

EuroSAT'tan Gerçeę: Bir Tarım Arazisi Sınıflandırma Modelinin %96'dan %70'e Düşüş Hikayesi

Bir bitirme projesinin, "benchmark skoru her şeyi anlatmaz" dersini nasıl öğrettięi üzerine.

Başlangıç: Basit Bir Soru

Bir uydu görüntüsüne bakıp "bu tarım arazisi mi?" diye sorabilen bir model yapmak istedim. Kulaęa basit geliyor — ama işin içine girince fark ettim ki asıl zor kısım modeli eğitmek deęil, modelin gerçekten güvenilir olup olmadığını anlamakmış.

Bu yazı, EuroSAT veri setiyle başlayıp %96,3 doğruluęa ulaşan, sonra kendi topladıęım gerçek Sentinel-2 görüntüleriyle test edilince %70,5'e düşen bir modelin hikayesi. Düşüşün kendisi kötü haber deęildi aslında — tam tersine, projenin en deęerli bulgusu oradan çıktı.

Problem: 10 Sınıf, Bir Karar

EuroSAT veri seti, Sentinel-2 uydusundan alınmış 64×64 piksellik görüntüleri 10 arazi örtüsü sınıfına ayırıyor: AnnualCrop, Forest, HerbaceousVegetation, Highway, Industrial, Pasture, PermanentCrop, Residential, River, SeaLake.

Ama benim asıl istedięim tek bir soruya cevap vermektir: bu arazi tarım arazisi mi, deęilse ne? Bunun için üç sınıfı (AnnualCrop, PermanentCrop, Pasture) "tarım", geri kalan yediyi "tarım deęil" olarak tanımladım — ve modelin ürettięi 10 sınıflı olasılık dağılımının üzerine, güven skoruna dayalı basit bir karar katmanı ekledim:

```
if guven < 0.5:
    "Emin deęilim"
elif sinif in {"AnnualCrop", "PermanentCrop", "Pasture"}:
    "Tarım arazisi"
else:
    "Tarım arazisi deęil"
```

Adım Adım: Basitten Karmaşıęa

Sıfırdan büyük bir modele atlamak yerine, her adımın gerçek bir katkı sağladığını ölçerek ilerlemeyi tercih ettim:

1. Önce EDA. Veri setini kör kör kullanmadan önce eksik dosya, bozuk dosya, kopya görüntü taraması yaptım — hepsi 0 çıktı, veri temizdi. Ama daha ilginç bir şey buldum: Pasture ve HerbaceousVegetation sınıflarının ortalama parlaklık dağılımları örtüşüyordu. Bu, ileride model karışacak demekti — ve tahmin doğru çıktı.

2. Basit bir CNN, referans olarak. Üç konvolüsyon bloęu, GlobalAveragePooling, 95 bin parametre. %90,1 test accuracy. Küçük ama iş görüyor.

3. Transfer learning — ama bir tuzakla. MobileNetV2 ve ResNet50'yi ImageNet ağırlıklarıyla denedim. İlk denemede MobileNetV2 sadece %84 civarında takılı kaldı — sebebini anlamam biraz sürdü: 64×64 gibi küçük bir görüntüde, bu modellerin ImageNet'ten öğrendięi özellik haritaları verimli çalışmıyordu. Çözünürlüğü 96×96'ya çıkarınca aynı model %92'ye fırladı. Küçük bir detay, büyük bir fark.

4. Fine-tuning.** ResNet50, MobileNetV2'yi geçince (macro F1: 0,941 vs 0,907) onu seçip son 20 katmanını çok düşük bir öğrenme oranıyla ince ayara soktum. Sonuç: **%96,3 accuracy, %98,5 tarım/deęil ikili doğruluk.

Model	Test Accuracy	Macro F1
Basit CNN	%90,1	0,898
MobileNetV2 (frozen)	%91,3	0,907
ResNet50 (frozen)	%94,3	0,941
ResNet50 (fine-tuned)	**%96,3**	**0,961**

5. Grad-CAM ile modelin "gözünden" bakmak. Modelin bir görüntüde tam olarak nereye baktığını görselleştirdim — sadece doğru tahmin etmesini değil, doğru nedenle tahmin etmesini de merak ettim.

