MAUN Fen Bil. Dergi.,13, 1, 128-137 Araştırma Makalesi/ Research Article
MAUN J. of Sci., 13, 1, 128-137 DOI : 10.18586/msufbd.1532112

MR Altyazılama için Çoklu Dikkat Tabanlı Derin Öğrenme Modeli

Burkay MARAŞ[1], Serhat KARATORAK[1], Kevser ÖZDEM KARACA[1]✉, A. Orkun GEDİK[2], M. Ali AKCAYOL[1]
[1] Bilgisayar Mühendisliği, Mühendislik Fakültesi, Gazi Üniversitesi, Ankara, Türkiye
[2] KoçSistem A.Ş., İstanbul, Türkiye

✉: kevserozdem@gazi.edu.tr [1]0009-0004-2967-0486 [2]0009-0001-3586-7056 [3]0000-0002-6695-200X
[4]0009-0004-2623-8073 [5] 0000-0002-6615-1237

Geliş (Received): 14.08.2024 Düzeltme (Revision): 21.09.2024 Kabul (Accepted): 09.10.2024

ÖZ

Son yıllarda sağlık alanında yapay zeka teknolojilerinin kullanımı hızla artmaya başlamıştır. Manyetik rezonans (MR) raporlarının manuel olarak tıp hekimleri tarafından oluşturulması oldukça zor, uzun zaman alan ve hatalı olma olasılığı yüksek bir süreçtir. Bu çalışmada, bu problemleri adreslemek amacıyla beyin MR görüntülerinden otomatik rapor üretecek derin öğrenme tabanlı görüntü altyazılama modeli önerilmiştir. Geliştirilen modelde, görüntü işleme, doğal dil işleme ve derin öğrenme yöntemleri birlikte kullanılarak tıbbi görüntüdeki içerik ve tanılara yönelik metin üretilmektedir. Öncelikle MR görüntüleri için, rastgele açılarla döndürme, boyut değiştirme, kırpma, parlaklık ve kontrast değiştirme, gölge ekleme ve aynalama gibi önişlemler yapılmıştır. Ardından Bootstrapping Language Image Pre-Training (BLIP) modeli ve modelin transformer mimarisinden faydalanılarak rapor üreten bir model geliştirilmiştir. Yapılan deneysel çalışmalarda, geliştirilen modelin farklı metrikler için başarılı sonuçlar verdiği, üretilen raporların orijinal raporlara yüksek oranda benzer olduğu ve tıp alanında yardımcı öneri sistemi olarak kullanılabileceği görülmüştür.

Anahtar Kelimeler: MR, BLIP, Derin Öğrenme, Görüntü Altyazılama, Transformer.

Multiple Attention-Based Deep Learning Model for MRI Captioning

ABSTRACT

In recent years, the use of artificial intelligence in medicine has begun to increase considerably. Creating magnetic resonance (MR) reports manually by medical doctors is very difficult, time-consuming, and error-prone process. This study proposes a deep learning-based image captioning model to automatically generate reports. In the developed model, image processing, natural language processing, and deep learning methods are used to produce text for the medical image. First of all, pre-processing, such as rotating at random angles, changing size, cropping, changing brightness and contrast, adding shadows, and mirroring, were performed for MR images. Then, a model was developed by utilizing the Bootstrapping Language Image Pre-Training (BLIP) model and the transformer architecture of the model to generate a report. The experimental studies showed that the proposed model had successful results; the produced reports were highly similar to the original reports and could be used as a tool in medicine.

Keywords: Brain MRI, BLIP, Deep Learning, Medical Image Captioning, Transformer.

GİRİŞ

Tıbbi görüntü altyazılama, medikal görüntülerden derin öğrenme ve doğal dil işleme teknikleri kullanılarak otomatik olarak rapor üretilmesini ifade etmektedir. Bu konuda yapılan çalışmalar ve araştırmalarda, görüntülerden üretilen bu metinlerin anlamlı cümlelerden oluşması ve tıbbi teşhis ve tedavi süreçlerinde faydalı olması amaçlanmaktadır. Beyin MR görüntülerinden üretilen raporlar, nörolojik anormalliklerin otomatik olarak tespit edilmesi ve bu tespitlerin doğal dilde ifade edilmesi sürecine yardımcı olmaktadır. Bu nedenle, MR görüntülerinden rapor üretmeye yönelik çalışmalar son yıllarda artarak devam etmektedir [1].

Paspula ve Saravanan tarafından 2023 yılında yayınlanan çalışmada, tıbbi görüntü altyazılama üzerinde çalışılmıştır [1]. Çalışmada vücudun farklı bölgelerine ait 21 kategoride 7.442 adet tıbbi görüntü içeren Pathology Education Informational Resource (PEIR) veri kümesi kullanılmıştır. Çalışmada bölgelerin tespiti için önceden eğitilmiş olan You Only Live Once v4 (YOLOv4) modeli kullanılarak ince ayar (fine tuning) yapılmıştır. Model, birinci aşamada nesne tespiti ve lokalizasyon, ikinci aşamada ise Long-Short Term Memory (LSTM) birimleri ve dikkat mekanizması ile dil bilgisine uygun cümle üretimi sağlanmıştır. Yapılan değerlendirmeler sonucu modelin Bilingual Evaluation Understudy (BLEU) skoru %81,78, Metric for Evaluation of Translation with Explicit ORdering (METEOR) skoru ise %78,56 olarak elde ettiği görülmüştür.

Wang ve arkadaşlarının yaptığı çalışmada radyoloji görüntülerinden otomatik rapor üretme amaçlanmıştır

[3]. COVID-19 CT Report ve Indiana University chest X-Ray veri kümelerinden faydalanılmıştır. Resnet101 modelinden ve Recurrent Neural Network (RNN) mimarilerinden yararlanılarak üretilen raporlarda, diğer çalışmalardan farklı olarak raporların benzerliğinin ölçülmesi için Sentence Matched Adjusted Semantic Similarity (SMAS) teknolojisi kullanılmıştır. Raporlar vücudun çeşitli bölgeleri için başarılı çıktılar vermiştir. BLEU, METEOR ve ROUGE-L gibi başarı metriklerinde de literatürdeki çalışmalar ile karşılaştırıldığında başarılı sonuçlara ulaşılmıştır.

William ve arkadaşları tarafından 2020 yılında yayınlanan çalışmada ise yine göğüs X-ray görüntülerinden rapor üretmek amacıyla derin öğrenme mimarilerinden yararlanılmıştır [4]. Çalışmada eğitim ve test işlemleri için MIMIC Chest X-ray (MIMIC-CXR) veri kümesi kullanılmıştır. CNN-RNN mimarileri kullanılan çalışmada, özelliklerin çıkarımı için DenseNet121 modelinden faydalanılmıştır. Projede Beam Search algoritmasının da aktif kullanıldığı çalışmada BLEU ve Consensus-based Image Description Evaluation (CIDEr) performans metrikleri aracılığıyla diğer çalışmalarla kapsamlı bir kıyaslama yapılmıştır.

Liu ve arkadaşlarının yaptığı çalışmada, geleneksel rapor üretimi algoritmalarının üzerine eklenen Contrastive Attention (Karşılaştırmalı Dikkat) isimli bir mekanizmadan bahsedilmiştir [5]. Bu mekanizma kısaca giriş görüntüsünü almakta ve referans görüntüsü üzerinden karşılaştırma yapmaktadır. Referans görüntüsü, normal yani sağlıklı bir kişinin X-ray görüntüsünü temsil etmektedir. 2 görüntünün karşılaştırılmasıyla birlikte giriş görüntüsü üzerinden anomali içeren bölge tespit edilmekte ve mevcut rapor bu şekilde güncellenmektedir. Bu mekanizma temelde 3 aşamadan oluşmaktadır. İlk aşamada normal görüntüler eğitim verisi üzerinden ayıklanarak elde edilmektedir. Daha sonra Aggregate Attention (Toplu Dikkat) mekanizmasıyla giriş görüntüsüne yakın olan normal görüntüler belirlenmektedir. Son aşamada ise Differentiate Attention (Dikkat Farklılaştırma) ile 2 görüntü arasındaki benzerlikler çıkartılmakta ve geriye sadece farklı olan kısım kalmaktadır. Elde edilen bu bölge anomali içeren bölgeyi temsil etmektedir. IU-X-ray ve MIMIC-CXR veri kümeleri üzerinde test edilen bu mekanizmayla birlikte BLEU-4 metriği için %14 ve %17 performans artışı elde edilmiştir.

