Learning Transferable Visual Models From Natural Language Supervision (CLIP)

Emir Özaktaş

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

1. Giriş ve Motivasyon

Geleneksel computer vision modelleri, fixed label set ile eğitilir. Ancak bu yöntem, modelin yeni kavramları öğrenmesini zorlaştırır. Her yeni görev için ek labeled data'ya ihtiyaç duyulur. Natural language supervision'ın bu zorlukları aşabileceği öngörülmektedir. Bu makalede *Contrastive Language-Image Pre-training (CLIP)* modeli tanıtılmaktadır. CLIP, internetten alınan 400 milyon görsel-metin çiftinden oluşan bir veri setiyle (WIT, WebImageText) eğitilmiş ve zero-shot öğrenme (sıfır atışlı öğrenme) kapasitesi test edilmiştir.

2. Model ve Yöntem

2.1. Natural Language Supervision

CLIP; görselleri, onlarla ilgili olan metin açıklamalarıyla öğrenir. Geleneksel modellerin aksine etiketlere bağımlı değildir.

2.2. Veri Seti

Mevcut çalışmalar genellikle MS-COCO veya YFCC100M gibi veri setlerini kullanmaktadır. Ancak yukarıda da belirtildiği gibi bu çalışmada internet üzerinden toplanmış 400 milyon görüntü-metin çiftinden oluşan bir veri seti (WIT) kullanılmıştır. Bu veri seti 500.000 sorgu (quest) kullanılarak oluşturulmuştur ve her bir sorgu için maksimum 20.000 görsel-metin çifti içermektedir.

2.3. Pre-Training

Çalışmada Contrastive Language-Image Pre-Training yöntemi kullanılmıştır. CLIP, görüntü-metin çiftlerinden oluşan bir eğitim setinde hangi görüntünün hangi metinle eşleştiğini tahmin etmek üzere eğitilir. Bu yöntem image ve text encoder'ları aynı anda eğiterek gerçek çiftlerin benzerliğini maksimize ederken yanlış çiftleri minimize eder.

2.4 Model Mimarisi

Çalışmada ResNet-50 ve Vision Transformer (VIT) yapıları kullanılmıştır. Text encoder olarak ise Transformer mimarisi kullanılmıştır.

3. Deney ve Gözlemler

3.1. Zero-Shot Transfer

CLIP daha önce hiç görmediği görevleri hiçbir ek eğitim veya ayar olmadan yapabiliyor. Bunu yapabilmesini sağlayan şey ise önceden eğitildiği sırada edindiği yetenekleri zero shot transfer ile farklı computer vision görevlerinde kullanabiliyor olması. Örneğin, bir nesne tanıma görevinde, o nesneyi hiç görmemiş olsa bile metinlerle olan ilişkisini kullanarak görevi başarıyla tamamlayabiliyor.

Çalışmada 30'dan fazla veri seti üzerinde testler yapılmış ve CLIP'in bu görevlerin çoğunda, spesifik olarak o konular için eğitilmiş modellerle rekabet edebildiği gözlemlenmiştir. Örneğin ImageNet adlı büyük bir image recognition veri setinde zero-shot ile ResNet-50 modelinin doğruluk oranını yakalayabilmiştir. Bu deneyler CLIP'in esnekliği ve gücü hakkında önemli kanıtlar sunmuştur.

3.2. Representation Learning

CLIP'in representation learning yetenekleri lineer problar kullanılarak değerlendirilmiştir. Bu değerlendirme, CLIP'in ImageNet modelini hem accuracy hem de computational efficiency açısından geride bıraktığını göstermiştir.

4. Sonuçlar

Bu çalışma, NLP alanındaki pre-training yöntemlerinin computer vision alanına uyarlanabileceğini göstermiştir. CLIP pre-training esnasında çeşitli görevleri öğrenmekte ve bu görevleri zero-shot transfer ile mevcut veri setlerine uygulayabilmektedir. Yeterli ölçekte, bu yaklaşım, göreve özel denetimli modellerle rekabet edebilmektedir.

Kaynakça

1. https://arxiv.org/pdf/2103.00020