TÜBİTAK–2242 ÜNİVERSİTE ÖĞRENCİLERİ ARAŞTIRMA PROJE YARIŞMALARI

BİLGİ VE İLETİŞİM TEKNOLOJİLERİ

GÖRME ENGELLİ BİREYLER İÇİN AKILLI TELEFONLARDA ÇOK KATMANLI GRU TABANLI OTOMATİK GÖRÜNTÜ ALTYAZILAMA

BAŞVURU SAHİBİ: Rumeysa KESKİN
DANIŞMAN: Dr. Öğr. Üyesi Volkan KILIÇ

2021

1. Dönem Başvurusu

İÇİNDEKİLER

ŞEKİLLER LİSTESİ

TABLO LİSTESİ

KISALTMALAR

BLEU	Bilingual Evaluation Understudy
CIDEr	Consensus-Based Image Description Evaluation
ESA	Evrişimsel Sinir Ağ
GRU	Gated Recurrent Unit
LSTM	Long Short-Term Memory
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SPICE	Semantic Propositional Image Caption Evaluation
TSA	Tekrarlayan Sinir Ağ

ÖZET

Görüntü altyazılama, bilgisayarlı görü ve doğal dil işleme alanlarını kullanarak, bir görüntünün doğal dil ifadeleriyle tanımlanmasıdır. Son teknolojik gelişmelerle birlikte donanım ve işlemci gücünün akıllı telefonlarda ileri düzeye taşınması, görüntü altyazılama ile ilgili birçok uygulamanın geliştirilmesinin önünü açmıştır. Bu çalışmada, akıllı telefonlarda uygulanabilecek kodlayıcı-kod çözücü yaklaşımına dayanan özgün bir otomatik görüntü altyazılama sistemi önerilmektedir. Kodlayıcı kısmında ResNet152V2 evrişimsel sinir ağı ile yüksek seviyeli görsel bilgiler çıkarılırken önerilen kod çözücü, çıkarılan görsel bilgileri görüntülerin doğal ifadelerle tanımlanmış altyazılara dönüştürmektedir. Önerilen kod çözücüde tekrarlayan sinir ağı çok katmanlı kapılı tekrarlayan hücre tabanlı yapısı, en faydalı görsel bilgileri kullanarak daha anlamlı altyazı üretilmesini sağlamaktadır. Önerilen sistem, MSCOCO veri kümesi üzerinde farklı performans metrikleri kullanılarak test edilmiş ve literatürdeki çalışmalarla kıyaslanarak sağladığı üstünlük gösterilmiştir. Evrişimsel Sinir Ağları ve Tekrarlayan Sinir Ağları yüksek işlem gücü gerektiren derin sinir ağları kullanır. Ancak düşük işlem gücü ve düşük bellek alanına sahip mobil cihazlar bu yapıların doğrudan cihaz üzerinde çalışma ihtiyacını karşılayamaz. Bu çalışmada önerilen sistem, cihaz üzerinde internet bağlantısı olmadan, hızlı, düşük güç tüketimiyle altyazı üretebilecek şekilde geliştirilmiş ve *IMECA* adlı Android uygulama ile birleştirilmiştir. Uygulamanın hareket tabanlı ekran okuma, sesli olarak komut alma ve bilgilendirme özellikleri ile görme engeli veya kısmi görme kaybı olan bireyler için yardımcı bir platform sunulmuştur. Ayrıca farklı dil seçenekleriyle görüntü altyazılama uygulamasının daha geniş bir hedef kitlesine ulaşması amaçlanmıştır.

Anahtar kelimeler: Derin öğrenme, doğal dil işleme, Android uygulama

1. GİRİŞ

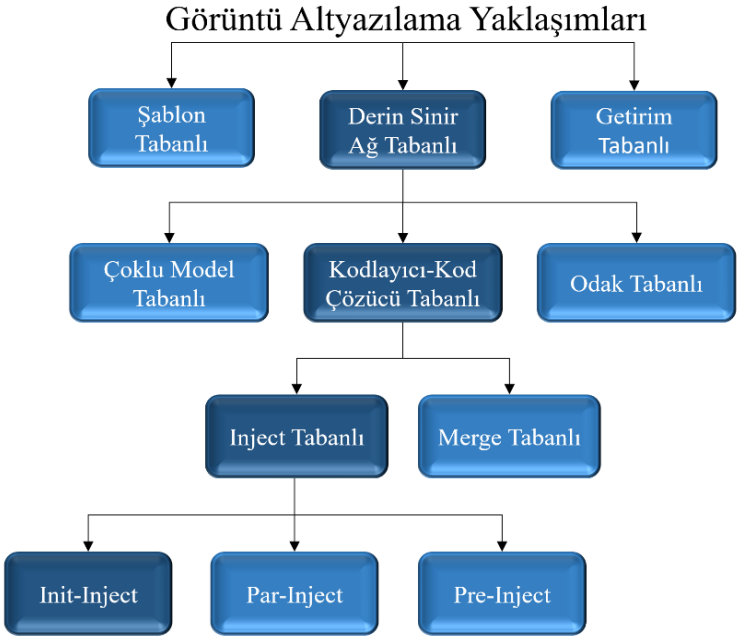
Gelişen teknolojiyle modern yaşamın her alanında ürünler sunmakta olan yapay zekânın birçok görme engelli bireyin yaşam kalitelerini iyileştirme yönündeki çalışmaları oldukça önem arz etmektedir. Bilgisayarlı görü ve doğal dil işleme teknikleri kullanılarak bir görüntünün dil bilgisi kurallarına uygun ve anlamlı bir şekilde tanımlanmasını amaçlayan görüntü altyazılama problemi, son yıllarda görsel arama, görüntü dizinleme ve görme engelli bireyler için sanal asistanlar gibi uygulamalarıyla ilgi çekmektedir (Makav & Kılıç, 2019a), (Gong & Zhang, 1994), (Makav & Kılıç, 2019b). Literatürde görüntü altyazılama problemine yönelik çalışmaların çoğu, en başarılı ve en iyi bilinen yöntem olarak benimsenmiş derin sinir ağlarına dayanmaktadır. (Literatürdeki görüntü altyazılama yaklaşımları için bkz. Şekil 1.)

