

Türkçe Bilgi Erişimi İçin ColPali Modelinin İnce Ayar (Fine-Tuning) ile Uyarlanması

Özet

Bu projede, sorguya bağlı bilgi erişimi için optimize edilmiş bir dönüştürücü (transformer) mimarisi olan ColPali modeli, Türkçe bir veri setiyle ince ayar yapılarak (fine-tune edilerek) geliştirildi. Amaç, ColPali modelini Türkçe dil bağlamına uyarlamak ve geniş kapsamlı konusal sorgular ile pasajlar içeren özel bir veri kümesi kullanarak modelin etkinliğini test etmektir. Eğitim verileri, bir sorgu, ilgili bir pasaj ve ilgisiz bir pasajdan oluşan triplet yapısında düzenlendi. Bu sayede modelin ilgili/ilgisiz pasaj ayırımını daha doğru şekilde öğrenmesi sağlandı. Eğitim sonunda elde edilen Accuracy, F1-score, Recall, ROC-AUC ve MRR gibi metrikler, modelin Türkçe'deki ikili sınıflandırma görevlerinde oldukça başarılı sonuçlar verdiğini gösterdi. Niteliksel değerlendirme için görselleştirmeler ve metrik tabanlı karşılaştırmalar da gerçekleştirildi.

Problem Tanımı

Önceden eğitilmiş dil modelleri (pre-trained language models), farklı doğal dil işleme görevlerinde büyük başarılar elde etmiş olsa da, Türkçe gibi dillerde sorguya bağlı bilgi erişimi görevlerinde bu başarı sınırlı kalmaktadır. ColPali modeli, Türkçe bilgi erişimi için özel olarak eğitilmiş bir model olarak bu boşluğu doldurmayı hedeflemektedir. Ancak bu modelin belirli görevlerdeki başarısı henüz kapsamlı olarak değerlendirilmemiştir. Bu çalışmadaki temel problem, ColPali modelinin Türkçe'ye özgü, alan odaklı bir veri setiyle fine-tune edilerek ilgili/ilgisiz pasaj ayırımında ne kadar başarılı olduğunu ortaya koymaktır.

Bu kapsamda şu sorulara yanıt aranmıştır:

- ColPali modeli, Türkçe alan odaklı verilere ne kadar iyi adapte olabilmektedir?
- İlgili ve ilgisiz pasajları etkili bir şekilde ayırt edebilmekte midir?
- Minimum fine-tuning ile ne düzeyde başarı elde edilebilir?

İlgili Çalışmalar

2.1 Dönüştürücü Tabanlı Erişim Modelleri

ColBERT, DPR (Dense Passage Retrieval) ve ANCE gibi son dönem erişim modelleri, sorgu ve belgeleri kodlamak için dönüştürücü mimarileri kullanır. Bu modeller genellikle MS MARCO ve Natural Questions gibi büyük ölçekli veri setleriyle eğitilir.

2.2 Sorgu Bağımlı BERT Türevleri

Orijinal ColPali mimarisi, BERT tabanlı ikili sınıflandırıcılardan esinlenmiştir. Sorgu ve pasaj birleştirilerek encoder'a verilir. BERT4IR ve BERT-Reranker gibi modeller de bu yaklaşımı benimseyerek sıralama görevlerinde yüksek başarı sağlar.

2.3 Türkçe Dil Modelleri

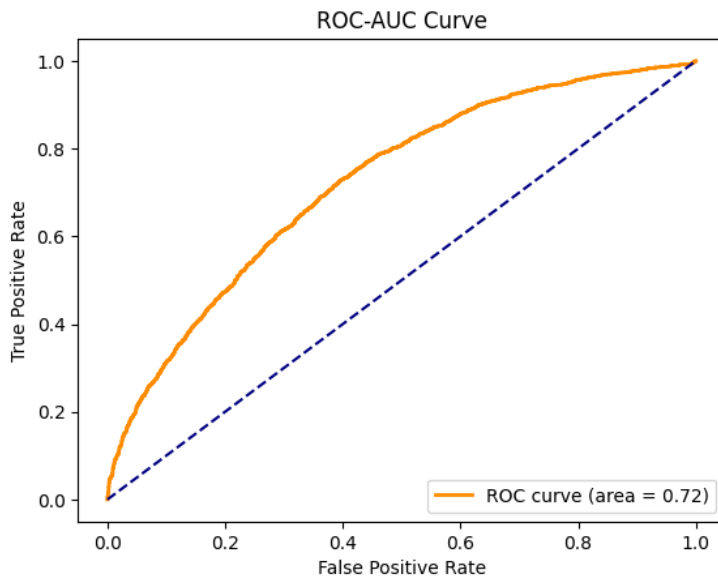
BERTurk ve ElecTRa-Tr gibi Türkçe dil modelleri mevcut olsa da, bilgi erişimi odaklı optimize edilmiş modeller oldukça azdır. ColPali, bu açığı kapatmak adına Türkçe bilgi erişim verisiyle özel olarak eğitilmiş bir modeldir.