Lovelace ve arkadaşlarının yaptığı çalışmada, mevcut rapor üretimlerine ek olarak transformer mimarisi kullanılarak üretilen raporların önceki bahsedilen yöntemlere göre daha akıcı ve tutarlı olması sağlanmıştır [6]. Model performansının artırılması amacıyla da görüntüdeki bilgilerin diferansiyel olarak elde edilmesi sağlanmıştır.

Encoder-decoder (kodlayıcı-kod çözücü) mimarisiyle üretilen raporlar düzgün bir rapor çıktısı verse de bu mimaride bazı problemler ortaya çıkmaktadır. İlk olarak görüntüdeki anomali içeren bölgeler görüntünün sadece küçük bir bölümünü kapsamaktadır ve encoder-decoder mimarisi bu kısma yeterince dikkat edememektedir. Bu sebeple bu mimari, anomali içeren bölgelerin tespitinde başarısız olabilmektedir. Mimarinin diğer bir problemi ise uzman bilgisini modele dahil edememesidir. Bu nedenle de bu mimari üzerinden üretilen araçlar kliniklerde yaygın olarak kullanılamamaktadır.

Bu problemlerin çözülebilmesi için Yang ve arkadaşlarının çalışmasında ele alınan yaklaşım, bilgiyi modele dahil etmek amacıyla genel ve spesifik olmak üzere 2 ayrı kategoriye ayırmaktadır [7]. Genel bilgi tüm görüntünün geneliyle ilgilenirken spesifik bilgi ise görüntüye benzer farklı bir görüntü üzerinden terminolojideki terimleri kapsamaktadır. Bu iki bilginin de aynı anda kullanılması adına bilgiyle geliştirilmiş çok kafalı dikkat (knowledge-enhanced multi-head attention) mekanizması geliştirilmiştir. Bu mekanizma 2 ayrı bilgiyi birleştirmektedir. Mekanizma IU-Xray ve MIMIC-CXR veri kümeleri üzerinde test edilmiştir. Sonuç olarak, örnek bir çıktıda 3 adet rapor bilgisi elde edilirken bu mekanizma sayesinde 5 farklı bilgi de elde edilmiştir.

Önceki çalışmalarda bahsedilen, mevcut görüntü altyazılama modellerinin üzerine eklenen bilgiyle geliştirilmiş çok kafalı dikkat mekanizması sayesinde daha anlamlı ve uzun medikal çıktılar üretilmiştir. Fakat bu modellerdeki en büyük problem, başka hastalıkların tespiti söz konusu olduğunda dikkat mekanizmasının el ile güncellenmesini gerektirmesidir. Yang ve arkadaşlarının diğer çalışmasında, bu problem aşılmaya çalışılarak bu sürecin otomatik olarak öğrenilip hafızada saklanması sağlanmıştır. Çalışmada temel olarak 2 aşama üzerinde durulmuştur. İlk olarak medikal bilgiyi işleyip hafızada saklayan bilgi güncelleme mekanizmasından (knowledge updating mechanism) bahsedilmiştir. Bu mekanizma sayesinde görüntü üzerindeki anomali içeren bölgelerin bilgisi otomatik olarak eğitim aşamasında güncellenmektedir. Diğer bir aşama ise görüntü, rapor ve hastalık bölgelerinin semantik olarak eşleştirilmesini sağlayan çok modelli hizalama mekanizmasıdır (multi-model alignment mechanism). Bu model, IU-Xray ve MIMIC-CXR veri kümeleri kullanılarak test edilmiştir. Sonuçlar incelendiğinde BLEU, CIDEr gibi metrikler için önceki çalışmalara göre çok daha başarılı değerler elde edilmiştir.

2020 yılında yayınlanan, göğüs X-ray görüntüleri ve raporlarını içeren bir veri kümesi ile yapılan bir çalışmada, diğer çalışmalardan farklı bir yaklaşım kullanılmıştır [13]. Doğrudan rapor üretmek yerine, patolojik tanıyı tespit edip bunu üretilen rapor ile birleştirerek doğruluk oranları artırılmaya çalışılmıştır. Görüntü altyazılama kavramı medikal amaçlı rapor yazılmasında önemli bir gelişme kat etse de mevcut yaklaşımlar, görüntü altyazılama algoritmalarını kullanarak medikal görüntüleri girdi olarak alıp rapor çıktısı üretebilmekte fakat bu raporlarda patolojik bilgilerin doğruluğunu dikkate almamaktadır. Bu sebeple, her ne kadar rapor üretilebilse de bu raporların güvenilirliği şüphelidir. Çalışmada bu problemi çözmek adına Semantic Fusion Network (SFNet) isimli, lezyonlu bölgeyi tespit eden ve tanı raporu üreten bir model öne sürülmüştür. Lezyonlu alanı tespit eden model görüntüden görsel ve patolojik bilgi çıkarımı yapmakta,

tanı raporu üretilmesini sağlayan model ise bu 2 bilgiyi birleştirerek raporların üretilmesini sağlamaktadır. Bu şekilde tanı raporlarının doğruluk oranları artırılmıştır. Deneysel sonuçlarda, doğruluk skoru için Ultrason veri kümesinde %1,2'lik, Open-i X-ray görüntü veri kümesinde ise %2,4'lük bir artış gözlemlenmiştir. Ayrıca lezyonlu bölgenin tespit edildiği sınıflandırma modellerinde doğruluk oranları %80,2-96,9 arasında değişmektedir. Karşılaştırılan modeller arasında Attention, NIC, m-RNN, Co-Attention, Faster RCNN+LSTM, SFNet, Attention, NIC, m-RNN modelleri bulunmaktadır [13].

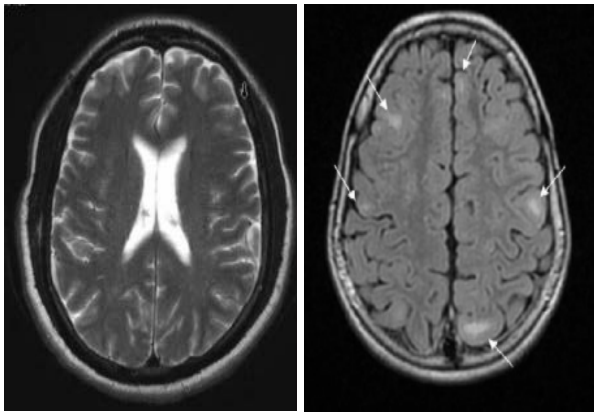
Ankit tarafından 2023 yılında yapılmış çalışmada BLIP modelinden faydalanılmıştır [2]. Veri kümesi olarak Indiana University Chest X-Ray veri kümesi kullanılmıştır. Veriler üzerinde sıkı bir önişleme süreci yürütülen çalışmada karşılaştırmalı ve kapsamlı bir çalışma ile BLIP modelinin diğer modellere göre avantaj ve dezavantajları vurgulanmıştır [2].

