0.1037	0.0411	0.0178	0.1108	0.2624	0.0423	-
[137]	2023	Derin öğrenme (ResNext101 (32x8d) + LSTM)	MS COCO (eğitim ve test)	0.3105	0.1061	0.0453	0.0205	0.1112	0.2636	0.0465	-
[137]	2023	Derin öğrenme (ResNext101 (64x4d) + LSTM)	MS COCO (eğitim ve test)	0.3109	0.1055	0.0450	0.0214	0.1106	0.2633	0.0448	-
[137]	2023	Derin öğrenme (Swin (base) + LSTM)	MS COCO (eğitim ve test)	0.3122	0.1040	0.0433	0.0179	0.1106	0.2639	0.0432	-
[137]	2023	Derin öğrenme (Swin (small) + LSTM)	MS COCO (eğitim ve test)	0.3079	0.1045	0.0440	0.0198	0.1104	0.2627	0.0443	-
[137]	2023	Derin öğrenme (Swin (tiny) + LSTM)	MS COCO (eğitim ve test)	0.3103	0.1043	0.0433	0.0196	0.1100	0.2628	0.0423	-
[138]	2023	Derin öğrenme (TRCaptionNet)	MS COCO (eğitim ve test)	0.5773	0.4146	0.2839	0.1954	0.2546	0.4652	0.6552	-
[138]	2023	Derin öğrenme (TRCaptionNet)	MS COCO (eğitim) ve Flickr30K (test)	0.5311	0.3587	0.2293	0.1395	0.2146	0.4192	0.3798	-
[138]	2023	Derin öğrenme (TRCaptionNet)	Flickr30K (eğitim) ve MS COCO (test)	0.4213	0.2576	0.1518	0.0900	0.1933	0.3619	0.3145	-
[138]	2023	Derin öğrenme (TRCaptionNet)	Flickr30K (eğitim ve test)	0.5477	0.3771	0.2535	0.1671	0.2271	0.4393	0.4668	-
[138]	2023	Derin öğrenme (TRCaptionNet)	MS COCO + Flickr30K (eğitim) ve MS COCO (test)	0.5761	0.4124	0.2803	0.1905	0.2523	0.4609	0.6437	-
[138]	2023	Derin öğrenme (TRCaptionNet)	MS COCO + Flickr30K (eğitim) ve Flickr30K (test)	0.5713	0.4056	0.2789	0.1843	0.2330	0.4491	0.5154	-
[139]	2023	Derin öğrenme (ORTPiece (voc. size 19k))	MS COCO (eğitim ve test)	0.3300	-	-	0.0550	0.1720	0.3250	0.5260	0.0770
[139]	2023	Derin öğrenme (ORTPiece (voc. size 24k))	MS COCO (eğitim ve test)	0.3330	-	-	0.0530	0.1740	0.3290	0.5530	0.0770
[139]	2023	Derin öğrenme (ORTPiece (voc. size 39k))	MS COCO (eğitim ve test)	0.3310	-	-	0.0530	0.1720	0.3260	0.5240	0.0740
[140]	2023	Derin öğrenme (VGG-16 + LSTM)	TasvirEt (eğitim ve test)	0.3316	0.1118	0.0687	0.0571	-	-	-	-
[140]	2023	Derin öğrenme (DenseNet + LSTM)	TasvirEt (eğitim ve test)	0.3068	0.0946	0.0617	0.0421	-	-	-	-
[141]	2024	Derin öğrenme (TRCaptionNet (CLIP ViT-B/16))	TasvirEt (eğitim ve test)	0.2880	0.1662	0.0915	0.0459	0.1392	0.2738	0.2709	-
[141]	2024	Derin öğrenme (TRCaptionNet (CLIP ViT-B/32))	TasvirEt (eğitim ve test)	0.3093	0.1862	0.1063	0.0577	0.1457	0.2890	0.3135	-
[141]	2024	Derin öğrenme (TRCaptionNet (CLIP ViT-L/14))	TasvirEt (eğitim ve test)	0.3069	0.1841	0.1042	0.0555	0.1480	0.2889	0.3078	-
[141]	2024	Derin öğrenme (TRCaptionNet (CLIP ViT-L/14@336px))	TasvirEt (eğitim ve test)	0.3406	0.2110	0.1253	0.0690	0.1610	0.3145	0.3879	-

3.3. Türkçe Görüntü Altyazılama için Veri Setleri

Otomatik Türkçe görüntü altyazılama üzerine mevcut çalışmaların ve bu çalışmaların dâhil olduğu özel uygulama alanlarının ele alındığı bir önceki bölümde de ifade edildiği üzere, Türkçe görüntü altyazılama literatürü yapılan çalışmaların niceliği ve çeşitliliği açısından henüz yüksek bir seviyede değildir. Mevcut literatürün bu durumunda, her ne kadar Türkçe üzerine yapılan ilk çalışmaların oldukça yakın tarihli olmasının etkisi olsa da, kanaatimizce, Türkçe görüntü altyazılama için oluşturulmuş veri setlerinin az ve kısıtlı olmasının da etkisi vardır. Nitekim mevcut çalışmalar

incelendiğinde, otomatik Türkçe görüntü altyazılama ile ilgili problemlerin çözümünde istifade edilebilecek veri setlerinin yalnızca TasvirEt ve Türkçe MS COCO veri setleri olduğu görülmektedir. Bu iki veri seti de, her ne kadar Türkçe görüntü-altyazı seti olsa da, esasında MS COCO ve Flickr gibi bilinen görüntü-altyazı setleri referans alınarak oluşturulmuştur.