Derin sinir ağlarına dayalı görüntü altyazılama yaklaşımlarında, görüntü özniteliklerini çıkararak sabit boyutlu bir vektöre "kodlayan" evrişimsel sinir ağı ve bu görüntü bilgisini istenen görüntü açıklamasında "kodu çözmek" için tekrarlayan sinir ağı kullanır. Teorik ilerlemeler ile birlikte kodlayıcı tasarımlarında kullanılabilecek Inception-v3, ResNet gibi gelişmiş evrişimsel sinir ağları (ESA) mimarileri ortaya çıkmıştır. Ancak, kod çözücüde geleneksel tekrarlayan sinir ağları (TSA) yapıları kaybolan veya patlayan gradyan problemlerine yol açtığından, geliştirilen uzun kısa dönemli bellek (long short-term memory - LSTM) ve kapılı tekrarlayan hücre (gated recurrent unit - GRU) yapıları kullanılmaktadır (Hossain, Sohel, Shiratuddin, & Laga, 2019). LSTM yapısı, bilgiyi depolamak için uzun periyotlarda hafıza hücreleri kullanırken GRU yapısı ek hafıza hücrelerine gerek duymadan bilgi akışına izin vererek hesaplama verimliliği ve işlem kolaylığı sağlamaktadır (Wang, Wang, & Xu, 2020). TSA tabanlı kod çözücüler, kodlayıcıdan gelen görsel bilgiyi TSA yapısına farklı durumlarda besleyerek işlem yapabilmektedirler (Tanti, Gatt, & Camilleri, 2018). Örneğin, kod çözücü mimarilerinden biri olan init-inject mimari yaklaşımı, görsel bilgiyi bir görüntü vektörü olarak RNN başlangıç gizli katmanına besleyerek daha fazla görsel bilgiden yararlanmaktadır.

Araştırmacılar İngilizce için çok sayıda çözüm önermiş olsalar da, Türkçe görüntü altyazılama modellerini eğitmek için uygun veri kümelerinin bulunmamasından dolayı TasvirEt (Unal et al., 2016), (Kuyu, Erdem, & Erdem) gibi çok az sayıda çalışmada Türkçe veri kümesi oluşturularak görüntü altyazılama yaklaşımı önerilmektedir. Veri işleyen yaklaşımların başarısı, veri miktarıyla doğru orantılı olduğundan, bu çalışmada veri kümesinin zenginleştirilmesi ve verilerin çoğaltılması faydalı olacaktır. Ayrıca, Türkçe İngilizce'ye göre farklı dil bilimsel özelliklere sahiptir ve dilbilgisine uygun altyazılama algoritmalarının geliştirilmesinde daha fazla iş gücü gerektirmektedir. Yüksek işlem gücü gerektiren derin sinir ağlarının büyük veri kümeleriyle eğitimleri sonucunda modellerin başarılı sonuçlar sunmasına karşın bu sistemler düşük güçlü gömülü cihazlar üzerinde doğrudan çalışma ihtiyacını karşılayamazlar. Literatürde, Makav vd. (2019b) ve Çaylı vd., (2020) tarafından geliştirilmiş sistemde, akıllı telefon uygulamasında görüntü altyazısı üretimi için Firebase bulut sistemini kullanılmaktadır. Bulut sistemi, her ne kadar büyük model boyutuna sahip derin öğrenme algoritmalarının düşük güçlü cihazlarda çalışabilmeleri için etkin bir yöntem olsa da, fazla güç tüketimi, uzun işlem süresi, internet gereksinimi, veri gizliliği gibi sorunları gündeme getirir.

Önerilen çalışma, MSCOCO veri kümesinin görüntü ve İngilizce altyazıları üzerinde gerçekleştirilmiştir. Bu çalışma şu katkıları sağlamaktadır:

- Kodlayıcı-kod çözücü yaklaşımı kullanılarak init-inject yapısında geliştirilen n-katmanlı (n = 3, 6, 9, 12, 15) GRU-tabanlı modeller önerilmiştir. Eğitilen modeller çeşitli performans metrikleri ile değerlendirilmiş ve n-katmanlı GRU-tabanlı modellerin bir noktaya kadar verilerin karmaşık ilişkilerini öğrenme becerisini artırdığı sonucuna ulaşılmıştır.
- Görme engelli bireylerin sesli komut ve sesli bilgilendirme ile kullanabilecekleri Android-tabanlı akıllı telefon uygulaması geliştirilmiş, Türkçe dahil 44 farklı dilde çevrimdışı çeviri özelliği oluşturulmuştur.
- Önerilen görüntü altyazılama sistemi, akıllı telefon uygulamasına sunucu kullanılmadan uygulanabilir hale getirilmiş ve altyazı üretimi, internete ihtiyaç duymaksızın, hızlı, düşük güç tüketimi ve veri gizliliği sağlanarak cihaz üzerinde gerçekleştirilmiştir.



Şekil 1 Literatürdeki görüntü altyazılama yaklaşımları.

1. 1. Projenin Amacı ve Önemi

Gelişen ve değişen teknolojinin sunduğu kolaylıklar en çok engelli bireyler için önem arz etmektedir. Bu bireylerin toplum içinde çevresindeki bireylere bağımlı olmadan, olabildiğince toplumsal yaşama aktif biçimde uyum sağlayabilmeleri için eğitimden sağlığa, ulaşımdan sosyo-kültürel faaliyetlere kadar ciddi bir şekilde çözüm bekleyen sorunları bulunmaktadır. Bu sorunlardan yola çıkarak, çalışmamızda görme engelli bireylerin çevresinde olanları daha rahat anlayabileceği, kullanıcılarına etrafında gelişen olayları doğal bir dille ifade edebilecek bir akıllı telefon uygulaması yapılması hedeflenmiştir. Bu projede, yapay zekânın doğal dil işleme ve derin öğrenme uygulama alanları kullanılarak görüntüyü anlamlı bir cümleyle ifade etme olan görüntü altyazılama problemi üzerinde çalışılmıştır. Geliştirilmiş

Android-tabanlı uygulama ile kullanıcı ve telefon arasında sesli iletişim sağlanarak görüntü altyazılama yapan bir platform sunulmuştur. Yazılım ile geliştirilen ürünlerin donanımsal ürünlere göre düşük maliyetli ve ekonomik kâr oranının yüksek olması, projemizin büyük ekonomik paya sahip olan yazılım ürünlerine katkıda bulunmasını sağlayacaktır.

Bu çalışmada gerekli altyapının oluşması adına AdresGezgini A. Ş. ile ortak bir çalışma yürütülmüştür. AdresGezgini A. Ş., dijital pazarlama çözümleri ve web tabanlı yazılım geliştirme projeleri alanında faaliyet göstermekte ve aktif olarak doğal dil işleme, sinyal işleme, pekiştirmeli öğrenme alanlarında TÜBİTAK TEYDEB projeleri yürütmektedir. Bu projeyle görüntü altyazılama modelleri üzerinde gerekli bilgi birikimi akademi-sanayi iş birliği çatısı altında firmaya kazandırılması sağlanacaktır. Bu sayede, chatbot uygulamalarının uluslararası alanda rekabet edebilecek seviye gelmesi ve ülke ekonomisine katkı sağlayacak yerli ve milli yazılımların geliştirilmesine katkı sağlanacaktır.