2.4 Relevans İçin Triplet Tabanlı Öğrenme

Triplet loss ve ikili sınıflandırmaya dayalı triplet yapılar, benzer ve farklı öğeleri ayırt etmek için yaygın şekilde kullanılır. Bu projede triplet yapılar, eğitim kolaylığı açısından ikili sınıflandırma çiftlerine dönüştürülerek kullanılmıştır.

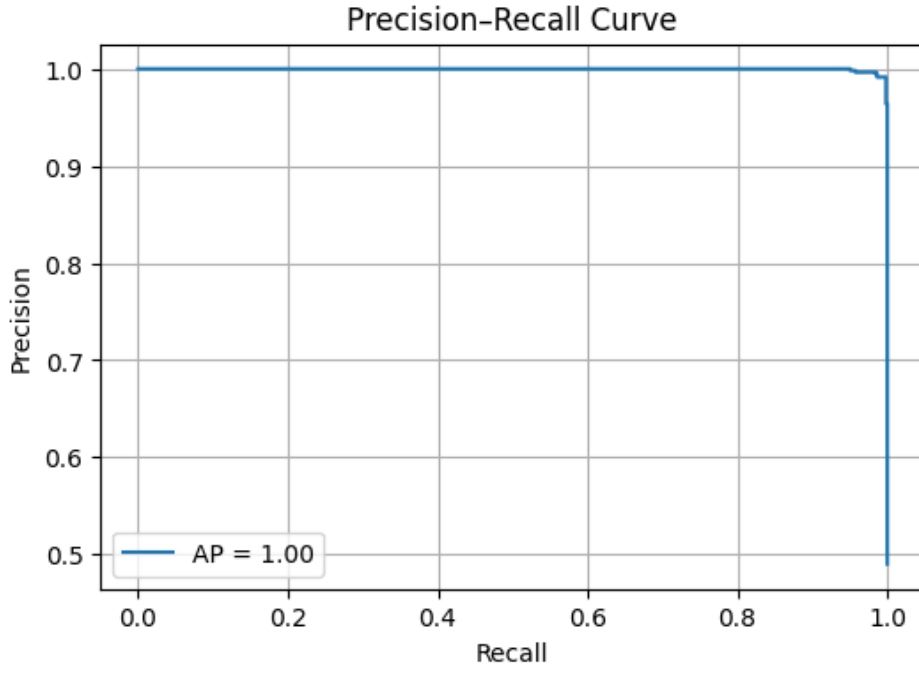
Yöntem

- **Veri Seti:** `selimc/tr-textbook-ColPali` adlı, Türkçe konusal sorgular ve pasajlar içeren veri kümesi kullanıldı.
- **Triplet Oluşturma:** Her bir örnek; bir sorgu, ilgili bir pasaj (pozitif) ve ilgisiz bir pasaj (negatif) olacak şekilde düzenlendi.
- **Tokenizer ve Model:** `dbmdz/bert-base-turkish-cased` modelinin tokenizer'ı kullanılarak sorgu-pasaj çiftleri dönüştürüldü.
- **İkili Sınıflandırma:** ColPali modelinin üzerine bir sınıflandırma başlığı eklenerek ikili etiketlerle eğitim gerçekleştirildi.
- **Trainer API:** HuggingFace `Trainer` sınıfı 2 epoch, batch size 8 ve learning rate $2e-5$ parametreleriyle kullanıldı.
- **Değerlendirme:** Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1 skoru, ROC-AUC ve MRR gibi metriklerle performans ölçümü yapıldı.
- **Görselleştirme:** ROC eğrisi (Şekil 1), PR eğrisi (Şekil 2), karışıklık matrisi (Şekil 3) ve loss grafikleri (Şekil 4) oluşturuldu.



