ColPali: Efficient Document Retrieval With Vision Language Models

Emir Özaktaş

**E-Mail: ozaktasemir@gmail.com*

Giriş

Belge erişimi, kullanıcı sorularını (*query*) bir veritabanındaki verilerle eşleştirmeyi amaçlayan bir sistemdir. Arama motorları ve çeşitli soru cevaplama uygulamalarında kritik role sahiptir. Mevcut belge erişim sistemleri metin tabanlıdır ve görselleri (grafik vs.) göz ardı eder, bu şekilde bazı önemli ipucu ve detayları kaçırır. Bu makale Görsel Dil Modelleri (*Vision Language Models, VLM's*) ve yeni bir kıyaslama seti (ViDoRe) kullanarak yeni bir model (ColPali) sunmayı amaçlamaktadır.

ViDoRe Benchmark

Görsel olarak zengin belgeler için mevcut erişim sistemleriyle (retrieval systems) değerlendirmek üzere tasarlanmış yeni bir karşılaştırmalı değerlendirme veri seti sunmaktadır. Bu veri seti çeşitli alanlardan, dillerden ve görsel öğelerden (grafik, tablo, şema vs.) belgeler içerir. Mevcut sistemlerin görsel detayları ne kadar etkili kullanabildiğini ölçmek için tasarlanmıştır. Bu benchmark, sayfa düzeninde belge erişimini değerlendirir. Yani sorgu verildiğinde, sistemin ilgili sayfa/sayfaları ne kadar iyi bulduğunu ölçer. Hem metin hem de görsel bulunan belgeler üzerinde değerlendirme yapılır.

Makalede akademik görevler (DocVQA, InfoVQA, TAT-DQA, arXivQA, TabFQuAD) ve pratik görevler (Enerji, devlet raporları, sağlık, yapay zeka, çevre raporları) kategorileri altında toplanmış veri setleri görülmektedir.

- DocVQA: UCSF Endüstri Belge Kütüphanesi'nden taranmış belgeler.
- InfoVQA: Web'den toplanan infografikler.
- TAT-DQA: Yüksek kaliteli finansal raporlar.
- arXivQA: arXiv'den alınan bilimsel figürler.
- TabFQuAD: Fransız endüstriyel PDF'lerden çıkarılan tablolar.

Sistemde nDCG, Recall@K, MRR gibi standart erişim metrikleri kullanılmış, nDCG@5 değerleri raporlanmıştır.

Dataset	Language	# Queries	# Documents	Description
Academic Tasks				
DocVQA	English	500	500	Scanned documents from UCSF Industry
InfoVQA	English	500	500	Infographics scrapped from the web
TAT-DQA	English	1600	1600	High-quality financial reports
arXiVQA	English	500	500	Scientific Figures from arXiv
TabFQuAD	French	210	210	Tables scrapped from the web
Practical Tasks				
Energy	English	100	1000	Documents about energy
Government	English	100	1000	Administrative documents
Healthcare	English	100	1000	Medical documents
AI	English	100	1000	Scientific documents related to AI
Shift Project	French	100	1000	Environmental reports

ColPali Modeli

ColPali, belge sayfalarının görüntülerinden yüksek kaliteli çok vektörlü gömmeler üretmek için eğitilmiş bir VLM'dir. ColPali, PaliGemma-3B adlı bir VLM'yi temel alır. Bu model görsel ve metinsel bilgileri birleştirir. Özellikle belge anlama konusunda başarılıdır. Mevcut modeller belgeleri tek bir vektörle temsil ederken, ColPali belgenin her sayfasını birden fazla vektörle temsil eder. Çoklu vektörle temsil etmesi metnin farklı bölümlerinin (örneğin: metin, tablo, figür) ayrı ayrı işlenmesini sağlar.

Geç etkileşim (late interaction) mekanizması, sorgu ve belgeler arasındaki benzerliği hesaplamak için kullanılır. Sorgu ve belge önce bağımsız olarak işlenir. Ardından bu gömmeler (embeddings) arasındaki benzerlik hesaplanır.