1. 2. Projenin İçerdiği Yenilik Unsuru

Günümüz en popüler konularından biri olan ve binlerce alanda kullanılan yapay zekâ uygulamalarının özellikle akıllı telefonlarda kullanımının artması, bu ürünleri insan hayatının önemli bir parçası haline getirmiştir. Görüntü tanıma, konuşma tanıma, doğal dil işleme alanlarında giderek artan çalışmalar yapılmaktadır (Mathur, Gill, Yadav, Mishra, & Bansode, 2017), (Karpathy & Fei-Fei, 2015). Microsoft tarafından geliştirilen CaptionBot (Qiuyuan, Pengchuan, Oliver, & Lei, 2018) ve Çin'de Baidu A.Ş. tarafından görme engelli bireylerin hayatlarını kolaylaştırması amacıyla geliştirilen DuLight (DuLight, 2015) gibi uygulamalar bu alandaki çalışmalara örnek olarak gösterilebilir. Ancak bu uygulamalar Türk kullanıcılara hitap etmemekte ve bilgimiz dahilinde Türkiye'de böyle bir uygulama bulunmamaktadır.

Önerilen görüntü altyazılama sistemi için deneyler MSCOCO (Chen et al., 2015) veri kümesinin mevcut görüntü ve İngilizce görüntü altyazılarıyla gerçekleştirilmiş ve görüntü altyazılama modellerinin performanslarını iyileştirmek için farklı yöntemler uygulanmıştır. Önerilen sistem, akıllı telefonlarda internet bağlantısı olmadan altyazı üreten geliştirilen *IMECA* adlı Android uygulama ile birleştirilmiştir. Böylece geliştirilen model, görme engelli bireyler için geliştirilmiş bulut sistemi kullanarak altyazı üreten uygulamalardan (Makav & Kılıç, 2019b), (Çaylı et al., 2020) farklı olarak internet bağlantısına ihtiyaç duymadan, hızlı, veri gizliliğine uygun ve düşük güç tüketimiyle tamamen cihaz üzerinde çalışmaktadır. Hareket tabanlı ekran okuma, sesli olarak komut alma ve bilgilendirme özellikleri ile görme engeli veya kısmi görme kaybı olan bireyler için geliştirilmiş bu uygulama, kullanıcılara farklı dil seçenekleri sunarak uluslararası düzeyde kullanıcı kitlesine ulaşmayı hedeflemiştir.

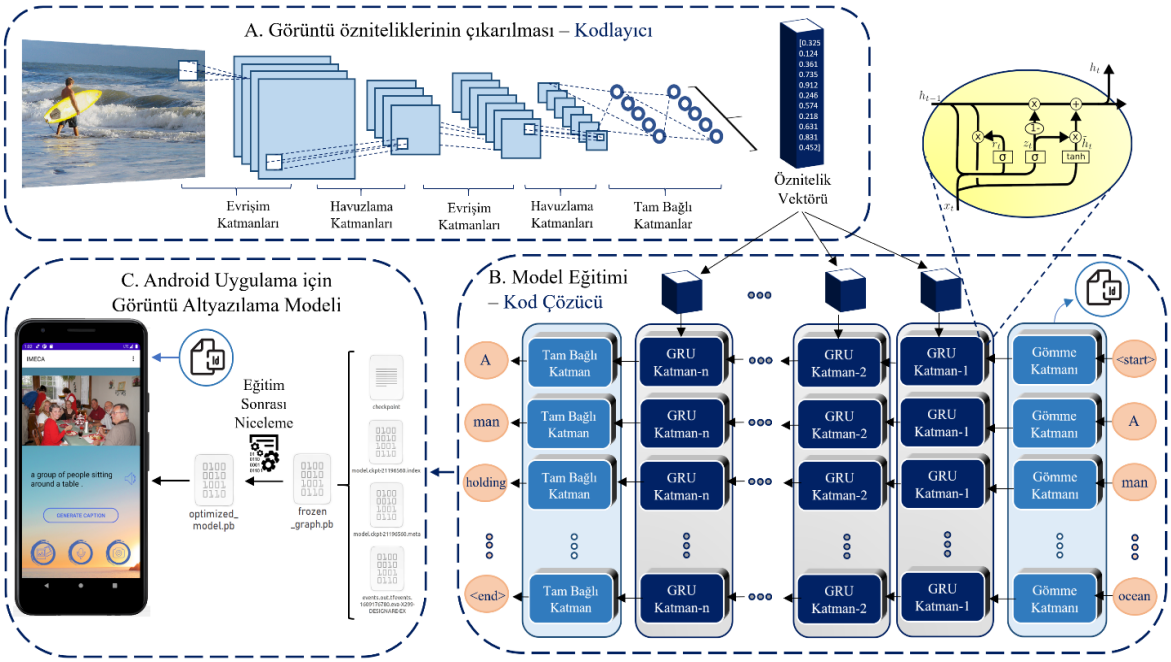
Amerika, İsviçre, Çin ve Kanada gibi bazı ülkelerde birçok üniversite ve büyük çaplı şirketler bilgisayarlı görü, doğal dil işleme, görüntü işleme alanlarında çalışmalar yapmakta ve bu alanlardaki başarılı uygulamaların ticarileştiği görülebilmektedir (Şeker, Diri, & Balık, 2017).

1. 3. Projenin İlgili Olduğu Teknolojik Alanı

Gelişen teknoloji daha hızlı, güvenli, verimli, kişiye özel ve düşük maliyetli çözümler aramaktadır. Yapay zekânın yenilikçi çözümleri hem dijital hem fiziksel hayatta farklı alanlardaki problemlere uygulanmakta ve büyük başarılar elde etmektedir. Çalışmamızda, derin öğrenme ve doğal dil işleme alanlarının bir arada kullanıldığı görüntü altyazılama sistemi, Türkçe dahil birçok dilde kullanıma uygun bir platform üzerinden hedeflenen kullanıcı kitlesine rehberlik edecek, bireylerin bir dizi basit hareketle ve/veya konuşarak kontrol edebilecekleri aynı zamanda bilgilendirilecekleri, kullanıcı ile uyumlu bir uygulama üzerinden sunulmaktadır. Bilgi ve iletişim teknolojilerinin sunduğu imkanları kullanarak, görme engelli bireylerin ihtiyaçları doğrultusunda insan ve teknoloji unsurlarını bir araya getiren bu proje, Dijital Dönüşüm teknolojik alanına dahildir.

2. YÖNTEMLER VE TEKNİKLER

Bu bölümde, akıllı telefonlarda uygulanabilecek kodlayıcı-kod çözücü yaklaşımına dayanan otomatik görüntü altyazılama sistemi, Android-tabanlı uygulama ve model optimizasyonu sunulmaktadır. Çalışmaya ait iş akış şeması Şekil 2'de gösterilmektedir.



Şekil 2 Önerilen görüntü altyazılama sistemi ve Android uygulaması.