Son yıllarda, görsel-dil ön eğitiminde kaydedilen önemli ilerlemeler, görsel bağlama dayalı dil anlama ve üretme yeteneklerini önemli ölçüde geliştirmiştir. Contrastive Language-Image Pre-training (CLIP) ve A Large-scale ImaGe and Noisy-Text embedding (ALIGN) gibi önde gelen modeller, büyük ölçekli veri kümeleri ve karşıt öğrenme tekniklerinden yararlanarak, görüntü-metni alma ve sıfırdan öğrenme gibi görevlerde dikkate değer performanslar elde etmiştir. Ancak, bu modeller, görüntü ve metin gömmelerini hizalamaya odaklanmakta olup, birleşik anlama ve üretme konularına yeterince değinmemektedir. Junnan Li ve arkadaşlarının geliştirdiği BLIP Transformer modeli hem kendi kendine denetimli öğrenmeyi hem de karşıt öğrenmeyi tek bir çerçevede birleştirerek yeni bir yaklaşımı temsil etmektedir. Bu ikili metodoloji, BLIP Transformer'ın her iki öğrenme paradigmasının güçlü yönlerini etkili bir şekilde kullanmasını sağlayarak, çeşitli görsel-dil görevlerinde üstün performans elde etmesine olanak tanımaktadır. Önceki modellerin sıklıkla ayrı ayrı görüntü ve dil bileşenleri için ön eğitim gerektirdiği durumların aksine, BLIP, anlama ve üretme yeteneklerini aynı anda adım adım geliştirmek için bir bootstrapping tekniği kullanmaktadır. Bu kapsamlı ön eğitim stratejisi, model mimarisini basitleştirmenin yanı sıra, görsel soru yanıtlama ve görüntü altyazı oluşturma gibi çeşitli alt görevler için adaptasyonunu da geliştirerek alanda yeni bir standart belirlemektedir [8].

Bu çalışmada, modelin görsel-dil görevlerinde sağladığı üstün performans ve esneklikten dolayı BLIP Transformer tercih edilmiştir. Beyin MR görüntülerinden rapor üretme gibi karmaşık ve yüksek doğruluk gerektiren bir görevde, BLIP Transformer'ın birleşik anlama ve üretme yetenekleri büyük bir avantaj sağlamaktadır. BLIP'in bootstrap (yeniden örnekleme) tekniği, modelin beyin MR görüntülerini doğru bir şekilde anlamasını ve bu görüntülerden anlamlı açıklamalar üretmesini mümkün kılmaktadır. Böylece, tıbbi görüntüleme alanında daha etkili ve güvenilir çözümler geliştirilmesine katkıda bulunmaktadır. Literatürde BLIP modelini tıbbi görüntü altyazılama amacıyla kullanan mevcut çalışmalar bulunmakla birlikte, bu çalışmalar daha çok X-ray görüntüleri üzerinde yoğunlaşmaktadır. Görüntü altyazılama problemine yönelik yeterli miktarda MR görüntüsü bulunduran veri kümelerine erişim zorluğundan dolayı çalışmaların bu şekilde şekillendiği düşünülmektedir. Bu çalışmada, literatürde farklı veri kümeleri üzerinde başarılı bir şekilde uygulanmış BLIP modeli beyin MR görüntülerini altyazılama için kullanılmıştır.

MR ALTYAZILAMA

Beyin manyetik rezonans görüntüleme teknikleri, nörolojik hastalıkların tanısında ve tedavisinde hayati bir rol oynamaktadır. Beyin MR görüntüleri, beyin dokularının detaylı ve yüksek çözünürlüklü görüntülerini sunarak, çeşitli nörolojik durumların ve anormalliklerin tespit edilmesini sağlar. Sağlıklı ve tümörlü örnek MR görüntüleri Şekil 1'de sunulmuştur. Ancak, bu görüntülerin analiz edilmesi ve anlamlı raporlar üretilmesi, uzman radyologlar tarafından yapılan zaman alıcı ve karmaşık bir süreçtir. Radyologlar, her bir görüntüyü dikkatlice inceleyerek, bulguları detaylı bir şekilde raporlamalıdır. Bu süreç, insan hatalarına açık olup radyologların iş yükünü önemli ölçüde artırmaktadır.



Şekil 1. Örnek MR Görüntüleri

Bu sorunlar göz önünde bulundurulduğunda beyin MR görüntülerinden otomatik rapor üretme sistemlerinin geliştirilmesi büyük bir ihtiyaç haline gelmiştir. Otomatik rapor üretme, görüntülerin hızlı ve doğru bir şekilde analiz edilmesini sağlayarak teşhis sürecini hızlandırabilir ve radyologların iş yükünü azaltabilir. Ancak, görüntülerin karmaşıklığı ve çeşitli nörolojik durumların yüksek bir doğruluk ile tanımlanma gerekliliği bu alanda karşılaşılan zorluklardır.

Bu çalışmanın temel motivasyonu, beyin MR görüntülerinden otomatik rapor üretme sistemlerinin sağlık hizmetlerinde yaratabileceği potansiyel faydalara dayanmaktadır. İlk olarak, otomatik rapor üretme sistemleri, radyologların iş yükünü önemli ölçüde azaltarak zaman ve iş gücü tasarrufu sağlar. Radyologlar, her bir MR görüntüsünü manuel olarak analiz etmek yerine otomatik sistemler sayesinde daha fazla hastaya

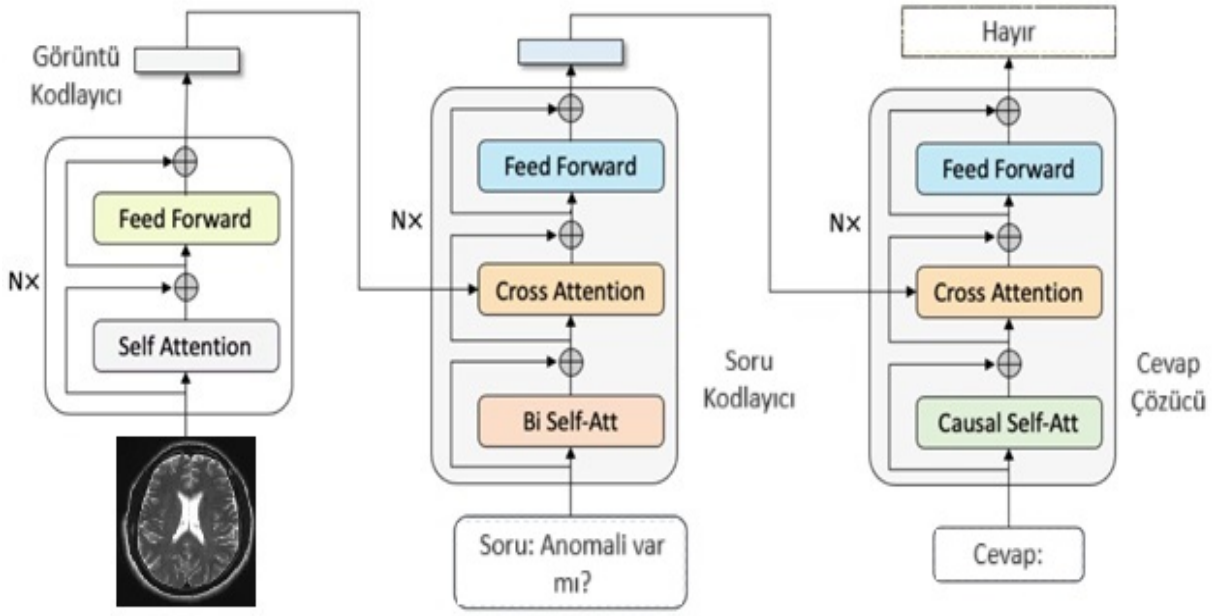
hizmet verebilir. Bu durum, sağlık hizmetlerinin daha verimli bir şekilde sunulmasına katkıda bulunur ve hastaların daha hızlı teşhis ve tedavi almasını sağlar.
İkinci olarak, bu sistemler insan hatalarını minimize ederek tanı süreçlerinin doğruluğunu artırır. Manuel raporlama süreçleri, radyologların dikkat ve konsantrasyon seviyesine bağlı olarak değişkenlik gösterebilmektedir. Buna bağlı olarak bu süreçte hata yapma olasılığı da yüksektir. Otomatik sistemler, algoritmaların tutarlılığı sayesinde standart ve doğru raporlar üreterek yanlış teşhislerin önüne geçebilir. Bu, hastaların doğru tedaviye yönlendirilmesi açısından kritik bir öneme sahiptir.
Üçüncü olarak, otomatik rapor üretme sistemleri klinik karar destek araçları olarak büyük bir potansiyele sahiptir. Tıp hekimleri, bu sistemler sayesinde karmaşık ve çok sayıda görüntüyü hızlı bir şekilde analiz edebilir ve daha bilinçli kararlar alabilir. Otomatik olarak üretilen raporlar, tıp hekimlerinin teşhis sürecinde ek bir bilgi kaynağı olarak kullanabileceği değerli bir katkı sağlar.
Son olarak, bu tür sistemler tıbbi araştırma ve eğitim alanlarında da önemli avantajlar sunar. Araştırmacılar, geniş veri kümelerini hızlı bir şekilde analiz ederek araştırma süreçlerini hızlandırabilir. Aynı zamanda, tıbbi eğitimde otomatik rapor üretme sistemleri, öğrencilere pratik yapma ve gerçek veriler üzerinden öğrenme imkanı sunarak eğitim materyallerini zenginleştirir.
Bu nedenlerle, beyin MR görüntülerinden otomatik rapor üretme çalışmaları, sağlık hizmetlerinin kalitesini ve verimliliğini artırma potansiyeline sahip olup hem klinik hem de akademik alanlarda önemli katkılar sunmaktadır.

