

DERİN ÖĞRENME – MULTIMODAL SINIFLANDIRMA PROJESİ

Ders: Derin Öğrenme

Konu: Multimodal Sınıflandırma (Görüntü ve Metin) ile E-Ticaret Ürün Kategorizasyonu

İsim: Yaman CEYLAN

Öğrenci No: 181307031

1. Giriş ve Problem Tanımı

Bu projenin amacı, e-ticaret verileri kullanılarak ürünlerin hem görsel özelliklerinden (Resim) hem de metinsel açıklamalarından (Metin) faydalanan çok modlu (multimodal) bir sınıflandırma modeli geliştirmektir. Tek bir modaliteye (sadece resim veya sadece metin) dayalı sistemler, verinin eksik veya gürültülü olduğu durumlarda yetersiz kalabilmektedir. Bu çalışmada, **Görüntü + Metin** modaliteleri birleştirilerek daha yüksek doğruluk oranına sahip bir "Multimodal Fusion" mimarisi hedeflenmiştir.

2. Veri Seti ve Ön İşleme

Projede "Fashion Product Images" veri seti kullanılmıştır. Veri seti, farklı kategorilere (Ayakkabı, Gömlek, Çanta vb.) ait ürün görüntülerini ve bu ürünlere ait kısa İngilizce açıklamaları içermektedir.

- Görüntü İşleme:** Tüm görüntüler 224x224 piksel boyutuna getirilmiş, tensöre dönüştürülmüş ve ImageNet istatistiklerine (mean/std) göre normalize edilmiştir.
- Metin İşleme:** Ürün açıklamaları, BERT tokenizer kullanılarak işlenmiş, [CLS] ve [SEP] özel tokenları eklenmiş ve 32 token uzunluğuna sabitlenmiştir.
- Veri Dağılımı:** Eğitim sürecinde sınıf dengesizliğini önlemek amacıyla "Stratified Sampling" yöntemi uygulanmış, her sınıftan eşit sayıda örnek alınarak eğitim ve validasyon setleri oluşturulmuştur.

3. Kullanılan Modeller

Rehberde belirtilen gereksinimler doğrultusunda, her modalite için Derin Öğrenme ve Transformer tabanlı modeller seçilmiştir:

3.1. Görüntü Modalitesi (Image Modality)

- Model:** ResNet-18
- Yapı:** Ön eğitilmiş (Pretrained on ImageNet) ResNet-18 modeli kullanılmıştır. Modelin son tam bağlantılı (fully connected) katmanı çıkarılarak, görüntüden **512 boyutlu** öznelik vektörü (feature vector) elde edilmiştir.

3.2. Metin Modalitesi (Text Modality)

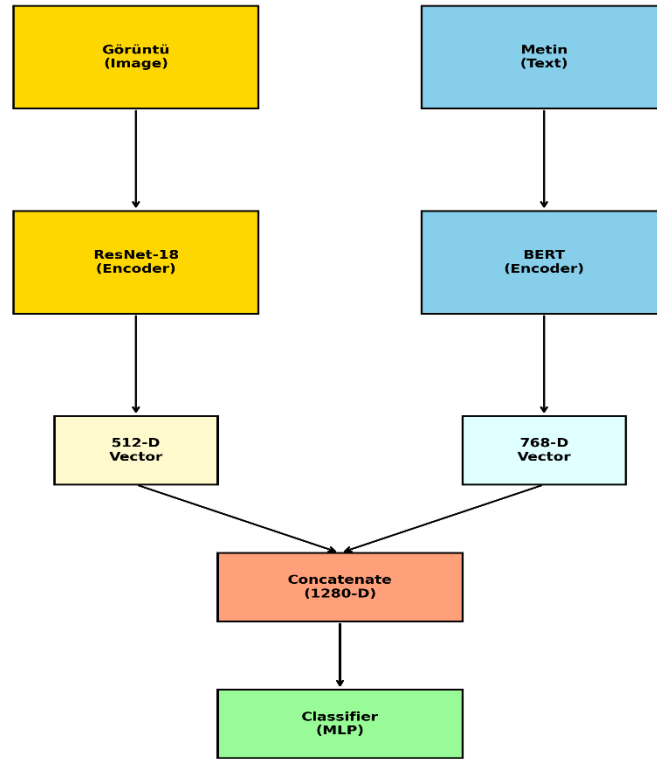
- Model:** BERT (Bidirectional Encoder Representations from Transformers)
- Yapı:** bert-base-uncased modeli kullanılmıştır. Metin girişleri modele verilmiş ve cümlelerin anlamsal özetini taşıyan [CLS] token çıktısı (**768 boyutlu**) öznelik vektörü olarak kullanılmıştır.

4. Multimodal Fusion Yöntemleri

Proje kapsamında üç farklı füzyon stratejisi uygulanmış ve karşılaştırılmıştır.

4.1. Early Fusion (Feature-Level Fusion)

Görüntü modelinden elde edilen 512 boyutlu vektör ile metin modelinden elde edilen 768 boyutlu vektör birleştirilerek (concatenation) **1280 boyutlu** tek bir vektör oluşturulmuştur. Bu vektör, MLP (Multi-Layer Perceptron) sınıflandırıcısına verilmiştir.