2. 1. Önerilen Kodlayıcı-Kod Çözücü Yaklaşımı

Önerilen kodlayıcı-kod çözücü yaklaşımındaki görüntü altyazılama sisteminde, kodlayıcı olarak en iyi görüntü özniteliklerini çıkarma kapasitesinden dolayı ESA mimarisi ve anlamlı altyazılar üretmek için sıralı verileri işleyebilen TSA mimarisi kullanılmaktadır.

2. 1. 1. Evrişimsel Sinir Ağları

Evrişimsel sinir ağları, görüntü analizi ve sınıflandırma, nesne bulma, nesne takip etme, doğal dil işleme gibi uygulama alanlarında çoğunlukla görsel görüntüleri analiz etmek için kullanılan derin sinir ağları sınıfıdır. ESA'lar, bu işlevselliği elde etmek için görüntüyü; evrişim katmanları, havuzlama katmanları ve tam bağlı katmanlardan oluşan birçok yapı bloğunu kullanarak işler. Evrişim katmanında evrişim işlemi yapan filtreler ile görüntüde tespit edilmesi beklenilen nesnelere ait bir öznitelik haritası çıkarılır. Havuzlama katmanı ağın hesaplama karmaşıklığını azaltmak ve modelin etkili bir şekilde eğitilme sürecini sürdürmek için öznitelik haritasının uzamsal boyutunu düşürür. Tam bağlı katman, bir önceki katmandaki verilere bağlı olarak son çıktıyı üretir.

Derin sinir ağları büyük veri kümeleri üzerinde eğitildiklerinde daha iyi performans gösterirler. ESA modelini sıfırdan eğitmek hesaplama açısından uzun bir süreç gerektirir ve maliyetli olmaktadır. Bu nedenle Inception, ResNet gibi büyük veri kümelerinde (örn. ImageNet) önceden eğitilmiş ESA mimarileri kullanılmaktadır. Bu çalışmada, 152 katmandan oluşan ResNet152V2 mimarisi kullanılmıştır. Bu mimari, ResNetV1 mimarisinden farklı olarak her ağırlık katmanından önce yığın normalleştirmesi kullanır. Sınıflandırma katmanı çıkarılmış ResNet152V2 mimarisi evrişim ve havuzlama katmanları kullanarak girdi görüntüsünün yüksek seviyeli görüntü öznitelik vektörünü çıkarır. ESA katmanları Şekil 2 - A kısmında gösterilmektedir.

2. 1. 2. Tekrarlayan Sinir Ağları

Cümlelerdeki kelimeler gibi birbirleri ile ilişkili olan dizisel girdi ve çıktıların olduğu problemlerde tekrarlayan sinir ağları kullanılmaktadır. TSA gömme, tekrarlayan gizli katmanlar ve tam bağlı katmanlardan oluşur. Gömme katmanında sözcükler vektör olarak ifade edilirken tekrarlayan gizli katmanlar işlenen dizinin her sözcüğünün yüksek bir kesinlikle tahmin edilebilmesi için kendisini eğitir ve tam bağlı katman görüntü özniteliklerine karşılık gelen en uygun sözcüğü tahmin eder. Katmanların her biri birbirine ağırlıklı kenarlarla bağlanmıştır ve her sinir ağının öğrenim aşamasında gradyan hesaplaması yapılırken türev alınarak en uygun ağırlıklar, yani görüntü açıklamasını, sınıfları temsil eden sabit uzunluklu bir vektör üzerinde sınıflandıracak en uygun değerler bulunmaya çalışılmaktadır. Ancak öğrenilen dizilerin artmasıyla alınan türev değerleri kaybolan gradyan problemine neden olmaktadır. Bu problemden kaynaklanan öğrenmenin yavaşlaması ve alakasız noktaların öğrenilmeye başlanması gibi sorunlara çözüm üretmek için geliştirilen GRU yapısı dizisel verileri modellemede büyük başarı göstermiştir.

Literatürde tek katmanlı TSA ile yapılandırılmış kod çözücü mimarileri mevcuttur (Mellit, Kalogirou, Hontoria, Shaari, & Reviews, 2009). Mevcut çalışmalar birden fazla katmanla yapılandırılmış TSA'ların daha karmaşık gösterimleri hesaplamasına ve sıralı verileri öğrenme becerilerinin artarak daha başarılı tahminler yapmasına olanak sağladığını göstermektedir (Sutskever, Vinyals, & Le, 2014), (Rahman, Srikumar, & Smith, 2018). Önerilen çalışmada, verilerin karmaşık ilişkilerini öğrenme becerisini artırmak ve modelin daha başarılı altyazı üretmesini sağlamak için çok katmanlı GRU yapısı geliştirilmiştir. ResNet152V2 ile çıkarılan öznitelik vektörü katman sayısına bağlı olarak eşit boyutta n-vektöre

ayrılmıştır. Her bir vektör sırasıyla n-katmanlı GRU yapısının başlangıç gizli katmanlarını beslemektedir. Önerilen model yapısı Şekil 2 - B kısmında gösterilmektedir. Gerçekleştirilen deneylerde, katman sayısının altyazı üretme başarısına olan etkisini gözlemlemek için 3, 6, 9, 12, ve 15 katmanlı GRU tabanlı kod çözücüleri değerlendirilmiştir.

2. 2. Android Tabanlı Uygulama ve Model Optimizasyonu

Önerilen otomatik görüntü altyazılama sisteminin mobil ve gömülü cihazlarda verimli bir şekilde çalışmasını sağlayan TensorFlow kütüphanesinden yararlanılmıştır. Model kodlayıcısında görüntü öznitelik çıkarımı için TensorFlow tarafından sunulan dondurulmuş CNN modelleri (Marks, 2020) mevcuttur. Inception-v3, diğer CNN modellerinden farklı olarak dondurulmuş model dosyasına sahiptir ve 3 katmanlı GRU tabanlı dil modeli ile kodlayıcı-kod çözücü yapısında kullanılmıştır. Çalışmanın eğitim süreci ve modelin oluşturulmasında uygulanan adımlar şu şekilde devam etmektedir: Modelin her eğitim epoğundan sonra eğitim ağırlıklarını içeren kontrol noktaları dosyası oluşturulur. Bununla birlikte modelin çıkarımı ve eğitiminde yeniden kullanılması için model ağırlıklarının anlamlı bir şekilde ilişkilendirilmesiyle ilgili bilgileri içeren hesaplama grafiğinin tanım dosyaları ve meta verileriyle ilgili dosyalar oluşturulur. Eğitimde elde edilen modele ait tüm bu parametre değerleri dondurulmuş bir ProtoBuf dosyasında kaydedilir. Dondurma işlemindeki amaç, tüm bilgilerin tek bir dosyada kolayca kullanılabilmesini sağlamaktır.