VERİ KÜMESİ

Bu çalışmada, tıbbi görüntülerin otomatik olarak açıklamalara dönüştürülmesi amacıyla Radiology Objects in Context (ROCO) veri kümesi kullanılmıştır [9]. ROCO, tıbbi görüntülerin ve ilgili açıklamaların geniş bir koleksiyonunu içeren kapsamlı bir veri kümesidir. Veri kümesinde bulunan açıklamalar İngilizce dilindedir.
Çalışmanın odak noktası olan beyin MR görüntülerini elde etmek amacıyla, ROCO veri kümesi içerisinden spesifik olarak beyin MR görüntüleri filtrelenmiş ve kullanılmıştır. Filtreleme işlemi sonucunda toplamda 642 beyin MR görüntüsü elde edilmiştir. Bu görüntüler, çeşitli nörolojik durumları temsil etmekte olup çalışmada kullanılan modellerin eğitim ve değerlendirme süreçlerinde temel veri kaynağı olarak hizmet etmiştir.
Veri kümesindeki her bir beyin MR görüntüsü, ilgili İngilizce açıklamalarla birlikte gelmekte olup, bu açıklamalar doğal dil işleme modelleri için hedef modeller olarak kullanılmıştır. Bu sayede hem görüntü işleme hem de metin oluşturma modelleri, aynı veri kümesi üzerinden ortak bir çalışma alanı bulmuştur.

MR ALTYAZILAMA İÇİN ÇOKLU DİKKAT TABANLI DERİN ÖĞRENME MODELİ

Bu çalışmada, beyin MR görüntülerinden otomatik rapor üretmek amacıyla BLIP modeli kullanılmıştır. BLIP modeli, görüntülerin içeriğini anlamlı metinlere dönüştürmek için derin öğrenme ve doğal dil işleme tekniklerini bir araya getiren bir yaklaşımdır. Modelin temel bileşenleri, görüntü işleme ve metin oluşturma modüllerini içermektedir.
BLIP modeli, iki aşamalı bir ön eğitim sürecinden geçer. İlk aşamada, görüntü-dil temsil öğrenimi, dondurulmuş bir görüntü kodlayıcı (image encoder) kullanılarak gerçekleştirilir. Bu süreçte, görüntülerden çıkarılan özellikler, metinle uyumlu hale getirilir. Bu uyum, görüntülerin içeriği ile ilgili anlamlı metinler üretmeyi amaçlar. İkinci aşamada ise, dil üretme yeteneğine sahip büyük bir dil modeli ile entegrasyon sağlanarak görsel verilerden metin oluşturma işlemi gerçekleştirilir. Dil modeli olarak Bidirectional Encoder Representations from Transformers (BERT) [15] kullanılmaktadır. Bu iki aşama, BLIP modelinin, görüntü ve metin verilerini etkili bir şekilde entegre etmesini sağlamaktadır.
BLIP modelinin mimarisi, özellikle Q-Former adı verilen bir bileşen üzerine kuruludur. Q-Former, görüntü kodlayıcıdan gelen görsel özellikleri işleyerek metin oluşturma sürecine hazır hale getirir. Bu bileşen hem görüntü hem de metin kodlama görevlerini yerine getiren iki ayrı transformer alt modülünden oluşur. Q-Former, görsel özellikleri metinle uyumlu hale getirmek için kendiliğinden dikkat (self-attention) katmanlarını kullanır. Bu katmanlar, metin ve görsel özellikler arasında çapraz dikkat (cross-attention) katmanları aracılığıyla etkileşim kurar.
Modelin eğitimi sırasında, çeşitli kayıp fonksiyonları kullanılarak model optimize edilir. Görüntü-Metin Karşıt Öğrenimi (Image-Text Contrastive Learning) gibi yöntemlerle, görüntü ve metin temsilleri arasındaki uyum en üst düzeye çıkarılır. Ayrıca, Görüntüye Dayalı Metin Üretimi (Image-grounded Text Generation) ve Görüntü-Metin Eşleştirme (Image-Text Matching) gibi görevlerle modelin performansı artırılır.
BLIP modeli, büyük veri kümeleri kullanılarak eğitilmiş ve çeşitli görüntü-dil görevlerinde üstün performans göstermiştir. Modelin başarısı, mevcut yöntemlere kıyasla daha az sayıda parametre ile yüksek performans elde etmesinden kaynaklanmaktadır. Örneğin, BLIP-2 modeli, Flamingo80B modeline kıyasla %8,7 daha yüksek performans göstermiş ve 54 kat daha az eğitilebilir parametre kullanmıştır [10].
BLIP modelinin mimarisi Şekil 2 üzerinde özetlenmiştir. Şema, görüntü kodlayıcı (image encoder), soru kodlayıcı (question encoder), ve cevap çözücü (answer decoder) bileşenlerinden oluşmaktadır. Görüntü kodlayıcı, görüntüden özellikler çıkararak diğer bileşenlere aktarmaktadır. Soru kodlayıcı, çıkarılan özellikleri kullanarak ilgili soruyu kodlamakta ve son olarak cevap çözücü ise bu bilgileri kullanarak anlamlı bir cevap üretmektedir.



Şekil 2. BLIP Modelinin Mimari Şeması.

Şekil 2'de, modelin çeşitli dikkat mekanizmaları ve katmanları arasındaki etkileşimler vurgulanmıştır. Bu yapı, modelin hem görüntü hem de metin verilerini etkili bir şekilde işleyebilmesini sağlamaktadır [11].
Oluşturulan model, tahmin sırasında beam search algoritmasını kullanmaktadır. Beam search, olası tüm cümleleri değerlendirip en iyi sonucu bulmayı hedefleyen kapsamlı bir arama algoritmasıdır. Bu algoritma, her adımda en olası birkaç (beam width) cümle parçasını tutarak arama alanını daraltır ve böylece daha verimli bir şekilde en uygun cümleyi oluşturur. Beam search, modelin daha anlamlı ve doğru cümleler üretmesini sağlar. Bunun sebebi, rastgele seçimler yerine en iyi adayları sistematik bir şekilde değerlendirmesidir. Bu durum, özellikle karmaşık ve çok anlamlı görseller için daha yüksek doğruluk ve kalite sunmaktadır [12].