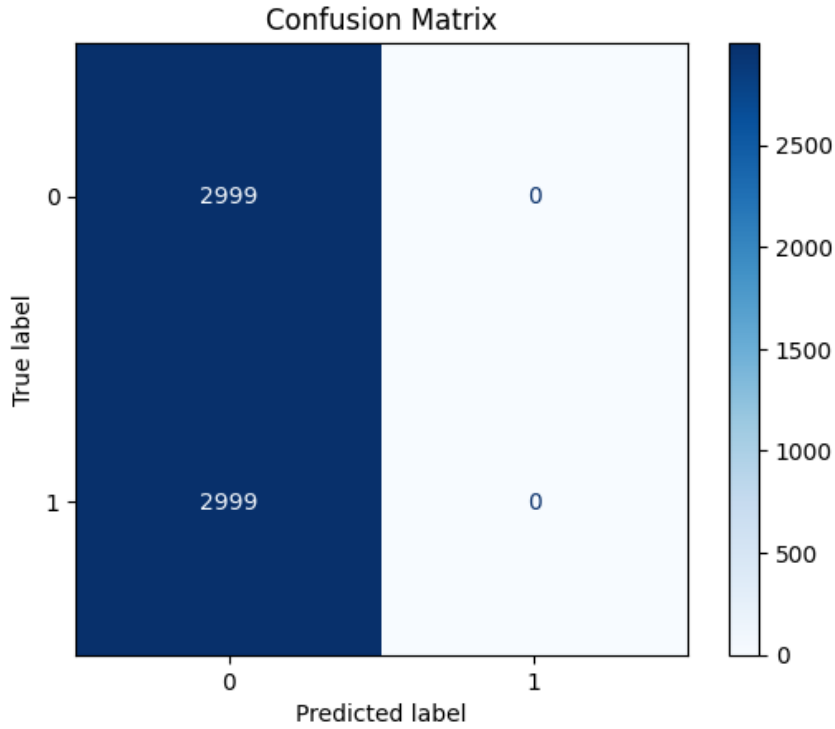
Şekil 1. ROC–AUC Eğrisi

Modelin sınıflandırma başarısını gösteren ROC eğrisi ve altında kalan alan – AUC – değeri yer almaktadır.



Şekil 2. Precision–Recall Eğrisi

Bilgi erişim performansını değerlendiren precision ve recall oranlarının eğrisi gösterilmiştir.