Modelin çalıştırılması istenilen mobil cihazların düşük işlem gücü ve düşük bellek alanına sahip olması ve işlem sonucunda elde edilen model boyutunun oldukça büyük olması, model boyutunda optimizasyon gerektirmektedir. Önerilen modelin optimizasyonu için eğitim sonrası 8-bit niceleme (Leon et al., 2020) yöntemi kullanılarak model eğitiminde veri kaybı olmadan model boyutunun ¼'üne düşürülmesi sağlanmıştır. Noktadan sonra 32-bit derin öğrenme dahil çoğu uygulama için varsayılan değerdir. Bunun yanı sıra derin sinir ağları 8 bitlik sayılarla da daha düşük hassasiyetle çalışabilmektedir. Önerilen niceleme yöntemi, float32 sayılarının en yakın küçük bitlere yuvarlanması işlemidir. Daha düşük bit derinliğinin aritmetiği, 32-bite göre her zaman daha hızlıdır, bellekte neredeyse 4 kat azalma sağlar ve düşük güçlü gömülü cihazlarda desteklenebilir. Optimize edilmiş gömülü model dosyası ve sözcük öznitelik dosyası Android Studio ile geliştirilen *IMECA* uygulamasına entegre edilmiştir. Uygulama, ayrıca, görme engelli bireyler için hareket tabanlı ekran okuma, sesli komut alma ve bilgilendirme, farklı dil ve ses seçenekleri sunmaktadır. Cihaz üzerinde dil çevirisi için Google tarafından geliştirilen ML kit kullanılmıştır. Bu bölümde uygulanan adımlar Şekil 2 - C kısmında gösterilmektedir.

2. 3. Veri Kümesi

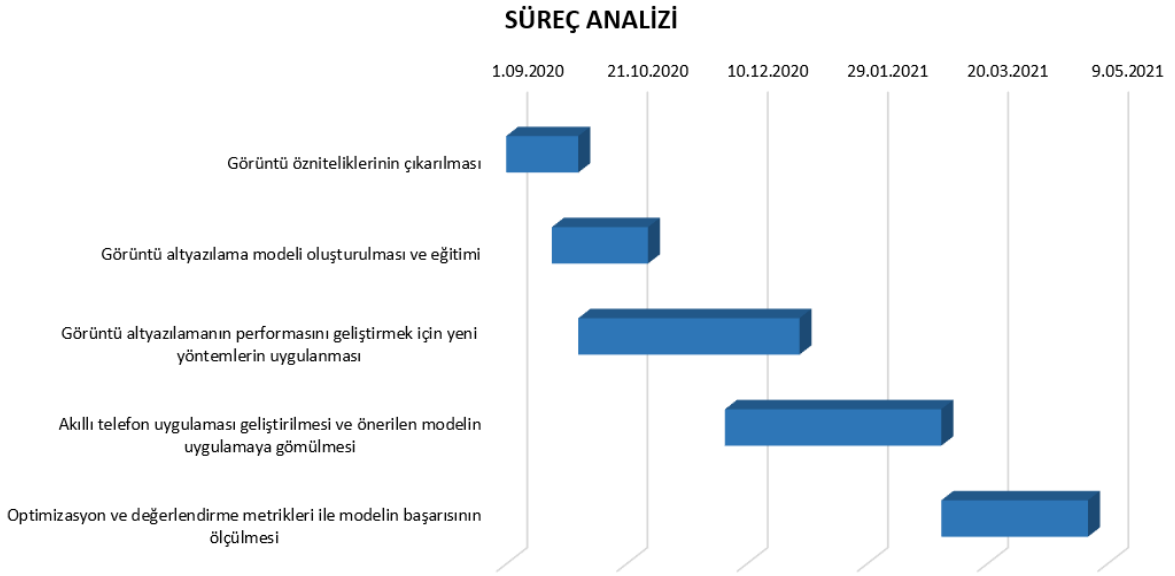
Flickr30k (Young, Lai, Hodosh, & Hockenmaier, 2014), VizWiz (Bigham et al., 2010), MSCOCO gibi veri kümeleri içerikleri itibariyle görüntü altyazılama sistemlerinde yaygın olarak kullanılan veri kümeleridir. Flickr30k, Flickr web sitesinden alınmış yaklaşık 32 bin görüntü ve her görüntü için beş adet referans altyazı içerir. Flickr30k veri kümesindeki altyazılar, birden fazla nesnenin olduğu bir görselde, ana nesneyi veya ön planda olan iki nesneye odaklanırlar. Yaklaşık olarak 39 bin görüntü ve her biri için beş referans altyazı içeren Vizwiz, görme engelli kişilerin yaklaşık 10 yıldır çektikleri fotoğraflardan oluşan bir veri kümesi sunmaktadır. Önerilen görüntü altyazılama sisteminin değerlendirilmesi için referans altyazılarla birlikte geniş sayıda görüntü içeren bir veri kümesi gereklidir. Mevcut durumda 123 bin görüntü ve her görüntü için beş adet altyazı içermekte olan MSCOCO veri kümesi, özellikle piksel düzeyindeki görüntülerde nesnelerin yerini belirleme, nesnelerin ikonik olmayan görünümlerini algılama gibi problemler için Microsoft Teams tarafından geliştirilmiştir ve doğal sahnelere odaklanan referans altyazılar sunmaktadır. Görüntü altyazılama çalışmasında MSCOCO veri kümesi dilbigisel ve anlamsal olarak başarılı performans sergilemesi nedeniyle tercih edilmiştir.

2. 4. Performans Metrikleri

Çalışmalarımızda görüntü altyazılarının başarıları değerlendirmek için BLEU-n (n=1, …,4) (Papineni, Roukos, Ward, & Zhu, 2002), METEOR (Lavie & Agarwal, 2007), ROUGE-L (Lin, 2004), CIDEr (Vedantam, Lawrence Zitnick, & Parikh, 2015) ve SPICE (Anderson, Fernando, Johnson, & Gould, 2016) gibi otomatik değerlendirme metrikleri kullanılmıştır. BLEU-n, METEOR ve ROUGE-L, makine çevirisi sistemlerini değerlendirmek için geliştirilmiş metriklerdir ve makine tarafından üretilen cümle ile insan tarafından üretilen referans cümle arasındaki sözcüksel benzerliği dikkate alarak bir değerlendirme yapmaktadırlar. Buna karşılık, özellikle görüntü altyazılarını değerlendirmesi için geliştirilen CIDEr ve SPICE metrikleri insan yargısına değişken derecede benzerlik göstermekte ve anlatımda görüntüde göze çarpan nitelikleri ve nitelikler arasındaki ilişkiyi dilbilgisel ve anlamsal olarak değerlendirmektedir. Bu nedenle daha verimli altyazılar üreten modeller üzerinde çalışılması için sonuçların karşılaştırılmasında CIDEr ve SPICE metriklerine öncelik verilmiştir.