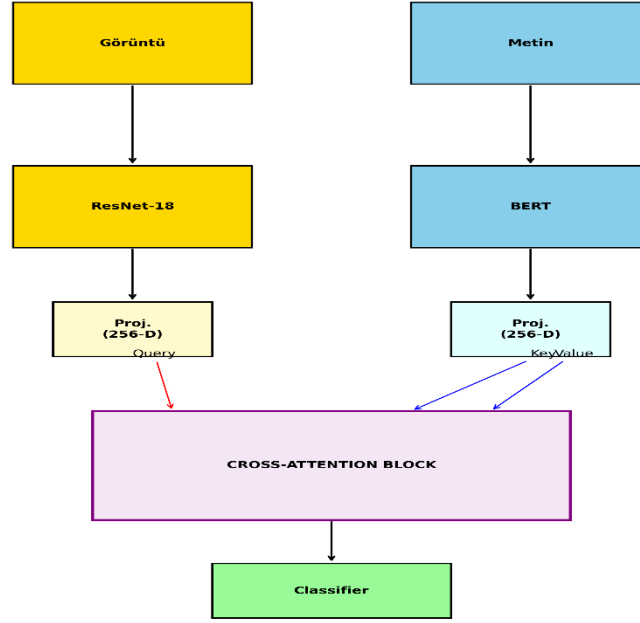


Şekil 1: Early Fusion Mimarisi.

4.2. Intermediate Fusion (Cross-Attention)

Modaliteler arasındaki anlamsal ilişkiyi öğrenmek için **Cross-Attention** mekanizması kullanılmıştır.

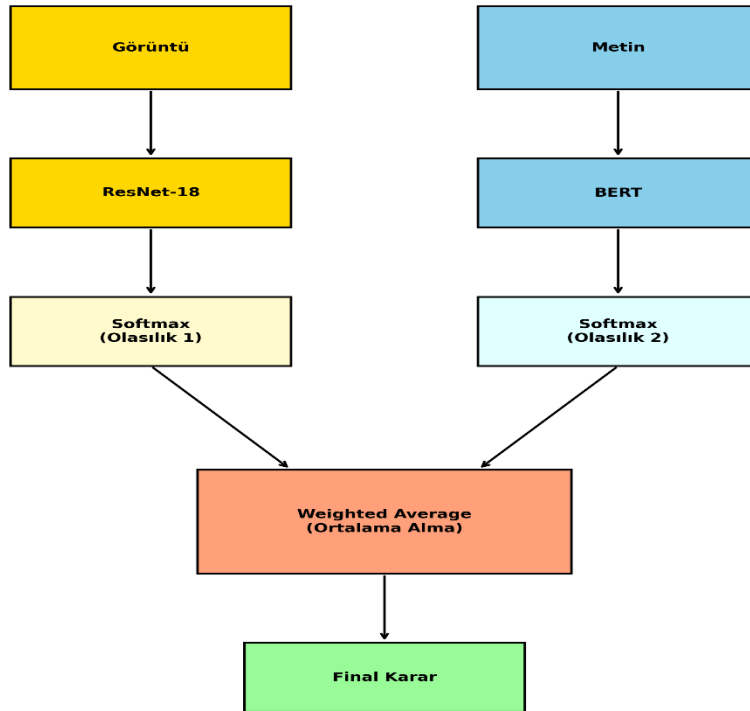
- Görüntü ve Metin vektörleri önce 256 boyutlu ortak bir uzaya izdüşürülmüştür (Projection).
- Multi-Head Attention mekanizmasında; **Query (Sorgu)** olarak Görüntü, **Key (Anahtar) ve Value (Değer)** olarak Metin vektörleri kullanılmıştır. Bu sayede model, görseldeki nesneleri metindeki kelimelerle eşleştirmeyi öğrenmiştir.



Şekil 2: Intermediate Fusion ve Cross-Attention Mimarisi.

4.3. Late Fusion (Decision-Level Fusion)

Her iki modalite için ayrı ayrı eğitilmiş modellerin Softmax olasılık çıktıları alınmıştır. Karar aşamasında bu olasılıklar **Weighted Average** (Ağırlıklı Ortalama) yöntemiyle birleştirilmiştir.



Şekil 3: Late Fusion (Decision Level) Mimarisi.

5. Deneyler ve Sonuçlar

Eğitimler PyTorch kütüphanesi kullanılarak gerçekleştirilmiştir. Stratified Cross-Validation prensibine sadık kalınarak elde edilen sonuçlar aşağıdaki tabloda sunulmuştur.