TasvirEt [130], Unal ve ark. tarafından Flickr8K görüntüleri üzerine Türkçe altyazıların toplanması ile oluşturulmuş bir veri setidir. TasvirEt veri kümesi 8091 adet görüntü ve bu görüntülere ait toplam 12222 adet Türkçe görüntü altyazısından meydana gelmektedir. TasvirEt veri setinin tanıtıldığı bildiride [130], görüntü altyazılarının toplanabilmesi için özel bir sistem geliştirildiği de bildirilmiştir. Golech ve ark. tarafından hazırlanan Türkçe MS COCO görüntü altyazılama veri seti ise [135], orijinal MS COCO veritabanındaki görsellere ait altyazıların Google Cloud Platform'un çeviri Uygulama Programlama Arayüzü (Application Programming Interface, API) kullanılarak Türkçe'ye çevrilmesiyle oluşturulmuştur. İlgili altyazı çevirilerinin biçimsel ve anlamsal doğruluğunun insan gözü ile doğrulanması amacıyla da bir ekip oluşturulmuş ve görselleri daha doğru bir şekilde tanımlamak için çeviri altyazılar bu ekipçe düzeltilmiştir. Türkçe MS COCO veri kümesi toplamda 123.287 adet görüntü ve bu görüntülere karşılık gelen toplam 616.767 açıklamadan (görüntü altyazısı) müteşekkildir. TasvirEt ve Türkçe MS COCO veri setlerine ait çok daha detaylı bilgilere ve istatistiklere bu veri setlerinin tanıtıldığı bilimsel raporlardan [130, 135] erişilebilmesi mümkündür.

Bilindiği üzere, görüntü altyazılama için veri seti oluşturmak oldukça yorucu ve iş gücü, zaman gibi kaynaklar açısından da oldukça maliyetli bir süreçtir. Bu zorluk ve maliyet her ne kadar tüm veri setlerinin oluşturulmasında genel bir problem olsa da, görüntü altyazılama için bunun bir parça daha üst seviyede olduğu söylenebilir. Bunun temel ve öne çıkan sebeplerinden bazıları ise şunlardır:

- Veri setinin kategori ve anlamsal içerik bakımından zengin olması gerekliliği: Modellerin görüntülerdeki birçok içeriği öğrenebilmesi ve buna bağlı olarak daha anlamlı görüntü altyazıları üretebilmesi.
- Görüntülere ait gerçek-referans altyazıların oluşturulmasının oldukça vakit alıcı olması: Görüntülerin önce insan gözüyle detaylı olarak incelenmesi ve görüntü altyazılarının belli bir düşünme sürecinden sonra yazılması.
- Gerçek-referans görüntü altyazılarının anlamlı, doğru, içeriği iyi tanımlayan ve yeterli bir kelime uzunluğunda oluşturulması gerekliliği: Modellerin iyi tanımlanmış verilerle eğitilmesi ve buna bağlı olarak anlamlı, doğruluk oranı yüksek altyazılar üretebilmesi.
- Subjektiflik: Bir görüntü için, belli bir benzerlik dâhilinde, herkesin farklı bir betimleme yapabilmesi. Aynı zamanda, bir kişinin bir görüntüyü farklı zamanlarda farklı biçimsel ifadelerle betimleyebilmesi.
- Her bir görüntü için birden fazla gerçek-referans altyazı oluşturulması ihtiyacı: Bir görüntüyü betimleyen birden fazla doğru altyazının olması.

Türkçe görüntü altyazılama literatüründe, bildiğimiz kadarıyla, bahsi geçen iki veri setinin (TasvirEt ve Türkçe MS COCO) haricinde genel kullanıma açık bir görüntü-altyazı seti bulunmamaktadır. Ayrıca, yine bilgimiz dâhilinde, tamamen Türkçe özelinde hazırlanmış orjinal bir görüntü-altyazı seti de şu an için mevcut değildir. Her ne kadar şu an için çeviri yoluyla elde edilmiş veri setleri üzerinden başarılı sayılabilecek görüntü altyazılama modelleri geliştirilebilse de, Türkçe'ye spesifik yeni görüntü-altyazı veri setlerinin oluşturulması bu alandaki araştırmalara ve uygulamalara önemli katkı sağlayacaktır.

3.4. Türkçe Görüntü Altyazılamanın Geleceği ve Olası Eğilimler

Literatürde otomatik Türkçe görüntü altyazılama üzerine yapılmış çalışmaların yıllara göre gelişimi ve dağılımı Şekil 4'teki bir zaman çizgisi üzerinde, bu çalışmalarda kullanılan yöntemler ile raporlanan başarımlar ise Tablo 2'de gösterilmişti. Şekil 4'e odaklanıldığında, her ne kadar çalışma sayısı kısıtlı olsa da, mevcut çalışmaların yaklaşık %50'sinin son birkaç yıl içinde yapıldığı ve raporlandığı dikkat çekmektedir. Bununla birlikte, son yıllarda yapılan çalışma sayısında bir artış da mevcuttur. Türkçe doğal dil modellerindeki gelişmeler [142-147] ve henüz Türkçe görüntü altyazılama ile ilgili birçok alt alanda çalışma yapılmamış olması da göz önüne alındığında, otomatik Türkçe görüntü altyazılama kapsamındaki çalışmaların artacağı öngörülmektedir.