2. 5. Proje İş-Zaman Çizelgesi

İş-Zaman çizelgesi yöntemlerde verilen alt başlıkları iş paketleri ile ilişkilendirilerek Şekil 3'te verilmiştir.



	Görüntü özniteliklerinin çıkarılması	Görüntü altyazılama modeli oluşturulması ve eğitimi	Görüntü altyazılamanın performasını geliştirmek için yeni yöntemlerin uygulanması	Akıllı telefon uygulaması geliştirilmesi ve önerilen modelin uygulamaya gömülmesi	Optimizasyon ve değerlendirme metrikleri ile modelin başarısının ölçülmesi
Başlangıç zamanı	1.09.2020	20.09.2020	1.10.2020	1.12.2020	1.03.2021
Süreç (gün)	30	40	92	90	61

Şekil 3 Projeye ait iş-zaman çizelgesi.

3. DENEYSEL SONUÇLAR

Önerilen görüntü altyazılama sisteminin kodlayıcı-kod çözücü modellerinin deneylerinde Keras kütüphanesi; geliştirilen Android uygulama için gerçekleştirilen çalışmalarda Android Inference Library ve Java API'nın destekledigi TensorFlow 1.13.1 kullanılmıştır.

Önerilen çok katmanlı GRU tabanlı kod çözücü, ResNet152V2 mimarisi ile beş farklı katman sayısı kullanılarak değerlendirilmiştir. Eğitilen farklı katmanlardaki GRU-tabanlı modellerin doğruluk performansları Tablo 1'de sunulmuştur. BLEU-n ve ROUGE-L metriklerinde 9 katmanlı GRU modeli; SPICE, METEOR, CIDEr metriklerinde 6 katmanlı GRU modeli en başarılı performansı göstermiştir. 3 katmanlı GRU-tabanlı model karmaşık özellikleri çıkarmak için yetersiz kalmış, artan katman sayısıyla geri besleme etkisinin ilk katmanlara daha az ulaşması, belli bir noktadan sonra model başarısında düşüşe neden olmuştur. Tablo 2'de önerilen sistemin ve MSCOCO veri kümesinde eğitilmiş farklı modellerin kıyaslaması verilmiştir. Önerilen görüntü altyazı sisteminin beş metrikte üstün olduğu görülmektedir. Şekil 4, MSCOCO doğrulama veri kümesinden alınmış iki örnek görsel için referans altyazılar ve önerilen katmanlardaki modeller tarafından üretilen altyazılar sunulmaktadır. 6 katmanlı GRU modeli tarafından üretilen altyazıların diğer altyazılara kıyasla

resimleri daha detaylı bir şekilde ifade ettiği ve daha başarılı altyazılar sunduğu gözlemlenmektedir. *IMECA* uygulaması ve üretilen altyazı örnekleri Şekil 5'te verilmiştir.

Tablo 1 ResNet152V2 mimarisi ile beş farklı GRU katmanıyla eğitilen modellerin doğruluk performansları.

	BLEU1	**BLEU2**	**BLEU3**	**BLEU4**	**ROUGE-L**	**SPICE**	**METEOR**	**CIDEr**
3-katman	0.679	0.494	0.350	0.244	0.488	**0.150**	0.219	0.782
6-katman	0.675	0.492	0.349	0.248	0.490	**0.150**	**0.221**	**0.786**
9-katman	**0.683**	**0.498**	**0.352**	**0.249**	**0.493**	0.148	0.219	0.778
12-katman	0.546	0.326	0.184	0.105	0.414	0.090	0.152	0.420
15-katman	0.544	0.325	0.182	0.104	0.411	0.093	0.151	0.421

Tablo 2 Farklı kodlayıcı-kod çözücü modellerinin performans metrik sonuçları.

	BLEU1	**BLEU2**	**BLEU3**	**BLEU4**	**ROUGE-L**	**SPICE**	**METEOR**
ResNet152-LSTM (You, Jin, & Luo, 2018)	0.512	0.306	0.188	0.116	0.384	0.172	0.611
VGG16-LSTM (Xu et al., 2019)	0.662	0.479	0.335	0.231	0.481	0.214	0.706
VGGNet-mRNN (Mao et al., 2014)	0.668	0.488	0.342	0.239	0.489	0.221	0.729
CNN-1 katmanlı GRU (Li, Yuan, Lu, & Applications, 2018)	0.668	0.459	0.331	0.229	-	**0.255**	0.733
Önerilen ResNet152 6-katmanlı GRU	**0.675**	**0.492**	**0.349**	**0.248**	**0.490**	0.221	**0.786**

Reference captions:

(1) Two zebras standing next to each other in a dry grass field.
(2) Two zebras standing in tall savannah grass near forest brush.
(3) Two zebras are walking through tall brown grass.
(4) Two zebras standing in a field of tall grass.
(5) Two zebras roaming through the terrain in a countryside setting.

Generated captions:

3-layer captioner: two zebras standing in the grass in the wild
6-layer captioner: two zebras standing in a grassy field with trees
9-layer captioner: two zebras are standing in a dry field
12-layer captioner: two zebras standing in a a in field
15-layer captioner: two zebras standing standing standing in in field

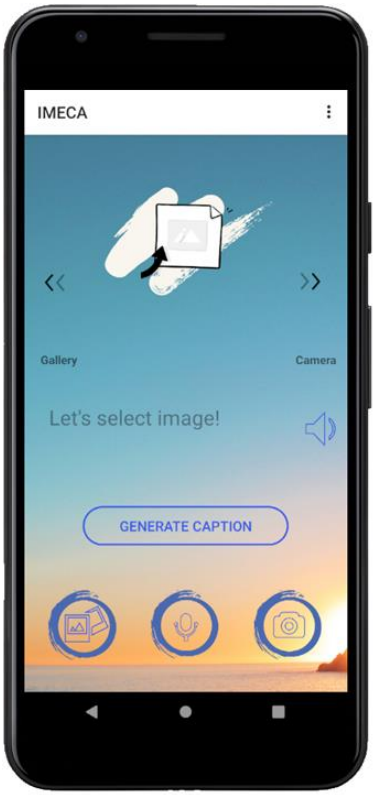
Reference captions:

(1) An old-fashioned steam engine train traveling down railroad tracks.
(2) A train engine that is letting out smoke travelling down a railroad track with multiple passenger cars attached.
(3) An old-fashioned train riding on train tracks in a wooded area.
(4) A train passing through wooded areas on a train track.
(5) A locomotive on train tracks in a wooded countryside.

