Hazırlayan: Murat Demirci

Tarih: 23 Ocak 2025

Büyük Dil Modelleri (BDM), doğal dil işleme alanında önemli bir rol oynamaktadır. Bu modellerin performansını değerlendirmek ve karşılaştırmak için çeşitli kıyaslama veri setleri kullanılmaktadır. Bu rapor, BDM kıyaslamasında kullanılan bazı önemli açık kaynak veri setlerini ve bu alandaki literatür çalışmalarını özetlemektedir.

Açık Kaynak Veri Setleri

- **MMLU (Massive Multitask Language Understanding):** Bu veri seti, dil modellerinin çok çeşitli konulardaki bilgi ve problem çözme becerilerini ölçmek için tasarlanmıştır. STEM, sosyal bilimler, beşeri bilimler gibi 57 farklı konuda sorular içermektedir. Sıfır-shot ve few-shot öğrenme senaryolarında modelleri değerlendirmek için kullanılır.
- **GSM8K (Grade School Math 8K):** Bu veri seti, ilkokul seviyesinde matematik problemi çözme yeteneğini değerlendirmek için kullanılır. 8.500 matematik kelime probleminden oluşmaktadır ve bu problemler 2 ila 8 adımda çözülebilmektedir.
- **Diğer Veri Setleri**
  - **Kaggle Veri Setleri:** Kaggle, veri bilimi yarışmalarının düzenlendiği bir platformdur ve çok çeşitli açık veri setlerine ev sahipliği yapar. Bu veri setleri farklı alanlardaki projeler için kullanılabilir.
  - **Hugging Face Datasets:** Hugging Face, doğal dil işleme (NLP) için çeşitli veri setlerini barındıran bir platformdur. Bu veri setleri, dil modellerinin eğitilmesi ve değerlendirilmesi için kullanılabilir.
  - **Türkiye'nin Açık Veri Platformları:** Türkiye'deki çeşitli kurum ve belediyelerin yayınladığı açık veri platformları ve veri setleri de farklı projelerde kullanılabilir.

Görsel Veri Setleri:

- **ImageNet:** 1000 kategoriye ayrılmış 1,2 milyon resimden oluşan, geniş çapta kullanılan bir veri kümesidir. Nesne tanıma görevleri için kullanılır.
- **COCO (Common Objects in Context):** Nesne tespiti, segmentasyonu ve başlıklandırmaya odaklanan çeşitli resimlerden oluşan bir veri kümesidir.
- **Open Images:** Nesne tespiti ve görsel ilişki tespiti için çeşitli alanlardan geniş bir resim koleksiyonu sunar.
- **PASCAL VOC (Visual Object Classes):** Nesne tespiti, sınıflandırması ve segmentasyon görevleri için kullanılan popüler bir veri kümesidir.
- **Cityscapes:** Kentsel ortamlarda anlamsal ve sahne anlayışı için tasarlanmış, piksel düzeyinde açıklamalar içeren yüksek çözünürlüklü resimlerden oluşur.
- **CIFAR-10 ve CIFAR-100:** 10 ve 100 kategoriye ayrılmış 32x32 renkli resimlerden oluşan veri kümeleridir.
- **Oxford-IIIT Pet Images Dataset:** Her sınıfta 200 resim olmak üzere 37 kategoride evcil hayvan resimleri içerir.
- **Google's Open Images:** Nesne tespiti, görsel ilişki tespiti ve daha fazlası için açıklamalarla birlikte çeşitli resimlerden oluşan geniş ölçekli bir veri kümesidir.

- **IMDB-WIKI Dataset:** Yaş, cinsiyet ve isimlerle etiketlenmiş yüzlere sahip popüler bir açık veri tabanıdır.
- **Celeb Faces:** Ünlülerin 200.000 açıklamalı görüntüsünü içeren büyük ölçekli bir veri kümesidir.
- **SA-1B Dataset:** Gelişmiş bilgisayar modellerini eğitmek ve değerlendirmek için uygun 11 milyon farklı ve yüksek çözünürlüklü görüntü ve 1.1 milyar piksel düzeyinde ek açıklamadan oluşur.

Video Veri Setleri:

- **Kinetics 700:** 700 farklı insan eylemi sınıfına ait 650.000'den fazla yüksek kaliteli klip içeren büyük bir veri kümesidir.
- **VoxCeleb2:** Açık kaynaklı medyadan otomatik olarak elde edilen büyük ölçekli bir konuşmacı tanıma veri kümesidir.
- **FaceForensics++:** Derin sahtekarlıklar, yüz değiştirme ve daha fazlasını içeren yüz manipülasyonu yöntemleriyle değiştirilmiş 1000 orijinal video dizisinden oluşan bir veri kümesidir.
- **IJB-C:** Yaklaşık 138.000 yüz görüntüsü, 11.000 yüz videosu ve 10.000 yüz dışı görüntü içeren video tabanlı bir yüz tanıma veri kümesidir.
- **100 Days Of Hands Dataset (100DOH):** 11 kategoriden 27,3K Youtube videosu içeren büyük ölçekli bir video veri kümesidir.
- **TrackingNet:** Vahşi doğada çekilen videoları içeren, ortalama 470,9 kare ile 30.643 videodan oluşan geniş ölçekli bir takip veri kümesidir.
- **Youtube-VOS:** Eğitim, doğrulama ve test için piksel düzeyinde temel gerçek açıklamalar içeren 4.453 video içeren bir video nesne segmentasyon veri kümesidir.
- **MSR-VTT (Microsoft Research Video to Text):** Açık alan video başlığı için büyük ölçekli bir veri kümesidir.
- **InternVid:** Video ve metin anlayışının sınırlarını zorlamak için titizlikle hazırlanmış, etkileyici bir 7 milyon video ölçeğine sahip çok modlu üretken yapay zeka veri kümelerinde önemli bir kaynaktır.
- **BDD100K:** Otonom sürüş, bilgisayarlı görü ve robotik araştırmalarını ilerletmek için değerli bir varlık olan 100.000'den fazla video içeren büyük ölçekli çeşitli bir sürüş videosu veri kümesidir.