2. Bölüm'deki "Görüntü Altyazılamanın Uygulama Alanları" başlığı altında, otomatik görüntü altyazılamanın hâlihazırdaki bazı genel ve spesifik uygulama alanları ve bu alanlarda yapılmış öne çıkan bazı güncel araştırmalar paylaşılmıştı. Literatürdeki otomatik Türkçe görüntü altyazılama çalışmaları, bazı çalışmalar haricinde (örneğin alışveriş kategorisindeki ürün resimleri için altyazı oluşturma gibi [136]) neredeyse tamamen genel-amaçlı görüntü altyazılamaya yöneliktir. Dolayısıyla, Türkçe görüntü altyazılama şu an için tıp, turizm, eğitim, eğlence, ulaşım ve engelli bireylerin hayat standartlarını arttırma gibi birçok alana özel uygulamaya açıktır. Bu alanlarda yapılmış araştırma çalışmalarında İngilizce altyazı üretici modellerin domine edici etkisi görülmektedir. Bu tarz modellerin, çok büyük miktardaki veriye olan ihtiyaçları göz önüne alındığında yapısal ve dilbilimsel açıdan Türkçe'ye uyarlanması güç gibi görünmesine rağmen, gelişmiş otomatik çeviri sistemleri ile entegre edilerek üretecekleri çıktıların Türkçe'ye uyarlanabilmesi mümkündür. Bu bağlamda, ChatGPT [148], Bard [149], Gemini [150] vb. gibi geniş dil modellerinden istifade edilebilir. Öte yandan, Türkçe doğal dil modellerinin geliştirilmesinde son yıllarda yaşanan umut verici gelişmeler de [142-147] hesaba katıldığında, yabancı diller için geliştirilmiş görüntü altyazılama modellerini Türkçe için kullanmanın ve alana özel spesifik Türkçe görüntü altyazılama uygulamaları geliştirmenin daha da kolaylaşacağı söylenebilir.

4. Sonuçlar

Bu makalede, otomatik Türkçe görüntü altyazılama üzerine bir inceleme çalışması sunulmuş olup, genel olarak Türkçe görüntü altyazılamadaki mevcut çalışmaların/uygulamaların ele alınması ve değerlendirilmesi amaçlanmıştır. Bununla birlikte, görüntü altyazılamada kullanılan genel yaklaşımlara, veri setlerine ve metriklere de makale kapsamında yer verilmiştir. Ayrıca, Türkçe görüntü altyazılamanın mevcut uygulama alanları da analiz edilmiş ve bu konu ile ilgili

araştırmaya açık alanlara da temas edilmiştir. Çalışma kapsamında yapılan incelemeler ve değerlendirmeler, son yıllarda otomatik Türkçe görüntü altyazılamada yaşanan önemli ve umut verici gelişmelere rağmen Türkçe görüntü altyazılamada henüz kaydadeğer bir literatür birikimi oluşmadığını ortaya koymaktadır. Literatürdeki mevcut çalışmaların çok büyük bir yüzdesinin (yaklaşık %50) son birkaç yılda yapıldığı da göz önüne alındığında bu durum daha da belirgin olarak göze çarpmaktadır. Dahası, ilgili çalışmaların neredeyse tamamının genel-amaçlı görüntü altyazılama problemine yoğunlaştığı, farklı uygulama alanlarına yönelik (tıp, eğitim, turizm, ulaşım vb. gibi) spesifik araştırma çalışmalarının çok kısıtlı olduğu tespit edilmiştir. Bu bağlamda, Türkçe görüntü altyazılama literatürünün bu alanda çalışma yürüten veya uygulama geliştirmeyi planlayan araştırmacıların hemen hemen her türlü katkısına açık olduğunu ifade etmek mümkündür.

5. Kaynaklar

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Advances in neural information processing systems, vol. 25, 2012.

[2] S. Yıldız, O. Aydemir, A. Memiş, and S. Varlı, "A turnaround control system to automatically detect and monitor the time stamps of ground service actions in airports: A deep learning and computer vision based approach," Engineering Applications of Artificial Intelligence, vol. 114, p. 105032, 2022, doi:10.1016/j.engappai.2022.105032.

[3] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," Medical image analysis, vol. 79, p. 102444, 2022, doi:10.1016/j.media.2022.102444.

[4] A. M. Ozbayoglu, M. U. Gudelek, and O. B. Sezer, "Deep learning for financial applications: A survey," Applied soft computing, vol. 93, p. 106384, 2020, doi:10.1016/j.asoc.2020.106384.

[5] X. Zhifeng, "Virtual entertainment robot based on artificial intelligence image capture system for sports and fitness posture recognition," Entertainment Computing, vol. 52, p. 100793, 2025, doi:10.1016/j.entcom.2024.100793.

[6] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," Engineering, vol. 18, pp. 143–153, 2022, doi:10.1016/j.eng.2021.03.023.

[7] R. K. Kaliyar, A. Goswami, and P. Narang, "FakeBERT: Fake news detection in social media with a BERT-based deep learning approach," Multimedia tools and applications, vol. 80, no. 8, pp. 11 765–11 788, 2021, doi:10.1007/s11042-020-10183-2.

[8] Y. Yang, K. Zhang, and P. Kannan, "Identifying market structure: A deep network representation learning of social engagement," Journal of Marketing, vol. 86, no. 4, pp. 37–56, 2022, doi:10.1177/00222429211033585.

[9] M. Soori, B. Arezoo, and R. Dastres, "Artificial intelligence, machine learning and deep learning in advanced robotics, a review," Cognitive Robotics, vol. 3, pp. 54–70, 2023, doi:10.1016/j.cogr.2023.04.001.

[10] M. Ozkan-Ozay, E. Akin, Ö. Aslan, S. Kosunalp, T. Iliev, I. Stoyanov, and I. Beloev, "A comprehensive survey: Evaluating the efficiency of artificial intelligence and machine learning techniques on cyber security solutions," IEEE Access, 2024, doi:10.1109/ACCESS.2024.3355547.

[11] Y. Matsuzaka and R. Yashiro, "Ai-based computer vision techniques and expert systems," AI, vol. 4, no. 1, pp. 289–302, 2023, doi:10.3390/ai4010013.

[12] S. V. Mahadevkar, B. Khemani, S. Patil, K. Kotecha, D. R. Vora, A. Abraham, and L. A. Gabralla, "A review on machine learning styles in computer vision—techniques and future directions," Ieee Access, vol. 10, pp. 107 293–107 329, 2022, doi:10.1109/ACCESS.2022.3209825.