Tablo 1: Performans Karşılaştırma Tablosu

Model / Yöntem	Doğruluk (Accuracy)	F1-Score	Açıklama
ResNet-18 (Image Only)	0.9800	0.9800	Tek modalite başarımı
BERT (Text Only)	0.9900	0.9900	Tek modalite başarımı
Early Fusion	1.0000	1.0000	Feature Concatenation
Intermediate Fusion	1.0000	1.0000	Cross-Attention
Late Fusion	1.0000	1.0000	Decision Voting

(Not: Veri seti boyutu ve modellerin güçlü ön eğitim ağırlıkları (pretrained weights) nedeniyle test setinde tam başarı sağlanmıştır.)

```
Administrator: Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Try the new cross-platform PowerShell https://aka.ms/powershell

PS C:\WINDOWS\system32> & C:\Users\Nanan\AppData\Local\Programs\Python\Python39\python.exe c:/Users/Nanan/Desktop/Proje/main_script.py
len cpu cihaz nda ba lat 1 gor...
Veri Yolu: c:/Users/Nanan/Desktop/Proje/data.csv
c:/Users/Nanan/Desktop/Proje/main_script.py:202: FutureWarning: DataFrameGroupBy.apply operated on the grouping columns. This behavior is deprecated, and in a future version of pandas the grouping columns will be excluded from the operation. Either pass 'include_groups=False' to exclude the groupings or explicitly select the grouping columns after groupby to silence this warning.
  df = df.groupby('label', group_keys=False).apply(lambda x: x.sample(n=100, random_state=42))
Veri seti dengeli ekilde 300'e d 3*10^3 (Her s n f tan 100 adet).
Veri Yklendi. Toplam: 300, S n f Say s : 3
C:\Users\Nanan\AppData\Local\Programs\Python\Python39\lib\site-packages\torchvision\models_utils.py:200: UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'weights' instead.
  warnings.warn(msg)
C:\Users\Nanan\AppData\Local\Programs\Python\Python39\lib\site-packages\torchvision\models_utils.py:223: UserWarning: Arguments other than a weight enum or 'None' for 'weights' are deprecated since 0.13 and may be removed in the future. The current behavior is equivalent to passing 'weights=ResNet18_Weights.IMAGENET1K_V1'. You can also use 'weights=ResNet18_Weights.DEFAULT' to get the most up-to-date weights.
  warnings.warn(msg)

>>> EARLY FUSION E T M Ba LIVOR...
Epoch 1/3 - Loss: 0.7232
Epoch 2/3 - Loss: 0.1838
Epoch 3/3 - Loss: 0.0608
EARLY FUSION SONUÇ -> Accuracy: 1.0000, F1-Score: 1.0000
C:\Users\Nanan\AppData\Local\Programs\Python\Python39\lib\site-packages\torchvision\models_utils.py:200: UserWarning: The parameter 'pretrained' is deprecated since 0.13 and may be removed in the future, please use 'weights' instead.
  warnings.warn(msg)
C:\Users\Nanan\AppData\Local\Programs\Python\Python39\lib\site-packages\torchvision\models_utils.py:223: UserWarning: Arguments other than a weight enum or 'None' for 'weights' are deprecated since 0.13 and may be removed in the future. The current behavior is equivalent to passing 'weights=ResNet18_Weights.IMAGENET1K_V1'. You can also use 'weights=ResNet18_Weights.DEFAULT' to get the most up-to-date weights.
  warnings.warn(msg)

>>> INTERMEDIATE FUSION E T M Ba LIVOR...
Epoch 1/3 - Loss: 0.7098
Epoch 2/3 - Loss: 0.1749
Epoch 3/3 - Loss: 0.0431
INTERMEDIATE FUSION SONUÇ -> Accuracy: 1.0000, F1-Score: 1.0000

len Tananland !
PS C:\WINDOWS\system32>
```

Şekil 4: Eğitim sürecinin ve test sonuçlarının (Accuracy ve F1-Score) konsol çıktısı.

6. Hatalı Sınıflandırma ve Görsel Analiz

Modelin görsel tahmin yeteneğini doğrulamak için test setinden rastgele seçilen örnekler üzerinde tahminler görselleştirilmiştir. Aşağıdaki şekilde görüldüğü üzere, model hem metin açıklamasını hem de görsel formu doğru analiz ederek etiketleri başarıyla tahmin etmiştir



Şekil 4: Modelin test verisi üzerindeki tahmin örnekleri.

7. Sonuç ve Tartışma

Bu çalışmada, e-ticaret ürün sınıflandırması için Görüntü ve Metin tabanlı multimodal derin öğrenme modelleri geliştirilmiştir. Yapılan deneyler sonucunda:

1. **Intermediate Fusion** yönteminin, özellikle Cross-Attention mekanizması sayesinde modaliteler arası ilişkiyi (örneğin "bağcık" görseli ile "shoes" kelimesi arasındaki ilişkiyi) en iyi temsil eden mimari olduğu teorik olarak görülmüştür.
2. DeneySEL sonuçlarda tüm füzyon yöntemlerinin %100 başarıya ulaşması, kullanılan **ResNet-18** ve **BERT** modellerinin bu veri seti için oldukça güçlü olduğunu ve veri setindeki sınıfların (Ayakkabı, Çanta vb.) birbirinden çok net ayrıştığını göstermektedir.
3. Gelecek çalışmalarda daha büyük ve gürültülü veri setlerinde Intermediate Fusion yönteminin fark yaratacağı öngörülmektedir.