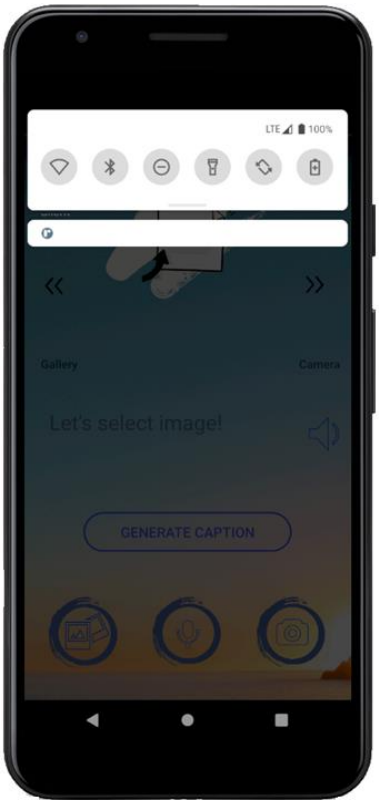
Generated captions:

3-layer captioner: a train traveling through a lush green forest
6-layer captioner: a steam train traveling down tracks next to a forest
9-layer captioner: a steam engine traveling through a lush green forest
12-layer captioner: a train train coming train the the tracks
15-layer captioner: a steam locomotive locomotive down down a tracks

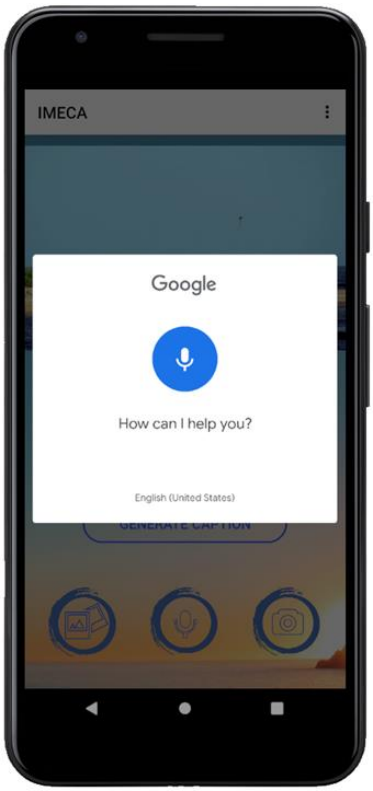
Şekil 4 MSCOCO doğrulama veri kümesinden alınmış resimlere (soldaki görsel: 000000052462.jpg, sağdaki görsel: 000000423229.jpg) ait referans ve üretilen altyazılar.



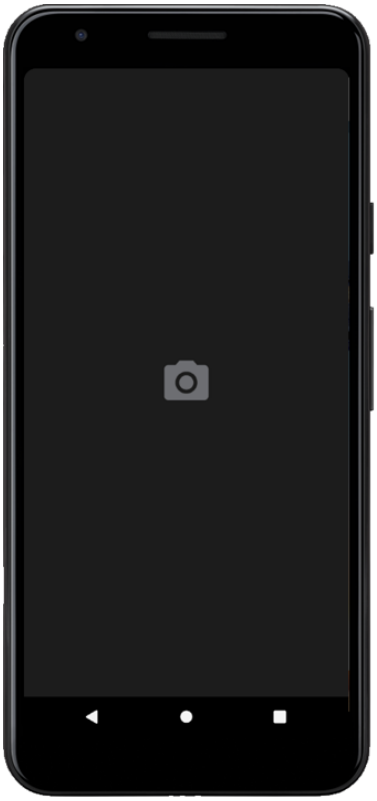
a. Giriş ekranı



b. Bağlantılar çubuğu

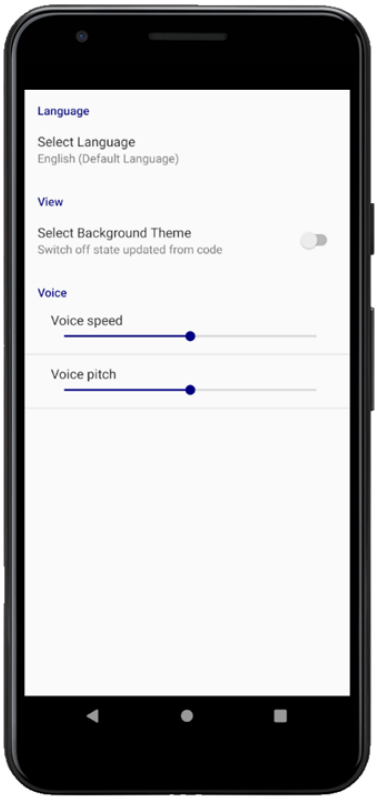


c. Sesli komut ekranı

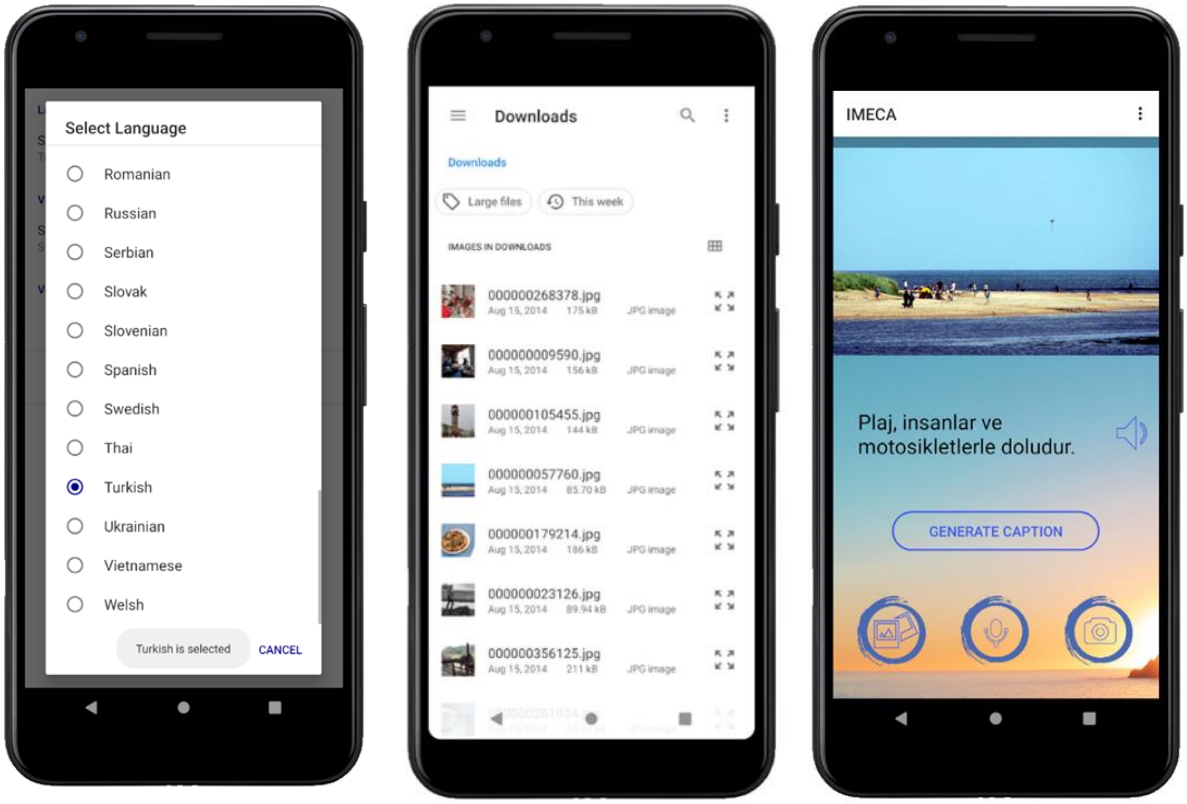


d. Kamera

e. İngilizce dilinde görüntü altyazısı üretilmiş ana ekran



f. Ayarlar ekranı



g. Dil seçenekleri h. Galeri i. Seçilen dilde görüntü altyazısı üretilmiş ana ekran

Şekil 5 *IMECA* uygulaması ekran görüntüleri.