[13] R. Kaur and S. Singh, "A comprehensive review of object detection with deep learning," Digital Signal Processing, vol. 132, p. 103812, 2023, doi:10.1016/j.dsp.2022.103812.

[14] D. Meimetis, I. Daramouskas, I. Perikos, and I. Hatzilygeroudis, "Real-time multiple object tracking using deep learning methods," Neural Computing and Applications, vol. 35, no. 1, pp. 89–118, 2023, doi:10.1007/s00521-021-06391-y.

[15] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao, "Review the state-of-the-art technologies of semantic segmentation based on deep learning," Neurocomputing, vol. 493, pp. 626–646, 2022, doi:10.1016/j.neucom.2022.01.005.

[16] S. Yıldız, "Turkish scene text recognition: Introducing extensive real and synthetic datasets and a novel recognition model," Engineering Science and Technology, an International Journal, vol. 60, p. 101881, 2024, doi:10.1016/j.jestch.2024.101881.

[17] D. Sarvamangala and R. V. Kulkarni, "Convolutional neural networks in medical image understanding: a survey," Evolutionary intelligence, vol. 15, no. 1, pp. 1–22, 2022, doi:10.1007/s12065-020-00540-3.

[18] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, "Deep image captioning: A review of methods, trends and future challenges," Neurocomputing, vol. 546, p. 126287, 2023, doi:10.1016/j.neucom.2023.126287.

[19] A. Verma, A. K. Yadav, M. Kumar, and D. Yadav, "Automatic image caption generation using deep learning," Multimedia Tools and Applications, vol. 83, no. 2, pp. 5309–5325, 2024, doi:10.1007/s11042-023-15555-y.

[20] L. Agarwal and B. Verma, "From methods to datasets: A survey on Image-Caption Generators," Multimedia Tools and Applications, vol. 83, no. 9, pp. 28 077–28 123, 2024, doi:10.1007/s11042-023-16560-x.

[21] S. Sreela and S. M. Idicula, "A Systematic Survey of Automatic Image Description Generation Systems," International Journal of Image and Graphics, p. 2650002, 2024, doi:10.1142/S0219467826500026.

[22] W. Li, Z. Qu, H. Song, P. Wang, and B. Xue, "The traffic scene understanding and prediction based on image captioning," IEEE Access, vol. 9, pp. 1420–1427, 2020, doi:10.1109/ACCESS.2020.3047091.

[23] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou, and H. Yu, "Visuals to text: A comprehensive review on automatic image captioning," IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 8, pp. 1339–1365, 2022, doi:10.1109/JAS.2022.105734.

[24] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," ACM Computing Surveys (CsUR), vol. 51, no. 6, pp. 1–36, 2019, doi:10.1145/3295748.
[25] D. Sharma, C. Dhiman, and D. Kumar, "Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey," Expert Systems with Applications, vol. 221, p. 119773, 2023, doi:10.1016/j.eswa.2023.119773.
[26] F. Chen, X. Li, J. Tang, S. Li, and T. Wang, "A survey on recent advances in image captioning," in Journal of Physics: Conference Series, vol. 1914, no. 1. IOP Publishing, 2021, p. 012053, doi:10.1088/1742-6596/1914/1/012053.
[27] H. Sharma and D. Padha, "A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues," Artificial Intelligence Review, vol. 56, no. 11, pp. 13 619–13 661, 2023, doi:10.1007/s10462-023-10488-2.
[28] X. Huang, K. Lu, S. Wang, J. Lu, X. Li, and R. Zhang, "Understanding remote sensing imagery like reading a text document: What can remote sensing image captioning offer?" International Journal of Applied Earth Observation and Geoinformation, vol. 131, p. 103939, 2024, doi:10.1016/j.jag.2024.103939.
[29] Y. Li, X. Zhang, X. Cheng, X. Tang, and L. Jiao, "Learning consensus-aware semantic knowledge for remote sensing image captioning ," Pattern Recognition, vol. 145, p. 109893, 2024, doi:10.1016/j.patcog.2023.109893.
[30] K. Zhao and W. Xiong, "Exploring region features in remote sensing image captioning," International Journal of Applied Earth Observation and Geoinformation, vol. 127, p. 103672, 2024, doi:10.1016/j.jag.2024.103672.
[31] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 895–903, doi:10.1109/CVPR.2017.681.
[32] N. Sanguansub, P. Kamolrungwarakul, S. Poopair, K. Techaphonprasit, and T. Siriborvornratanakul, "Song lyrics recommendation for social media captions using image captioning, image emotion, and caption-lyric matching via universal sentence embedding," Social Network Analysis and Mining, vol. 13, no. 1, p. 95, 2023, doi:10.1007/s13278-023-01097-6.
[33] J.-H. Wang, C.-W. Huang, and M. Norouzi, "Improving Rumor Detection by Image Captioning and Multi-Cell Bi-RNN With Self-Attention in Social Networks," International Journal of Data Warehousing and Mining (IJDWM), vol. 18, no. 1, pp. 1–17, 2022, doi:10.4018/IJDWM.313189.
[34] A. Tran, A. Mathews, and L. Xie, "Transform and tell: Entity-aware news image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13 035–13 045, doi:10.1109/CVPR42600.2020.01305.
[35] Z. Zhang, H. Zhang, J. Wang, Z. Sun, and Z. Yang, "Generating news image captions with semantic discourse extraction and contrastive style-coherent learning," Computers and Electrical Engineering, vol. 104, p. 108429, 2022, doi:10.1016/j.compeleceng.2022.108429.
[36] J. Chen and H. Zhuge, "A news image captioning approach based on multimodal pointer-generator network," Concurrency and Computation: Practice and Experience, vol. 34, no. 7, p. e5721, 2022, doi:10.1002/cpe.5721.
[37] A. Selivanov, O. Y. Rogov, D. Chesakov, A. Shelmanov, I. Fedulova, and D. V. Dylov, "Medical image captioning via generative pretrained transformers," Scientific Reports, vol. 13, no. 1, p. 4171, 2023, doi:10.1038/s41598-023-31223-5.
[38] G. Reale-Nosei, E. Amador-Domínguez, and E. Serrano, "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," Medical Image Analysis, p. 103264, 2024, doi:10.1016/j.media.2024.103264.
[39] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, "Diagnostic captioning: a survey," Knowledge and Information Systems, vol. 64, no. 7, pp. 1691–1722, 2022, doi:10.1007/s10115-022-01684-7.
[40] D. H. Fudholi, Y. Windiatmoko, N. Afrianto, P. E. Susanto, M. Suyuti, A. F. Hidayatullah, and R. Rahmadi, "Image captioning with attention for smart local tourism using efficientnet," in IOP Conference Series: Materials Science and Engineering, vol. 1077, no. 1. IOP Publishing, 2021, p. 012038, doi:10.1088/1757-899X/1077/1/012038.
[41] S. Watcharabutsarakham, S. Marukatat, K. Kiratiratanapruk, and P. Temniranrat, "Image Captioning for Thai Cultures," in 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP). IEEE, 2022, pp. 1–5, doi:10.1109/iSAI-NLP56921.2022.9960251.
[42] Y. Bounab, M. Oussalah, and A. Ferdenache, "Reconciling image captioning and user's comments for urban tourism," in 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2020, pp. 1–6, doi:10.1109/IPTA50016.2020.9286602.
[43] B. Zheng, F. Liu, M. Zhang, T. Zhou, S. Cui, Y. Ye, and Y. Guo, "Image captioning for cultural artworks: a case study on ceramics," Multimedia Systems, vol. 29, no. 6, pp. 3223–3243, 2023, doi:10.1007/s00530-023-01178-8.
[44] E. Cetinic, "Iconographic image captioning for artworks," in Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part III. Springer, 2021, pp. 502–516, doi:10.1007/978-3-030-68796-0_36.
[45] Y. Lu, C. Guo, X. Dai, and F.-Y. Wang, "Artcap: A dataset for image captioning of fine art paintings," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 576–587, 2022, doi:10.1109/TCSS.2022.3223539.
[46] J. Yi, C. Wu, X. Zhang, X. Xiao, Y. Qiu, W. Zhao, T. Hou, and D. Cao, "Micer: a pre-trained encoder–decoder architecture for molecular image captioning," Bioinformatics, vol. 38, no. 19, pp. 4562–4572, 2022, doi:10.1093/bioinformatics/btac545.
[47] A. Lozano, M. W. Sun, J. Burgess, L. Chen, J. J. Nirschl, J. Gu, I. Lopez, J. Aklilu, A. W. Katzer, C. Chiu et al., "BIOMEDICA: An Open Biomedical Image-Caption Archive, Dataset, and Vision-Language Models Derived from Scientific Literature," arXiv preprint