DENEYSEL SONUÇLAR

Veri Önişleme

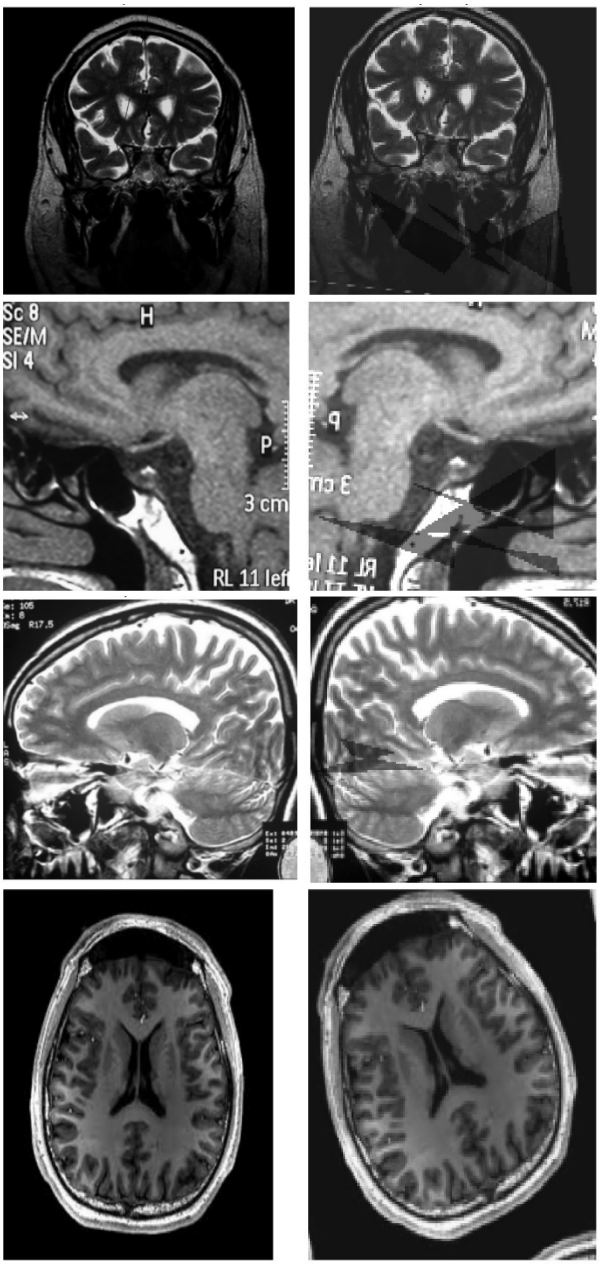
Derin öğrenme modellerinin doğruluk ve verimliliklerini artırmak için büyük miktarda veri kümesine ihtiyaç vardır. Fakat beyin MR görüntüleri ve açıklamalarını içeren büyük miktarda açık kaynaklı eğitim verisi bulunmamaktadır. Bu nedenle veri artırma teknikleri uygulanarak çalışmanın performansı artırılmıştır. Görüntü ve metinlerde ayrı ayrı yöntemler kullanılmıştır. Metin verilerinin artırılmasında *nlpaug* ve *textgenie* kütüphanelerinden; GPT 3.5 ve GPT 4.0 modellerinden faydalanılarak cümle seviyesinde işlem yapılmıştır. Tablo 1'de veri kümesinde bulunan bazı orijinal cümleler ve veri artırma işlemi sırasında bu cümleler kullanılarak oluşturulan yeni cümleler örnek olarak sunulmuştur.

Tablo 1. Metinsel Veri Artırım Örnekleri

Orijinal Metin	Üretilen Metin
MRI of the brain showing no mass or enhancing lesion.	Magnetic Resonance Imaging of the cerebral displaying without lesion or augmenting lesion.
Axial section (T2) showing optic nerve glioma.	Axial slice (T2) displaying optic nerve glioma.
Hemorrhagic infarction of the left temporal lobe	Bleeding stroke of the left temporal lobe
MRI brain showing oblitertion of transverse sinus in coronal section	Magnetic Resonance Imaging cerebral displaying oblitertion of transverse sinus in coronal slice
Brain MRI (sagittal plane) showing normal size of the pituitary.	Brain Magnetic Resonance Imaging (sagittal plane) displaying normal size of the pituitary.
In the left parasagital-frontoparietal convexity of brain MRI, a small meningioma (2.1 × 2.0 × 1.4 cm) was newly detected.	A small meningioma (2.1 × 2.0 × 1.4 cm) was newly detected in the left parasagital-frontoparietal convexity of cerebral MRI.

Görüntü artırma için *opencv* ve *albumentations* kütüphanelerinden faydalanılmıştır. Görüntülerin rastgele açılarla döndürülmesi, farklı boyutlarda yeniden ayarlanması ve belirli bölgelerinin kırpılması gibi işlemler uygulanmıştır. Bu teknikler, görüntülerin çeşitlendirilmesi ve modelin farklı açılardan gelen verileri öğrenmesi amacıyla kullanılmıştır. Ayrıca, görüntülerin parlaklık ve kontrast seviyelerinin rastgele

değiştirilmesi, bazı görüntülere rastgele gölgeler eklenmesi ve yatay aynalama işlemleri de gerçekleştirilmiştir. Bu yöntemler, görüntülerin daha çeşitli ve gerçek dünya koşullarına daha yakın hale getirilmesine yardımcı olmuştur. Şekil 3'te veri kümesinde bulunan orijinal görüntülere bahsedilen yöntemler uygulanarak elde edilmiş yeni görüntüler sunulmuştur. Soldaki görüntüler orijinal görüntüler, sağdakiler ise oluşturulan yeni görüntülerdir.



Şekil 3. Görsel Veri Artırım Örnekleri

Bu veri artırma teknikleri, sınırlı miktardaki veri kümesini genişleterek modelin genelleme yeteneğini ve performansını artırmayı hedeflemiştir. Bu sayede, beyin MR görüntülerine dayalı derin öğrenme modellerinin doğruluğu ve verimliliği önemli ölçüde iyileştirilmiştir.

Metrikler

Bu çalışmada, beyin MR görüntülerinden otomatik rapor üretme amacıyla geliştirilen modelin performansı çeşitli metrikler yardımıyla değerlendirilmiştir. Bunun için BLEU, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) ve CIDEr metrikleri kullanılmıştır. BLEU skoru tahmin edilen rapor ile orijinal rapor arasında kelime bazlı kıyaslama yaparak Eşitlik 1'de sunulan formül ile hesaplanan bir değerlendirme kriteridir.

$$BLEU\ Score\ (N) = Brevity\ Penalty \times GAP\ Score \quad (1)$$

Burada *BLEU Score (N)*, N-gram tabanlı BLEU skorunu; *Brevity Penalty*, kısalık cezasını; Geometric Average Precision *(GAP) Score* ise geometrik ortalama hassasiyeti ifade etmektedir.
Kısalık cezası, üretilen cümlenin referans cümleden daha kısa olması durumunda uygulanmaktadır. GAP, n-gram doğruluklarının geometrik ortalamasıdır. GAP hesaplaması Eşitlik 2'de verilmiştir.

$$GAP\ (N) = \exp\left(\sum_{n=1}^{N} w_n log p_n\right) \quad (2)$$

Bu işlem aynı zamanda Eşitlik 3'teki şekilde de ifade edilebilmektedir.

$$GAP\ (N) = \prod_{n=1}^{N} p_n^{w_n} \quad (3)$$

Burada N, kullanılan n-gramların sayısını; w_n, n-gramların ağırlıklarını; p_n ise n-gram doğruluklarını ifade etmektedir.
ROUGE, otomatik olarak oluşturulan metinleri değerlendirmek için kullanılan bir ölçüttür. Genellikle bir metin özetlemesi algoritması tarafından oluşturulan özetin kalitesini değerlendirmek için kullanılmaktadır. Bunun için, makine tarafından oluşturulan bir özeti, insan tarafından oluşturulan bir özetle karşılaştırmaktadır. Bu çalışmada da geliştirilen modelin oluşturduğu rapor ile gerçek raporu karşılaştırmak amacıyla kullanılmıştır.
ROUGE ölçümleri, farklı ayrıntı düzeylerine dayalı olarak çeşitli yöntemlerle hesaplanabilir. Literatürde en yaygın kullanılan ve bu çalışmaya dahil edilenler şu şekildedir:

- ROUGE-N, model tarafından oluşturulan metin ile referans metin arasındaki N-gramların (tek gram, çift gram, üç gram vb.) örtüşmesini ifade etmektedir. Literatürde yaygın olarak ROUGE-1 ve ROUGE-2 kullanılmaktadır.
- ROUGE-L, En Uzun Ortak Alt Dizi (Longest Common Subsequence - LCS) algoritması tarafından hesaplanan en uzun ortak kelime dizisini ölçmektedir [14].