4. VARGILAR

Bu çalışmada, ResNet152V2 kodlayıcı ve çok katmanlı GRU tabanlı kod çözücü yapısına dayanan otomatik görüntü altyazılama sistemi geliştirilmiştir ve gerçekleştirilen deneylerde katman sayısının üretilen görüntü tanımlarının başarısına olan etkisi gözlemlenmiştir. Önerilen sistemin akıllı telefonlarda çevrimdışı olarak çalışabilmesi için oluşturulan model, *IMECA* adlı geliştirdiğimiz Android uygulama ile birleştirilmiştir. Bu uygulama ile, görme engelli bireylere bir dizi basit hareket ve sesle kolayca kullanabilecekleri ve bilgilendirilecekleri bir platform sunulmuştur. Ayrıca 44 farklı dil seçeneğiyle uygulamanın daha geniş bir hedef kitlesine ulaşması sağlanmaktadır.

TEŞEKKÜR

Bu çalışma, TÜBİTAK 2209-B Sanayiye Yönelik Lisans Araştırma Projeleri Desteği Programı kapsamında 1139B412000694 numaralı proje tarafından desteklenmektedir.

KAYNAKLAR

Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). *Spice: Semantic propositional image caption evaluation.* Paper presented at the European conference on computer vision.

Bigham, J. P., Jayant, C., Ji, H., Little, G., Miller, A., Miller, R. C., . . . White, S. (2010). *Vizwiz: nearly real-time answers to visual questions.* Paper presented at the Proceedings of the 23nd annual ACM symposium on User interface software and technology.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. J. a. p. a. (2015). Microsoft coco captions: Data collection and evaluation server.

Çaylı, Ö., Makav, B., Kılıç, V., & Onan, A. (2020). *Mobile Application Based Automatic* Caption Generation for Visually Impaired. Paper presented at the International Conference on Intelligent and Fuzzy Systems.

DuLight, B. (2015).

Gong, Y., & Zhang, H. (1994). *An image database system with content capturing and fast image indexing abilities.* Paper presented at the 1994 Proceedings of IEEE international conference on multimedia computing and systems.

Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. J. A. C. S. (2019). A comprehensive survey of deep learning for image captioning. *51*(6), 1-36.

Karpathy, A., & Fei-Fei, L. (2015). *Deep visual-semantic alignments for generating image descriptions.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Kuyu, M., Erdem, A., & Erdem, E. Altsözcük Ögeleri ile Türkçe Görüntü Altyazılama Image Captioning in Turkish with Subword Units.

Lavie, A., & Agarwal, A. (2007). *METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments.* Paper presented at the Proceedings of the second workshop on statistical machine translation.

Leon, V., Mouselinos, S., Koliogeorgi, K., Xydis, S., Soudris, D., & Pekmestzi, K. J. T. (2020). A tensorflow extension framework for optimized generation of hardware cnn inference engines. *8*(1), 6.

Li, X., Yuan, A., Lu, X. J. M. T., & Applications. (2018). Multi-modal gated recurrent units for image description. *77*(22), 29847-29869.

Lin, C.-Y. (2004). *Rouge: A package for automatic evaluation of summaries.* Paper presented at the Text summarization branches out.

Makav, B., & Kılıç, V. (2019a). *A new image captioning approach for visually impaired people.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Makav, B., & Kılıç, V. (2019b). *Smartphone-based image captioning for visually and hearing impaired.* Paper presented at the 2019 11th International Conference on Electrical and Electronics Engineering (ELECO).

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. J. a. p. a. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn).

Marks, S. (2020). TensorFlow-Slim image classification model library.

Mathur, P., Gill, A., Yadav, A., Mishra, A., & Bansode, N. K. (2017). *Camera2Caption: a real-time image caption generator.* Paper presented at the 2017 International Conference on Computational Intelligence in Data Science (ICCIDS).

Mellit, A., Kalogirou, S. A., Hontoria, L., Shaari, S. J. R., & Reviews, S. E. (2009). Artificial intelligence techniques for sizing photovoltaic systems: A review. *13*(2), 406-419.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *Bleu: a method for automatic evaluation of machine translation.* Paper presented at the Proceedings of the 40th annual meeting of the Association for Computational Linguistics.

Rahman, A., Srikumar, V., & Smith, A. D. J. A. e. (2018). Predicting electricity consumption for commercial and residential buildings using deep recurrent neural networks. *212*, 372-385.

Sutskever, I., Vinyals, O., & Le, Q. V. J. a. p. a. (2014). Sequence to sequence learning with neural networks.

Şeker, A., Diri, B., & Balık, H. H. J. G. M. B. D. (2017). Derin öğrenme yöntemleri ve uygulamaları hakkında bir inceleme. *3*(3), 47-64.

Tanti, M., Gatt, A., & Camilleri, K. P. J. N. L. E. (2018). Where to put the image in an image caption generator. *24*(3), 467-489.

Unal, M. E., Citamak, B., Yagcioglu, S., Erdem, A., Erdem, E., Cinbis, N. I., & Cakici, R. (2016). *Tasviret: A benchmark dataset for automatic turkish description generation from images.* Paper presented at the 2016 24th signal processing and communication application conference (SIU).

Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). *Cider: Consensus-based image description evaluation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Wang, H., Wang, H., & Xu, K. J. N. (2020). Evolutionary recurrent neural network for image captioning. *401*, 249-256.

Xu, N., Liu, A.-A., Liu, J., Nie, W., Su, Y. J. J. o. V. C., & Representation, I. (2019). Scene graph captioner: Image captioning based on structural visual representation. *58*, 477-485.

You, Q., Jin, H., & Luo, J. J. a. p. a. (2018). Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. J. T. o. t. A. f. C. L. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *2*, 67-78.

Qiuyuan, H., Pengchuan, Z., Oliver, W., & Lei, Z. (2018). Turbo Learning for CaptionBot and DrawingBot.