arXiv:2501.07171, 2025, doi:10.48550/arXiv.2501.07171.

[48] D. Tran, N. T. Pham, N. Nguyen, and B. Manavalan, " Mol2Lang-VLM: Vision-and Text-Guided Generative Pre-trained Language Models for Advancing Molecule Captioning through Multimodal Fusion," in Proceedings of the 1st Workshop on Language+ Molecules (L+M 2024), 2024, pp. 97–102, doi:10.18653/v1/2024.langmol-1.12.

[49] H. Sharma and D. Padha, "Domain-specific image captioning: a comprehensive review," International Journal of Multimedia Information Retrieval, vol. 13, no. 2, pp. 1–27, 2024, doi:10.1007/s13735-024-00328-6.

[50] X. He and L. Deng, "Deep learning for image-to-text generation: A technical overview," IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 109–116, 2017, doi:10.1109/MSP.2017.2741510.

[51] S. Bai and S. An, "A survey on automatic image caption generation," Neurocomputing, vol. 311, pp. 291–304, 2018, doi:10.1016/j.neucom.2018.05.080.

[52] Y. Yang, C. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 444–454.

[53] R. Mason and E. Charniak, "Nonparametric method for data-driven image captioning," in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2014, pp. 592–598.

[54] T. Ghandi, H. Pourreza, and H. Mahyar, "Deep learning approaches on image captioning: A review," ACM Computing Surveys, vol. 56, no. 3, pp. 1–39, 2023, doi:10.1145/3617592.

[55] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5561–5570, doi:10.1109/CVPR.2018.00583.

[56] K. R. Suresh, A. Jarapala, and P. Sudeep, "Image captioning encoder–decoder models using cnn-rnn architectures: A comparative study," Circuits, Systems, and Signal Processing, vol. 41, no. 10, pp. 5719–5742, 2022, doi:10.1007/s00034-022-02050-2.

[57] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651–4659, doi:10.1109/CVPR.2016.503.

[58] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 4634–4643, doi:10.1109/ICCV.2019.00473.

[59] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 10 685–10 694, doi:10.1109/CVPR.2019.01094.

[60] N. Xu, A.-A. Liu, J. Liu, W. Nie, and Y. Su, "Scene graph captioner: Image captioning based on structural visual representation," Journal of Visual Communication and Image Representation, vol. 58, pp. 477–485, 2019, doi:10.1016/j.jvcir.2018.12.027.

[61] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," IEEE Transactions on Multimedia, vol. 21, no. 8, pp. 2117–2130, 2019, doi:10.1109/TMM.2019.2896516.

[62] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder-decoder network for image captioning," IEEE Transactions on Multimedia, vol. 21, no. 11, pp. 2942–2956, 2019, doi:10.1109/TMM.2019.2915033.