CIDEr skoru, genellikle bir referans (doğru veya insan yapımı açıklamalar) kümesi ve modelin ürettiği

açıklamalar arasındaki benzerlikleri hesaplamak için kullanılır. Matematiksel olarak nasıl hesaplandığı Eşitlik 4'te sunulmuştur.

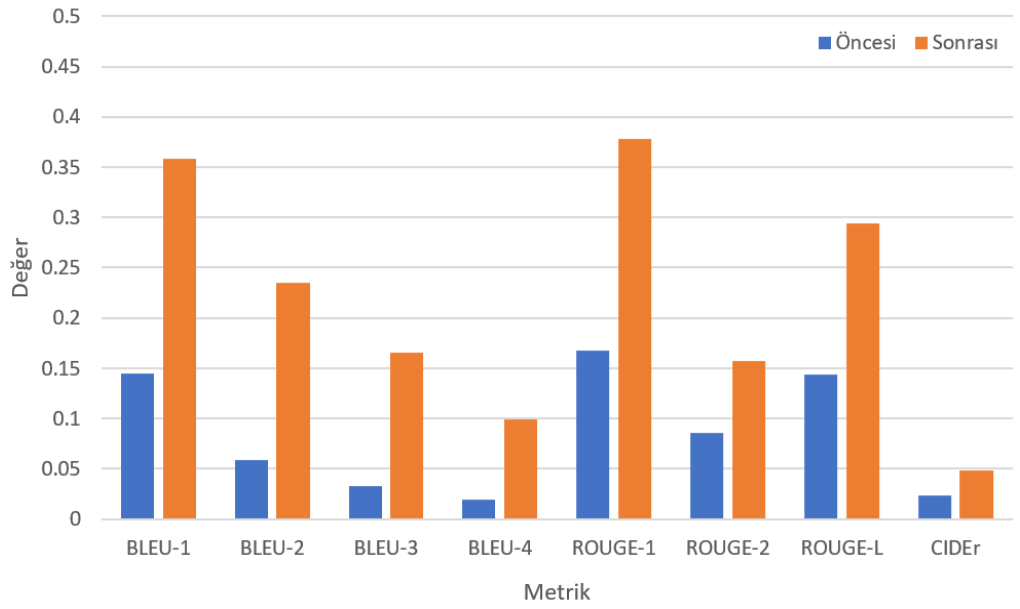
$$CIDEr = \frac{1}{m}\sum_{i=1}^{m}\frac{\min(rl, cl)}{\max(rl, cl)} \times \sum_{n=1}^{N}\frac{count_c}{count_r} \times TF\_c \qquad (4)$$

Burada m, değerlendirilen referans cümle sayısını; rl, referans cümlenin uzunluğunu; cl, üretilen cümlenin uzunluğunu; $count_c$ üretilen cümledeki n-gram sayısını; $count_r$ referans cümledeki n-gram sayısını; TF_c ise üretilen cümledeki n-gramların Term Frequency (TF) değerini ifade etmektedir.

Örneğin, BLEU skoru (Eşitlik 1) kullanılarak yapılan değerlendirmelerde, modelin referans açıklamalarına olan benzerliği ölçülmüştür. GAP (Eşitlik 2) kullanılarak hesaplanan BLEU skoru, modelin doğruluğunu yansıtan önemli bir metriktir. Dört n-gram kullanılarak yapılan bir hesaplama sonucunda, GAP skorunun (Eşitlik 4) 0,5 olduğu varsayılırsa, bu durumda BLEU skoru, Brevity Penalty ile çarpılarak nihai skoru verir.
CIDEr skoru ise (Eşitlik 5), referans açıklamalar ve model çıktıları arasındaki benzerlikleri detaylı bir şekilde değerlendirir. Bu metrik, referans uzunlukları ve n-gram sayımları gibi faktörleri dikkate alarak daha hassas bir değerlendirme sağlar. Örneğin, CIDEr skorunun 0,8 olduğu bir senaryoda, modelin referans açıklamalarla yüksek bir uyum içinde olduğu söylenebilir.
Bu metriklerin kullanımı, model performansının kapsamlı bir şekilde değerlendirilmesine olanak tanımakta ve sonuçların daha güvenilir bir şekilde yorumlanmasını sağlamaktadır.
Veri artırımı uygulanmadan önce ve uygulandıktan sonra elde edilen BLEU, ROUGE ve CIDEr skorları Tablo 2'de karşılaştırılmış ve görsel olarak Şekil 4'te sunulmuştur.

Tablo 2. Veri Artırımı İşlemi Öncesi ve Sonrası Sonuçlar

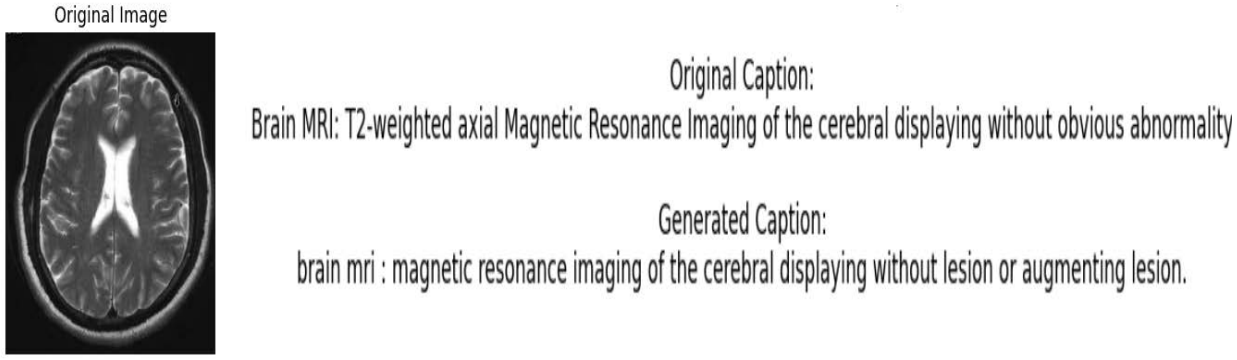
Metrik	Öncesi	Sonrası
BLEU-1	0,1444	0,3584
BLEU-2	0,0582	0,2347
BLEU-3	0,0324	0,1659
BLEU-4	0,0198	0,0990
ROUGE-1	0,1678	0,3779
ROUGE-2	0,0853	0,1576
ROUGE-L	0,1432	0,2943
CIDEr	0,0235	0,0487