Ses Veri Setleri:

- **LibriSpeech:** 1000 saatten fazla İngilizce konuşma içeren, sesli kitaplardan elde edilmiş popüler bir veri setidir.
- **Mozilla Common Voice:** Dünyanın dört bir yanından binlerce kişinin katkılarıyla oluşturulmuş, çok dilli ve büyük ölçekli bir veri setidir.
- **TED-LIUM:** Çeşitli konuşma konuları, aksanlar ve kayıt kaliteleri sunan, TED konuşmalarından elde edilmiş bir veri setidir.
- **Multilingual LibriSpeech:** İngilizce'nin yanı sıra Almanca, Hollandaca, İspanyolca, Fransızca, İtalyanca, Portekizce ve Lehçe gibi dilleri de içeren, LibriVox sesli kitaplarından elde edilmiş çok dilli bir veri setidir.

- **VoxForge:** Açık kaynaklı konuşma tanıma motorlarında kullanılmak üzere birden fazla dilde konuşma kayıtlarını içeren bir veri setidir.
- **TIMIT:** Fonetik ve dilbilimsel araştırmalar için detaylı fonetik ek açıklamalar sunan temel bir korpustur.
- **Speech Commands:** Anahtar kelime tanıma sistemlerini eğitmek ve değerlendirmek için tasarlanmış bir sesli kelime veri setidir.
- **VoxPopuli:** 23 dilde 100.000 saat etiketlenmemiş konuşma verisi sağlayan büyük ölçekli çok dilli bir korpustur.

Ek Veri Setleri:

- **HMDB51:** Filmler ve web videoları dahil olmak üzere çeşitli kaynaklardan gerçekçi videoların büyük bir koleksiyonudur.
- **UCF101:** Vahşi doğadan alınan 101 insan eylemi sınıfına ait bir veri kümesidir.
- **Sports-1M:** Evrişimli sinir ağları ile büyük ölçekli video sınıflandırması.
- **ActivityNet:** İnsan etkinliği anlama için geniş ölçekli bir video karşılaştırması.
- **MPII-Cooking:** El merkezli özellikler ve komut dosyası verileri kullanılarak ince taneli ve bileşik etkinliklerin tanınması.

Diğer Önemli Veri Setleri:

- **MR-GSM8K:** Büyük dil modellerinin (LLM'ler) meta-akıl yürütme yeteneklerini değerlendirmek için tasarlanmış zorlu bir kıyaslama.

Literatür Çalışmaları

- **BDM'lerin Matematiksel Muhakeme Becerileri:** GSM8K veri seti, BDM'lerin matematiksel muhakeme becerilerini test etmek için sıklıkla kullanılır. Çalışmalar, modellerin bu alanda hala gelişime ihtiyaç duyduğunu göstermektedir.
- **MMLU ile Geniş Kapsamlı Değerlendirme:** MMLU, farklı alanlardaki bilgi ve problem çözme becerilerini değerlendirerek BDM'lerin güçlü ve zayıf yönlerini ortaya çıkarmada önemli bir rol oynamaktadır.
- **Türkçe Dil Modelleri ve Kıyaslama:** Türkçe dil modellerinin performansını değerlendirmek için de çalışmalar yapılmaktadır. İngilizce veri setlerinden örnekleme yapılarak Türkçeye çevrilen ve soru-cevap veri setleri oluşturulmaktadır.
- **Veri Kalitesi ve Etik Hususlar:** BDM'lerin eğitimi için kullanılan verilerin kalitesi ve etik yönleri de önemli bir konudur. Veri setlerindeki tekrarların giderilmesi ve yanlış bilgilerin düzeltilmesi, model performansını artırmak için kritik öneme sahiptir.

Dokümantasyon

- **Veri Seti Kaynakları:** Yukarıda bahsedilen veri setlerine Papers With Code, Hugging Face ve GitHub gibi platformlardan erişilebilir. Ayrıca, Kaggle'da da veri setleri bulunmaktadır.
- **BDM Kıyaslama Araçları:** WhyLabs'in LangKit projesi gibi açık kaynaklı araçlar, BDM'lerin performansını izlemek ve analiz etmek için kullanılabilir.

Sonuç

BDM kıyaslaması için çeşitli açık kaynak veri setleri ve literatür çalışmaları mevcuttur. MMLU ve GSM8K gibi veri setleri, modellerin farklı becerilerini değerlendirmek için yaygın olarak kullanılmaktadır. Bu alandaki çalışmalar, BDM'lerin gelişimine katkıda bulunmakta ve model performansını artırmak için yeni yaklaşımlar sunmaktadır.