Sonuçlar

ColPali, ViDoRe benchmark'ta metin tabanlı sistemlere ve diğer VLM'lere göre üstün performans gösteriyor. Ayrıca görsel ağırlıklı belgelerde de oldukça başarılı. Ek olarak metin tabanlı öğelerde de mevcut sistemlerden önde.

Sorgulama gecikmesi oldukça düşük. Bu modelin bir sorguyu işlemesi ortalama 30 ms. sürüyor. İndeksleme süreleri ise geleneksel sistemlere göre daha hızlı çünkü belge sayfalarını doğrudan görsel olarak işliyor.

Ablasyon (Ablation) Çalışmaları

Ablasyon çalışmaları, modelin farklı bileşenlerinin performansa etkisini anlamak için yapılan deneylerdir. Bu bölümde model üzerinde yapılan bu çalışmalara değineceğiz.

- Daha büyük modeller daha iyi performans göstermiştir ancak train ve sorgulama süreleri artmıştır.
- Daha fazla görsel parça, modelin daha detaylı bilgi çıkarmasını sağlamıştır ancak bellek kullanımını artırmıştır.
- Görsel kodlayıcının train esnasında güncellenmesi performansta bir düşüşe neden oluyor. Bu da görsel bileşenin önceden eğitilmiş halinin yeterli olduğunu gösteriyor.
- Sorguya özel token'ların eklenmesi başka dillerdeki performansı etkiliyor. Ancak İngilizce görevlerde bir fark yaratmıyor.
- Sadece en zor negatif örnekleri dikkate alan bir kayıp fonksiyonu kullanmak performansta hafif bir düşüşe sebep oluyor. Bu, tüm negatif örneklerin dikkate alınmasının önemli olduğunu gösteriyor.
- Model yeni görevlere hızla uyum sağlayabiliyor.

- Daha gelişmiş VLM'ler, modelin performansını daha da artırıyor. Bu sayede daha iyi modellerin erişimde daha iyi sonuçlar veriyor.
- Eğitim verilerinde farklı veri dağılımlarına genelleme yapabiliyor.

Gelecek Çalışmalar

Makalede ColPali'nin gelecekte hangi alanlarda kullanılabileceği tartışılmıştır. Bu kısımda bu başlıkları basitçe ele alacağız:

- Görsel Tabanlı RAG Sistemleri: Görsel ve metin tabanlı sorgu cevap sistemlerini birleştirme.
- Güvenilirlik: Modelin ne zaman emin olmadığını belirleyen güven tahmin teknikleri.
- Çok Dilli Destek: Modelin çok daha fazla dilde çalışması, özellikle de kaynağı az olan dillerde çalışması.
- Yeni görevler: Belge özetleme, sınıflandırma gibi görevlerde test edilmesi.
- Veri ve Model İyileştirmeleri: Daha büyük ve çeşitli veri setleri ile daha güçlü modeller kullanma.
- Uzun Belgelerin işlenmesi: Uzun belgeler ve kitap raporları gibi karmaşık belgeleri daha iyi işleme.

Sonuç

i) Ana Bulgular

Colpali belge erişimindeki görsel özellikleri doğrudan kullanabilmesi sayesinde mevcut metin tabanlı modelleri geride bırakacak bir performans gösterdi. Görsel ağırlıklı belgelerde üstün başarıya sahip. Ayrıca metin ağırlıklı belgelerde, metin tabanlı modellere göre üstün gelmiştir.

ii) Endüstriyel Potansiyel

Belge sayfalarını doğrudan görsel olarak işlediği için indeksleme süresi önemli derecede düşük. Sorgu terimleri ile belge görselleri arasındaki benzerlik haritaları modelin nasıl çalıştığını anlamayı kolaylaştırıyor. Ek olarak ciddi derecede seri bir işleme hızına sahip. Sorguları ortalama 30 ms içerisinde işleyebiliyor.

iii) Açık Kaynak Katkısı

ViDoRe benchmark, ColPali modeli ve tüm kaynaklar açık kaynak olarak paylaşılmıştır.

Kaynakça

1. https://arxiv.org/pdf/2407.01449