[63] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," IEEE Access, vol. 10, pp. 33 679–33 694, 2022, doi:10.1109/ACCESS.2022.3161428.

[64] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10 578–10 587, doi:10.1109/CVPR42600.2020.01059.

[65] K. Qian, Y. Pan, H. Xu, and L. Tian, "Transformer model incorporating local graph semantic attention for image caption," The Visual Computer, pp. 1–12, 2023, doi:10.1007/s00371-023-03180-7.

[66] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 4125–4134, doi:10.1109/CVPR.2019.00425.

[67] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 290–298, doi:10.1109/CVPR.2017.128.

[68] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 17 980–17 989, doi:10.1109/CVPR52688.2022.01745.

[69] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[70] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation, parallel distributed processing, explorations in the microstructure of cognition," ed. de rumelhart and j. mcclelland. vol. 1. 1986, Biometrika, vol. 71, no. 599-607, p. 6, 1986, doi:10.7551/mitpress/4943.003.0128.

[71] A. Tsantekidis, N. Passalis, and A. Tefas, "Recurrent neural networks," in Deep learning for robot perception and cognition. Elsevier, 2022, pp. 101–115, doi:10.1016/B978-0-32-385787-1.00010-5.

[72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi:10.1162/neco.1997.9.8.1735.

[73] A. Sherstinsky, "Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, 2020, doi:10.1016/j.physd.2019.132306.

[74] K. Cho, "Learning phrase representations using rnn encoder-decoder for statistical machine translation,"

arXiv preprint arXiv:1406.1078, 2014, doi:10.48550/arXiv.1406.1078.
[75] O. Ondeng, H. Ouma, and P. Akuon, "A review of transformer-based approaches for image captioning," Applied Sciences, vol. 13, no. 19, p. 11103, 2023, doi:10.3390/app131911103.
[76] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, etT. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020, doi:10.48550/arXiv.2010.11929.
[77] M. J. Parseh and S. Ghadiri, "Graph-based image captioning with semantic and spatial features," Signal Processing: Image Communication, vol. 133, p. 117273, 2025, doi:10.1016/j.image.2025.117273.
[78] J. Jia, X. Ding, S. Pang, X. Gao, X. Xin, R. Hu, and J. Nie, "Image captioning based on scene graphs: A survey," Expert Systems with Applications, vol. 231, p. 120698, 2023, doi:10.1016/j.eswa.2023.120698.
[79] A. S. Al-Shamayleh, O. Adwan, M. A. Alsharaiah, A. H. Hussein, Q. M. Kharma, and C. I. Eke, "A comprehensive literature review on image captioning methods and metrics based on deep learning technique," Multimedia Tools and Applications, vol. 83, no. 12, pp. 34 219–34 268, 2024, doi:10.1007/s11042-024-18307-8.
[80] 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 Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755, doi:10.1007/978-3-319-10602-1_48.
[81] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," Journal of Artificial Intelligence Research, vol. 47, pp. 853–899, 2013, doi:10.1613/jair.3994.
[82] 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, doi:10.1162/tacl_a_00166.
[83] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128–3137, doi:10.1109/TPAMI.2016.2598339.
[84] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The iapr tc-12 benchmark: A new evaluation resource for visual information systems," in International workshop ontoImage, vol. 2, 2006.
[85] 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, 2010, pp. 139–147.
[86] 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 Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11. Springer, 2010, pp. 15–29, doi:10.1007/978-3-642-15561-1_2.
[87] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," Advances in neural information processing systems, vol. 24, 2011.
[88] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "Yfcc100m: The new data in multimedia research," Communications of the ACM, vol. 59, no. 2, pp. 64–73, 2016, doi:10.1145/2812802.
[89] A. Koochali, S. Kalkowski, A. Dengel, D. Borth, and C. Schulze, "Which languages do people speak on flickr? a language and geo-location study of the yfcc100m dataset," in Proceedings of the 2016 ACM Workshop on Multimedia COMMONS, 2016, pp. 35–42, doi:10.1145/2983554.2983560.
[90] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 11–20, doi:10.1109/CVPR.2016.9.
[91] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng, "Stylenet: Generating attractive visual captions with styles," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 3137–3146, doi:10.1109/CVPR.2017.108.
[92] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565, doi:10.18653/v1/P18-1238.
[93] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "Nocaps: Novel object captioning at scale," in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8948–8957, doi:10.1109/ICCV.2019.00904.
[94] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "Textcaps: a dataset for image captioning with reading comprehension," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 742–758, doi:10.1007/978-3-030-58536-5_44.
[95] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3558–3568, doi:10.1109/CVPR46437.2021.00356.
[96] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "Stair captions: Constructing a large-scale japanese image caption dataset," in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2017, pp. 417–421, doi:10.48550/arXiv.1705.00823.
[97] A. Alsayed, T. M. Qadah, and M. Arif, "A performance analysis of transformer-based deep learning models for arabic image captioning," Journal of King Saud

University-Computer and Information Sciences, vol. 35, no. 9, p. 101750, 2023, doi:10.1016/j.jksuci.2023.101750.

[98] B. Das, R. Pal, M. Majumder, S. Phadikar, and A. A. Sekh, "A visual attention-based model for bengali image captioning," SN Computer Science, vol. 4, no. 2, p. 208, 2023, doi:10.1007/s42979-023-01671-x.

[99] A. Rathi, "Deep learning apporach for image captioning in hindi language," in 2020 international conference on computer, electrical & communication engineering (ICCECE). IEEE, 2020, pp. 1–8, doi:10.1109/ICCECE48148.2020.9223087.

[100] A. Mathews, L. Xie, and X. He, "Senticap: Generating image descriptions with sentiments," in Proceedings of the AAAI conference on artificial intelligence, vol. 30, no. 1, 2016, doi:10.1609/aaai.v30i1.10475.