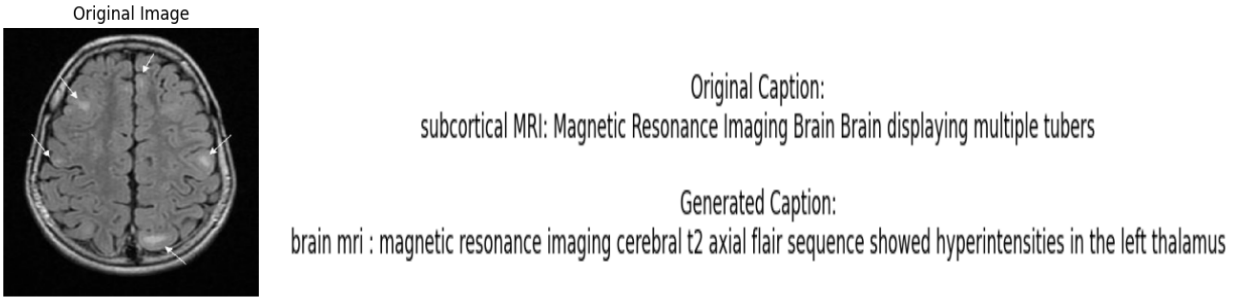
Şekil 4. Veri Artırımı İşlemi Öncesi ve Sonrası Sonuçlar

Bu iyileşmeler, modelin beyin MR görüntülerinden anlamlı ve doğru açıklamalar üretme yeteneğini büyük ölçüde artırmıştır.
Şekil 5'te T2 ağırlıklı bir beyin MR görüntüsü üzerinden elde edilen örnek bir çıktı ve model tarafından üretilen açıklama verilmiştir. Bu örnek, modelin beyin MR görüntülerini analiz ederek tam olarak doğru olmasa da anlamlı açıklamalar üretebildiğini göstermektedir.
Şekil 6'da ise Fluid Attenuated Inversion Recovery (FLAIR) ağırlıklı bir beyin MR görüntüsünden elde edilen çıktı gösterilmektedir. Orijinal açıklamada görüntünün FLAIR olduğu belirtilmese de modelin ürettiği çıktıda bunun tespit edildiği görülmektedir.
Şekil 7'de sella turcica olarak adlandırılan bölgenin MR görüntüsü ve çıktıları gösterilmektedir. Orijinal çıktıyla

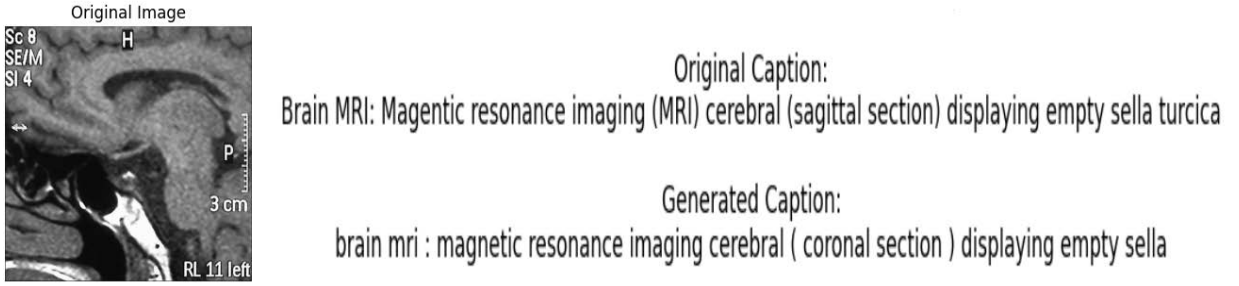
karşılaştırıldığında oldukça başarılı sonuç elde edilen çıktılardan biri olduğu görülmektedir.



Şekil 5. T2 ağırlıklı lezyonsuz bir beyin MR görüntüsü için bulgular: T2 ağırlıklı aksiyel manyetik rezonans görüntülemede beyinde belirgin bir anormallik görülmemektedir.



Şekil 6. FLAIR ağırlıklı bir beyin MR görüntüsü için bulgular: Serebral T2 aksiyel FLAIR dizisi sol talamusta hiperintensiteler göstermektedir. Beyinin subkortikal bölümünde çok sayıda tüberkül görülmektedir.



Şekil 7. Sella Turcica boşluğunu gösteren beyin MR görüntüsü için bulgular: Beyin manyetik rezonans görüntüsü serebral (koronal kesit) boş sella turcica'yı göstermektedir.

Veri artırımı ve önişleme tekniklerinin uygulanması, modelin performansını belirgin bir şekilde artırmıştır. BLEU ve ROUGE skorlarındaki iyileşmeler, modelin daha doğru ve anlamlı metinler üretebildiğini göstermektedir. CIDEr skorundaki artış ise, modelin metin üretiminde daha tutarlı olduğunu ve insan açıklamalarına daha yakın çıktılar ürettiğini göstermektedir.

Bu sonuçlar, beyin MR görüntülerinden otomatik rapor üretme konusundaki çalışmaların potansiyelini ortaya koymakta ve gelecekteki çalışmalar için önemli bir referans oluşturmaktadır.

TARTIŞMA

Performans Değerlendirmesi ve Sınırlamalar

Çalışmada elde edilen sonuçlar, ROCO veri kümesi kullanan literatürdeki diğer görüntü altyazılama çalışmaları ile farklı metrikler açısından kıyaslanmıştır. Tablo 3'te ImageCLEF görüntü altyazılama yarışmasına ait takımların farklı metrikler için elde ettikleri sonuçlar

ile aynı metrikler için bu çalışmada elde edilen sonuçlar verilmiştir [16]. Tablo 4'te ise literatürde bulunan farklı altyazılama çalışmalarına ait sonuçlar paylaşılmıştır.

Tablo 4. Literatürde farklı veri kümelerindeki altyazılamalara ait sonuçlar

Veri Kümesi	Model	ROUGE	BLEU-1	BLUE-2	BLEU-3	BLEU-4
MIMIC-CXR	**Noise-RNN**	0.272	0.269	0.172	0.113	0.074
	1-NN	0.244	0.305	0.171	0.098	0.057
	S&T	0.286	0.318	0.205	0.137	0.093
	SA&T	0.288	0.318	0.205	0.137	0.093
	TieNet	0.307	0.352	0.223	0.153	0.104
Open-I	**Noise-RNN**	0.291	0.233	0.130	0.087	0.061
	1-NN	0.201	0.232	0.132	0.082	0.018
	S&T	0.306	0.328	0.212	0.157	0.108
	SA&T	0.313	0.328	0.212	0.157	0.108
	TieNet	0.330	0.359	0.246	0.171	0.113
ImageCLEF	**BLIP**	**0.294**	**0.358**	**0.234**	**0.165**	**0.099**

Tablo 3. ImageCLEFmedical Task 2024 sonuçları

Takım	ROUGE	BLEU-1	CIDEr
pclmed	0.2726	0.2689	0.2681
CS_Morgaan	0.2508	0.2092	0.2450
DarkCow	0.2452	0.1950	0.2242
auebnlpgroup	0.2048	0.1110	0.1769
2Q2T	0.2477	0.2212	0.2200
MICLab	0.2135	0.1852	0.1582
Geliştirilen model	**0.2943**	**0.3584**	**0.0487**

Modelin performansı BLEU ve CIDEr gibi metriklerle değerlendirildiğinde anlamlı ama iyileştirilmeye açık sonuçlar elde edilmiştir. Örneğin, BLEU-4 skorunun düşük kalması modelin karmaşık cümle yapılarında zorlandığını göstermektedir. Bu durum özellikle karmaşık nörolojik anomalilerin tanımlanmasında sınırlamalar oluşturabilmektedir. Ayrıca kullanılan veri kümesinin büyüklüğü ve dengesi de modelin genelleme kabiliyetini etkileyen önemli faktörler arasındadır. Daha dengeli ve kapsamlı veri kümeleri kullanılması durumunda modelin doğruluk oranlarının artması beklenmektedir.