Asıl Test: Laboratuvarından Çıkmak

Buraya kadar her şey iyiydi. Ama EuroSAT'ın kendi test seti de sonuçta aynı dağılımdan geliyor — aynı sensör kalibrasyonu, aynı coğrafya havuzu, aynı zaman aralığı. Gerçek soru şuydu: bu model, hiç görmediği, farklı bir zamanda çekilmiş, farklı bir yerden gelen görüntülerde ne yapar?

Bunu test etmek için Copernicus Browser üzerinden, Türkiye'nin 10 farklı bölgesinden (Konya Ovası'ndan AnnualCrop, Rize'nin Çamlıhemşin ilçesinden Forest, Van Gölü'nden SeaLake, Fırat Nehri'nden River...) kendi Sentinel-2 test setimi topladım. Sentinel-2'yi bilinçli seçtim — çünkü EuroSAT'ın kendisi de bu sensörden türetildi, yani sensör farkı değil, sadece coğrafya ve zaman farkını test etmiş oluyordum.

288 görüntülük, dengeli bir test seti çıktı ortaya. Ve sonuç:

%70,5 accuracy. %96,3'ten 25,8 puanlık bir düşüş.

Düşüşün İçinde Saklı Olan Hikaye

İlk refleks "model kötüy mü" demek olabilir. Ama confusion matrix'e bakınca çok daha ilginç bir tablo çıktı:

Bildiğim bir zayıflık, gerçek dünyada da tekrar etti. Pasture görüntülerinin %43'ü HerbaceousVegetation olarak yanlış sınıflandırıldı — tam olarak EDA'da parlaklık analizinde önceden işaret ettiğim karışıklık. Bu beni rahatlatmıştı: model rastgele değil, tutarlı bir şekilde hata yapıyordu.

Ama bir de hiç beklemediğim bir şey çıktı. SeaLake görüntülerinin %57'si Forest olarak sınıflandırıldı. EuroSAT'ın kendi test setinde bu ikisi hiç karışmamıştı. Neden şimdi karışıyordu?

En olası açıklama: topladığım bazı SeaLake görüntüleri (özellikle koyu lacivert, derin göller) ile modelin ImageNet'ten öğrendiği "koyu, düşük doku farklılığı olan alan" örüntüsü arasında bir çakışma oluşmuş olabilir. Bu, EuroSAT'ın kendi içinde hiç test edilemeyecek bir zafiyetti — sadece gerçek dünya verisiyle ortaya çıkabilirdi.

Bir de güven skorlarına baktım: doğru tahminlerin ortalama güveni %91,3 iken, yanlış tahminlerin ortalama güveni %79,5. Yani model çoğu zaman yanlış ama emindi. Bu, 0,5'lik güven eşiğimin (EuroSAT üzerinde makul görünen) gerçek dünyada aslında pek işe yaramadığını gösteriyordu — model %79 güvenle yanılırken, eşik onu hiç yakalayamıyordu.

Çıkardığım Ders

Bu projeden önce "iyi model" demek benim için "yüksek test accuracy'si olan model" demektir. Şimdi böyle düşünmüyorum.

%96,3'lük bir EuroSAT skoru, gerçek dünyada %70,5'e düşebiliyor — ve bu düşüşün kendisi rastgele değil, sistematik ve açıklanabilir bir örüntü taşıyor. Bir modelin ne kadar iyi olduğunu, sadece kendi test setinde değil, o test setinin dışına çıktığında nasıl davrandığına bakarak anlayabiliyorsunuz.

Bu yüzden, eğer bir makine öğrenmesi projesi yapıyorsanız ve elinizde sadece tek bir benchmark skoru varsa — bir adım daha atıp, modelinizi o benchmark'ın hiç görmediği bir veriyle test etmenizi öneririm. Muhtemelen düşüş göreceksiniz. Ama o düşüşün şekli, modelinizin gerçekte ne öğrendiğini, sayısal bir skordan çok daha dürüst anlatacak.

Projenin tam kodu ve teknik raporu [GitHub reposunda](#) mevcut. Sorularınız veya benzer bir domain shift deneyiminiz varsa, yorumlarda paylaşmaktan çekinmeyin.

Etiketler: #MachineLearning #DeepLearning #RemoteSensing #ComputerVision #TransferLearning