[101] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International journal of computer vision, vol. 123, pp. 32–73, 2017, doi:10.1007/s11263-016-0981-7.

[102] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in 2016 International conference on computer, information and telecommunication systems (Cits). IEEE, 2016, pp. 1–5, doi:10.1109/CITS.2016.7546397.

[103] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," Journal of the American Medical Informatics Association, vol. 23, no. 2, pp. 304–310, 2016, doi:10.1093/jamia/ocv080.

[104] J. Vaishnavi and V. Narmatha, "Video captioning–a survey," Multimedia Tools and Applications, pp. 1–32, 2024, doi:10.1007/s11042-024-18886-6.

[105] M. Abdar, M. Kollati, S. Kuraparthi, F. Pourpanah, D. McDuff, M. Ghavamzadeh, S. Yan, A. Mohamed, A. Khosravi, E. Cambria et al., "A review of deep learning for video captioning," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, doi:10.1109/TPAMI.2024.3522295.

[106] M. S. Wajid, H. Terashima-Marin, P. Najafirad, and M. A. Wajid, "Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods," Engineering Reports, vol. 6, no. 1, p. e12785, 2024, doi:10.1002/eng2.12785.

[107] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies, 2011, pp. 190–200.

[108] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5288–5296, doi:10.1109/CVPR.2016.571.

[109] M. Ravinder, V. Gupta, K. Arora, A. Ranjan, and Y.-C. Hu, "Video-Captioning Evaluation Metric for Segments (VEMS): A Metric for Segment-level Evaluation of Video Captions with Weighted Frames," Multimedia Tools and Applications, vol. 83, no. 16, pp. 47 699–47 733, 2024, doi:10.1007/s11042-023-17328-z.

[110] 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 of the Association for Computational Linguistics, 2002, pp. 311–318, doi:10.3115/1073083.1073135.

[111] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[112] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81.

[113] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575, doi:10.1109/CVPR.2015.7299087.

[114] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. Springer, 2016, pp. 382–398, doi:10.1007/978-3-319-46454-1_24.

[115] J. Giménez and L. Màrquez, "Linguistic features for automatic evaluation of heterogenous mt systems," in Proceedings of the Second Workshop on Statistical Machine Translation, 2007, pp. 256–264.

[116] Y. Li, C. Wu, L. Li, Y. Liu, and J. Zhu, "Caption generation from road images for traffic scene modeling," IEEE Transactions on intelligent transportation systems, vol. 23, no. 7, pp. 7805–7816, 2021, doi:10.1109/TITS.2021.3072970.

[117] H. Zhang, C. Xu, B. Xu, M. Jian, H. Liu, and X. Li, "TSIC-CLIP: Traffic Scene Image Captioning Model Based on Clip," Information Technology and Control, vol. 53, no. 1, pp. 98–114, 2024, doi:10.5755/j01.itc.53.1.35095.

[118] Q. Cheng, H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang, "Nwpu-captions dataset and mlca-net for remote sensing image captioning," IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1–19, 2022, doi:10.1109/TGRS.2022.3201474.

[119] M. Nivedita, P. Chandrashekar, S. Mahapatra, Y. A. V. Phamila, and S. K. Selvaperumal, "Image captioning for video surveillance system using neural networks," International Journal of Image and Graphics, vol. 21, no. 04, p. 2150044, 2021, doi:10.1142/S0219467821500443.

[120] P. Liwen, "Image Captioning-based Smart Phone Forensics Analysis Model," in Proceedings of the 2023 6th International Conference on Big Data Technologies, 2023, pp. 372–376, doi:10.1145/3627377.3627435.

[121] M. Jeon, J. Ko, and K. Cheoi, "Enhancing Surveillance Systems: Integration of Object, Behavior, and Space Information in Captions for Advanced Risk Assessment," Sensors, vol. 24, no. 1, p. 292, 2024, doi:10.3390/s24010292.

[122] Y. Mori, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Image captioning for near-future events from

vehicle camera images and motion information," in 2021 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2021, pp. 1378–1384, doi:10.1109/IV48863.2021.9575562.

[123] H. Lee, J. Song, K. Hwang, and M. Kang, "Auto-Scenario Generator for Autonomous Vehicle Safety: Multi-modal Attention-based Image Captioning Model using Digital Twin Data," IEEE Access, vol. 12, pp. 159 670–159 687, 2024, doi:10.1109/ACCESS.2024.3487588.

[124] F. Chen, C. Xu, Q. Jia, Y. Wang, Y. Liu, H. Zhang, and E. Wang, "Egocentric Vehicle Dense Video Captioning," in Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 137–146, doi:10.1145/3664647.3681214.

[125] D. Ghataoura and S. Ogbonnaya, "Application of image captioning and retrieval to support military decision making," in 2021 international conference on military communication and information systems (ICMCIS). IEEE, 2021, pp. 1–8, doi:10.1109/ICMCIS52405.2021.9486395.

[126] S. Das, L. Jain, and A. Das, "Deep learning for military image captioning," in 2018 21st International Conference on Information Fusion (FUSION). IEEE, 2018, pp. 2165–2171, doi:10.23919/ICIF.2018.8455321.

[127] L. Pan, C. Song, X. Gan, K. Xu, and Y. Xie, " Military Image Captioning for Low-Altitude UAV or UGV Perspectives ," Drones, vol. 8, no. 9, p. 421, 2024, doi:10.3390/drones8090421.

[128] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023, doi:10.48550/arXiv.2303.18223.

[129] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang et al., "A survey on evaluation of large language models," ACM Transactions on Intelligent Systems and Technology, vol. 15, no. 3, pp. 1–45, 2024, doi:10.1145/3641289.

[130] 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). IEEE, 2016, pp. 1977–1980, doi:10.1109/SIU.2016.7496155.