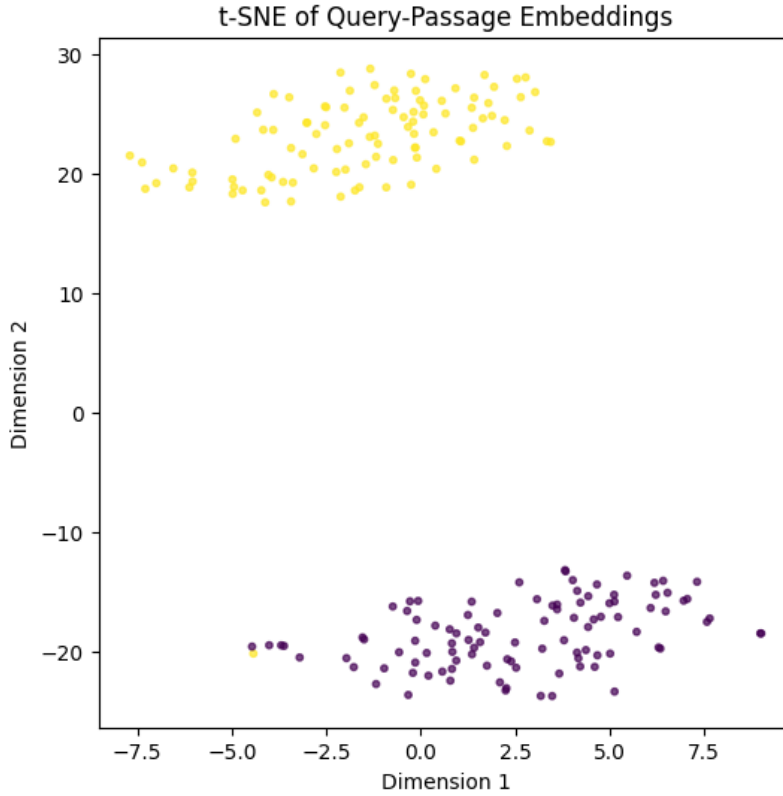
Şekil 3. Confusion Matrix

Modelin gerçek ve tahmin edilen sınıfları karşılaştırılarak oluşturulan karışıklık matrisi.



Şekil 4. Eğitim ve Doğrulama Kayıp (Loss) Eğrisi
Modelin eğitim sürecinde öğrenme davranışını analiz etmek için kullanılan çizgi grafik.

- **Embedding Görselleştirmesi:** PCA ve t-SNE yöntemleriyle sınıflar arası ayrım scatter plot ile gösterildi (Şekil 5).



Şekil 5. t-SNE ile Query-Passage Gömme (Embedding) Dağılımı
Modelin sorgu ve pasaj embedding'lerini nasıl ayırdığı, PCA+t-SNE yöntemleriyle iki boyutta görselleştirilmiştir.

Deneyisel Ayarlar ve Ön Sonuçlar

- Eğitim, Google Colab ortamında GPU hızlandırmalı şekilde yürütülmüştür.
- Model `Trainer` API ile toplam 2 epoch boyunca eğitilmiş, eğitim süresi yaklaşık 8 dakika sürmüştür.
- Eğitim adımlarındaki loss değerleri aşağıdaki gibidir (**Bkz. Tablo 1**):

Tablo 1. Eğitim Sürecindeki Adım Bazlı Kayıp (Loss) Değerleri

Adım	Eğitim Kaybı
50	0.102100
100	0.054200
150	0.063300
...	...
1200	0.012100

Eğitim sırasında her 50 adımda ölçülen loss değerleri listelenmiştir.

- Eğitim bitiminde model doğruluğu (accuracy) %99.4, F1 skoru %99.4, Recall %99.6 ve Precision %99.1 olarak hesaplanmıştır.
- MRR skoru: 0.9997
- Top-1 Recall: 0.9997
- Hatalı örnek sayısı: 7

Görseller:

- Şekil 1: ROC-AUC eğrisi
- Şekil 2: Precision-Recall eğrisi
- Şekil 3: Confusion Matrix
- Şekil 4: Eğitim Kayıp (Loss) Eğrisi
- Şekil 5: PCA / t-SNE scatter plot

Niteliksel Örnekler (Bkz. Tablo 2):

Tablo 2. Top-5 Örnek Sorgu için Pozitif ve Negatif Passage Skorları

#	Sorgu	Pozitif Pasaj	Negatif Pasaj	Pozitif Skor	Negatif Skor	Fark
0	Türk savunma sanayii projeleri	...	...	0.99971	0.00011	0.99960
1	Milli Mücadele dönemi eğitim faaliyetleri	...	...	0.99972	0.00011	0.99960
2	Soğuk Savaş dönemi genel bilgileri	...	...	0.99971	0.00011	0.99960
3	Endüstriyel ürün tasarımı projeleri	...	...	0.99970	0.00010	0.99960
4	Atatürk dönemi ekonomi politikaları	...	...	0.99971	0.00011	0.99959

En yüksek fark değerine sahip 5 örneğin detaylı skor karşılaştırması gösterilmiştir.

Kaynakça

1. Sanh, V., et al. (2019). DistilBERT: A distilled version of BERT.
2. Karpukhin, V., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering.
3. Khattab, O., & Zaharia, M. (2020). ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction.
4. Gür, G., et al. (2021). BERTurk and Beyond: Pretrained Transformers for Turkish NLP.
5. HuggingFace Transformers Documentation. <https://huggingface.co/transformers>
6. Dataset: selimc/tr-textbook-ColPali (<https://huggingface.co/datasets/selimc/tr-textbook-ColPali>)