Diğer modellerle kıyaslandığında, S&T ve SA&T modellerinin daha dengeli bir performans sergilediği görülmüştür. Bu modeller BLEU ile ROUGE metriklerinde kısmen daha istikrarlı sonuçlar elde etmiştir. Bununla birlikte, BLEU-4 skorlarının TieNet'e kıyasla daha düşük olması, bu modellerin daha kısa ve basit cümlelerde başarılı olduğunu ancak karmaşık yapıları oluşturmakta zorlandığını göstermektedir. Öte yandan, BLIP modeli özellikle BLEU-1 ve ROUGE metriklerinde yüksek skorlar elde etmiş olup, bu durum modelin daha geniş veri kümelerinde başarılı olabileceğini düşündürmektedir. Genel olarak, daha ileri seviye doğal dil işleme teknikleri içeren modeller, daha başarılı metin oluşturma kapasitesine sahipken, bazı klasik yöntemler belirli veri kümelerinde daha tutarlı sonuçlar verebilmektedir [17].

Görsel ve Metinsel Bilginin Entegrasyonu

BLIP modeli, görüntü ve dil entegrasyonunda güçlü bir yapı sergilemiştir. Ancak bu sürecin daha ileri seviyede optimize edilmesi mümkündür. Görüntülerdeki bağlamsal bilgilerin daha derinlemesine modellenmesi, üretilen açıklamaların zenginliğini artırabilir. Örneğin, görüntüler arasındaki ilişkileri daha iyi anlamak için çok katmanlı semantik bağlam analizleri uygulanarak modelin başarısı olumlu yönde geliştirilebilir.

Gelecekteki Geliştirme Olanakları

Bu çalışma, BLIP modeli tabanlı otomatik raporlama sistemlerinin gelecekteki uygulamaları için iyi bir temel oluşturmaktadır. Görsel-dil işleme yeteneklerinin çok modelli öğrenme teknikleriyle geliştirilmesi, daha büyük ve dengeli veri kümelerinin kullanılması ve klinik ortamlardan alınan geri bildirimlerin modelin rafine edilmesine yönelik olarak entegrasyonu, çalışmanın etki alanını genişletebilir. Bu iyileştirmeler, üretilen raporların nörolojik anomali tespitinde daha hassas ve klinik olarak daha uygulanabilir hale gelmesini sağlayacaktır.

Klinik Uygulamalara Etkisi

Modelin sağlık hizmetlerinde klinik karar destek sistemi olarak potansiyeli büyüktür. Bu tür sistemler, özellikle zaman tasarrufu ve teşhis süreçlerinin standartlaştırılması açısından önemli avantajlar sunabilir. Ancak, gerçek klinik ortamda uygulanabilirlik için model çıktılarının insan doğrulaması ile desteklenmesi gerekmektedir.

SONUÇ

Bu çalışmada, beyin MR görüntülerinden otomatik rapor üretme sürecini geliştirmek amacıyla BLIP modelinden yararlanılmıştır. Modelin eğitimi için beyin MR görüntüleri ile filtrelenmiş ROCO veri kümesi kullanılmıştır. BLEU, ROUGE ve CIDEr metrikleri kullanılarak model performansı değerlendirilmiş, önişleme ve veri artırım yöntemleri sayesinde bu skorlarda belirgin iyileşmeler gözlemlenmiştir.
BLEU, ROUGE ve CIDEr skorları, model tarafından üretilen açıklamaların veri kümesinde bulanan referans açıklamalara olan benzerliğini ölçmek için kullanılmıştır. Deneysel sonuçlarda, veri artırımı ve önişleme tekniklerinin uygulanmasıyla iyileşmeler sağlanmıştır. Veri kümesinin hem görüntü kısmında hem de metin kısmında veri artırımı uygulandıktan sonra elde edilen sonuçlar BLEU-1 için 0,3584, BLEU-2 için 0,2347, BLEU-3 için 0,1659, BLEU-4 için 0,0990 şeklinde olmuştur. Diğer skorlardaki başarı ise ROUGE-1, ROUGE-2, ROUGE-L ve CIDEr için sırasıyla 0,3779, 0,1576, 0,2943 ve 0,0487 olarak elde edilmiştir.
Çalışma sürecinde karşılaşılan en büyük zorluklardan biri sağlıklı sonuçlar alınabilecek düzenli bir veri kümesi elde etme aşamasında yaşanmıştır. Görüntü altyazılamaya yönelik yeterli sayıda MR görüntüsü içeren ve her bir görüntü için yeterli boyutta ve anlamlı açıklamalar bulunduran veri kümelerine ihtiyaç vardır. Bu ihtiyaçtan kaynaklanan kısıt, veri artırma teknikleri ile azaltılmaya çalışılmıştır. Çalışmada kullanılan veri kümesi ile oluşturulan model ve elde edilen sonuçlar, literatürde ağırlıklı olarak X-ray görüntülerine odaklanan çalışmalara çeşitlilik katmıştır. Beyin MR görüntülerinden otomatik rapor üretme çalışmaları, sağlık hizmetlerinin kalitesini ve verimliliğini artırma potansiyeline sahiptir. Elde edilen bu sonuçların daha büyük ve düzenli bir veri kümesi ile daha da iyileştirilebileceği düşünülmektedir.

TEŞEKKÜR

TÜBİTAK ve KoçSistem Bilgi ve İletişim Hizmetleri A.Ş.'ne 2209B Üniversite Öğrencileri Sanayiye Yönelik Araştırma Projeleri Desteği Programı kapsamında 1139B412304472 proje kodu ile yer alan bu çalışmaya sağladıkları destekten dolayı teşekkür ederiz.

KAYNAKÇA

[1] Ravinder, P., & Srinivasan, S. (2024). Automated Medical Image Captioning with Soft Attention-Based LSTM Model Utilizing YOLOv4 Algorithm.

[2] Aggarwal, A. K. Revealing AI-Driven Chest X-Ray Image Captioning Using Blip Transformer.

[3] Wang, Y., Lin, Z., Xu, Z., Dong, H., Luo, J., Tian, J., ... & He, Z. (2024). Trust it or not: Confidence-guided automatic radiology report generation. Neurocomputing, 127374.

[4] Boag, W., Hsu, T. M. H., McDermott, M., Berner, G., Alesentzer, E., & Szolovits, P. (2020, April). Baselines for chest x-ray report generation. In Machine learning for health workshop (pp. 126-140). PMLR.

[5] Liu, F., Yin, C., Wu, X., Ge, S., Zou, Y., Zhang, P., Zou, Y., & Sun, X. (2023). Contrastive attention for automatic chest X-ray report generation. arXiv. https://arxiv.org/abs/2106.06965.

[6] Lovelace, J., & Mortazavi, B. (2020, November). Learning to generate clinically coherent chest X-ray reports. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1235-1243).

[7] Yang, S., Wu, X., Ge, S., Zhou, S. K., & Xiao, L. (2022). Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical image analysis, 80, 102510

[8] Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., & Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. Medical Image Analysis, 86, 102798.

[9] Pelka, O., Koitka, S., Rückert, J., Nensa, F., & Friedrich, C. M. (2018). Radiology objects in context (roco): a multimodal image dataset. In Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3 (pp. 180-189). Springer International Publishing.

[10] Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." Proceedings of the 40th International Conference on Machine Learning, PMLR 202:19730-19742, 2023.

[11] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." Proceedings of the 39th International Conference on Machine Learning, PMLR 162:12888-12900, 2022.

[12] Lemons, S., Linares López, C., Holte, R., & Ruml, W. (2022). Beam search: Faster and monotonic. Proceedings of the International Conference on Automated Planning and Scheduling, 32(1), 222-230. https://doi.org/10.1609/icaps.v32i1.19805.

[13] Zeng, X., Wen, L., Xu, Y., & Ji, C. (2020). Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models. Computer methods and programs in biomedicine, 197, 105700.

[14] Barbella, M., & Tortora, G. (2022). Rouge metric evaluation for text summarization techniques. Available at SSRN 4120317.

[15] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[16] https://www.imageclef.org/2024/medical/caption

[17] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., & Ghassemi, M. (2019, October). Clinically accurate chest x-ray report generation. In Machine Learning for Healthcare Conference (pp. 249-269). PMLR.