Başlık: Prune Once for All: Sparse Pre-Trained Language Models

Yayın Tarihi: 10 Kasım 2021

Yazarlar: Ofir Zafrir, Ariel Larey, Guy Boudoukh, Haihao Shen, Moshe Wasserblat

Özet: Bu araştırma makalesi, seyrek önceden eğitilmiş dil modellerini eğitmek için "Prune Once for All" (Prune OFA) adlı yeni bir yöntem önermektedir. Ana fikir, ağırlık budama ve model damıtma işlemlerini ön eğitim aşamasında gerçekleştirmektir. Bu, seyrek önceden eğitilmiş modellerin oluşturulmasını sağlar ve bu modeller, yüksek seyreklik ve minimal doğruluk kaybını koruyarak çeşitli alt seviye görevler için ince ayar yapılabilir.

Ana katkılar şunlardır:

1. Önceden eğitim aşamasında seyrek önceden eğitilmiş dil modellerini eğitmek için yeni bir mimari-agnostik yöntem.
2. Bu seyrek önceden eğitilmiş modellerin, görev özel budama veya ayarlama gerektirmeden çeşitli alt seviye görevler için ince ayar yapılabilir olduğunu gösterme.
3. Diğer mimariler için sonuçları yeniden üretebilmek adına bir sıkıştırma araştırma kütüphanesi ve seyrek önceden eğitilmiş modellerin yayınlanması.

Yöntem, iki teknik içerir: ağırlık budama ve bilgi damıtma. Ağırlık budama, düşük büyüklükteki ağırlıkların kaldırılmasını içerir ve seyrek modeller oluşturur. Bilgi damıtma, bir öğrenci modelinin (seyrek model) bir öğretmen modelinin (orijinal yoğun model) tahminlerini taklit etmesini sağlar.

Yazarlar, yöntemlerini BERT-Base, BERT-Large ve DistilBERT'in ön eğitim aşamasında uygular ve bu modellerin seyrek önceden eğitilmiş versiyonlarını oluşturur. Daha sonra, seyrek önceden eğitilmiş modelleri SQuAD ve GLUE gibi çeşitli alt seviye görevler için ince ayar yaparlar ve standart yoğun modellerle karşılaştırıldığında minimal doğruluk kaybı gösterirler.

Yazarlar ayrıca, seyrek modelleri 8-bit hassasiyetine kullanarak kuantizasyon farkındalığı eğitimi kullanarak nasıl daha da sıkıştırılacağını gösterir ve BERT-Base, BERT-Large ve DistilBERT için bilinen en iyi sıkıştırma-doğruluk oranını gösterir.

Özetle, Prune OFA yöntemi, minimal doğruluk etkisiyle farklı görevler için ince ayar yapılabilen, önemli ölçüde sıkıştırılmış seyrek önceden eğitilmiş dil modelleri üretir ve dağıtım sırasında önemli hesaplama maliyetlerini tasarruf sağlar.

Başlık: Transformer-based approaches to Sentiment Detection

Yayın Tarihi: 13 Mart 2023

Yazarlar: Olumide Ebenezer Ojo, Hoang Thang Ta, Alexander Gelbukh, Hiram Calvo, Olaronke Oluwayemisi Adebanji, Grigori Sidorov

Özet: Çalışma, doğal dil işleme (NLP) alanındaki mevcut ilerlemelerin büyük ölçüde transfer öğrenme yöntemlerinin kullanılmasından kaynaklandığını vurgulamaktadır. Araştırmacılar, duygu tespiti problemine çözüm bulmak amacıyla, metin sınıflandırma için BERT, RoBERTa, DistilBERT ve XLNet gibi dört farklı transformer modelinin performansını incelemişlerdir. Bu modeller, metindeki felaket durumlarını tespit etme kapasitesi açısından karşılaştırılmıştır.

Tüm modeller, metindeki felaket tespiti için uygun olduğunu gösteren yeterince iyi performans göstermiştir. Özellikle RoBERTa transformer modeli, test veri setinde %82.6'lık bir skor ile en iyi performansı sergileyerek kaliteli tahminler için özellikle tavsiye edilmiştir.

Araştırmacılar ayrıca, öğrenme algoritmalarının performansının, ön işleme teknikleri, kelime dağarcığındaki kelimelerin doğası, dengesiz etiketleme ve model parametreleri gibi faktörler tarafından etkilendiğini keşfetmişlerdir. Bu bulgular, transformer tabanlı modellerin ve bu modellerin performansını etkileyen faktörlerin daha derinlemesine anlaşılmasına katkı sağlamaktadır.