[131] N. Samet, S. Hiçsönmez, P. Duygulu, and E. Akbaş, "Could we create a training set for image captioning using automatic translation?" in 2017 25th Signal Processing and Communications Applications Conference (SIU). IEEE, 2017, pp. 1–4, doi:10.1109/SIU.2017.7960638.

[132] M. Kuyu, A. Erdem, and E. Erdem, "Image captioning in Turkish with subword units," in 2018 26th Signal Processing and Communications Applications Conference (SIU). IEEE, 2018, pp. 1–4, doi:10.1109/SIU.2018.8404431.

[133] B. D. Yılmaz, A. E. Demir, E. B. Sönmez, and T. Yıldız, "Image captioning in turkish language," in 2019 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2019, pp. 1–5, doi:10.1109/ASYU48272.2019.8946358.

[134] T. Yıldız, E. B. Sönmez, B. D. Yılmaz, and A. E. Demir, "Image captioning in Turkish language: Database and model," Journal of the Faculty of Engineering and Architecture of Gazi University, vol. 35, no. 4, pp. 2089–2100, 2020, doi:10.17341/gazimmfd.597089.

[135] S. B. Golech, S. B. Karacan, E. B. Sönmez, and H. Ayral, "A complete human verified turkish caption dataset for ms coco and performance evaluation with well-known image caption models trained against it," in 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME). IEEE, 2022, pp. 1–6, doi:10.1109/ICECCME55909.2022.9988025.

[136] B. Atıcı and S. İlhan Omurca, "Generating classified ad product image titles with image captioning," in Trends in Data Engineering Methods for Intelligent Systems: Proceedings of the International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2020). Springer, 2021, pp. 211–219, doi:10.1007/978-3-030-79357-9_21.

[137] S. Yıldız, A. Memi¸s, and S. Varlı, "Automatic Turkish image captioning: The impact of deep machine translation," in 2023 8th International Conference on Computer Science and Engineering (UBMK). IEEE, 2023, pp. 414–419, doi:10.1109/UBMK59864.2023.10286693.

[138] S. Yıldız, A. Memiş, and S. Varlı, "TRCaptionNet: A novel and accurate deep Turkish image captioning model with vision transformer based image encoders and deep linguistic text decoders," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 31, no. 6, pp. 1079–1098, 2023, doi:10.55730/1300-0632.4035.

[139] A. Ersoy, O. T. Yıldız, and S. Özer, "ORTPiece: An ORT-based Turkish image captioning network based on transformers and wordpiece," in 2023 31st Signal Processing and Communications Applications Conference (SIU). IEEE, 2023, pp. 1–4, doi:10.1109/SIU59756.2023.10223956.

[140] M. A. C. Ertuğrul and S. İ. Omurca, "Generating image captions using deep neural networks," in 2023 8th International Conference on Computer Science and Engineering (UBMK). IEEE, 2023, pp. 271–275, doi:10.1109/UBMK59864.2023.10286622.

[141] S. Yıldız, A. Memiş, and S. Varlı, "Turkish image captioning with vision transformer based encoders and text decoders," in 2024 32nd Signal Processing and Communications Applications Conference (SIU). IEEE, 2024, pp. 1–4, doi:10.1109/SIU61531.2024.10600738.

[142] S. Schweter, "BERTurk - BERT models for Turkish," Apr. 2020.

[143] G. Uludoğan, Z. Y. Balal, F. Akkurt, M. Türker, O. Güngör, and S. Üsküdarlı, "Turna: A turkish encoder-decoder language model for enhanced understanding and generation," arXiv preprint arXiv:2401.14373, 2024, doi:10.48550/arXiv.2401.14373.

[144] H. T. Kesgin, M. K. Yuce, and M. F. Amasyali, "Developing and evaluating tiny to medium-sized turkish bert models," arXiv preprint arXiv:2307.14134, 2023, doi:10.48550/arXiv.2307.14134.

[145] N. Tas, "Roberturk: Adjusting roberta for turkish," arXiv preprint arXiv:2401.03515, 2024, doi:10.48550/arXiv.2401.03515.

[146] H. T. Kesgin, M. K. Yuce, E. Dogan, M. E. Uzun, A. Uz, H. E. Seyrek, A. Zeer, and M. F. Amasyali, "Introducing cosmosgpt: Monolingual training for turkish

language models," arXiv preprint arXiv:2404.17336, 2024, doi:10.48550/arXiv.2404.17336.
[147] H. Türkmen, O. Dikenelli, C. Eraslan, M. C. Callı, and S. S. Özbek, "Bioberturk: Exploring turkish biomedical language model development strategies in low-resource setting," Journal of Healthcare Informatics Research, vol. 7, no. 4, pp. 433–446, 2023, doi:10.1007/s41666-023-00140-7.
[148] OpenAI, "ChatGPT," https://chat.openai.com/chat, 2023.
[149] Google, "Bard," https://bard.google.com/, 2023.
[150] Google, "Gemini," https://gemini.google.com/, 2024.

Özgeçmişler

Dr. Abbas Memiş, 2010 yılında Yıldız Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Bilgisayar Mühendisliği Bölümü'nden mezun olmuştur. Yüksek lisans ve Doktora derecelerini ise sırasıyla 2013 ve 2020 yıllarında Yıldız Teknik Üniversitesi, Fen Bilimleri Enstitüsü, Bilgisayar Mühendisliği Anabilim Dalı'ndan almıştır. Şu anda İstanbul Üniversitesi, Bilgisayar ve Bilişim Teknolojileri Fakültesi, Bilgisayar Mühendisliği Bölümü'nde Dr. Öğr. Üyesi olarak çalışmakta ve ağırlıklı olarak bilgisayarla görme, görüntü işleme, medikal görüntü analizi, yapay zekâ ve tıpta/sağlıkta yapay zekâ konularında araştırma çalışmaları yürütmektedir.