DistilBERT Nedir?

DistilBERT, doğal dil işleme (NLP) alanında yaygın olarak kullanılan BERT (Bidirectional Encoder Representations from Transformers) modelinin, daha küçük ve daha verimli bir sürümüdür. Hugging Face tarafından geliştirilen DistilBERT, BERT'in sunduğu yüksek doğruluk seviyesini büyük ölçüde korurken, model boyutu ve hesaplama süresinde önemli kazanımlar sağlar. Özellikle mobil cihazlar, düşük kaynaklı sistemler veya hızın önemli olduğu senaryolarda tercih edilmektedir.

DistilBERT'in eğitimi, "knowledge distillation" (bilgi damıtımı) adı verilen bir yöntemle gerçekleştirilmiştir. Bu yöntem, büyük ve güçlü bir modelin ("öğretmen") bilgi ve davranışlarını, daha küçük ve hafif bir modele ("öğrenci") aktarmayı amaçlar. DistilBERT'in durumunda, öğretmen model olarak BERT kullanılmıştır. Öğrenci model, öğretmenin hem tahmin çıktılarından, hem de ara katman çıktılarından faydalanarak daha küçük bir yapı içerisinde benzer doğruluk performansına ulaşmayı öğrenir.

Modelin mimarisi, BERT'in 12 transformer katmanının yarısını yani sadece 6 katmanını kullanacak şekilde sadeleştirilmiştir. Bu sadeleştirmeye rağmen, DistilBERT modeli BERT'e kıyasla yaklaşık %40 daha az parametreye sahiptir ve %60 daha hızlı çalışır. Aynı zamanda, doğruluk kaybı %3'ten daha azdır, bu da pratik uygulamalarda performans farkını büyük ölçüde önemsiz hâle getirir.

DistilBERT, BERT'te kullanılan **Masked Language Modeling (MLM)** görevine dayalı olarak eğitilir. Bununla birlikte, BERT'te bulunan **Next Sentence Prediction (NSP)** görevine yer verilmemiştir. Eğitim sürecinde kullanılan toplam kayıp fonksiyonu, üç farklı bileşeni birleştirir: cross-entropy loss (L_{CE}), Kullback-Leibler divergence (L_{KLD}) ve cosine embedding loss (L_{COS}). Bu kayıp fonksiyonu aşağıdaki şekilde formüle edilir:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{KLD}} + \gamma \cdot \mathcal{L}_{\text{COS}}$$

Burada; cross-entropy loss, doğru etiketlere göre modelin tahmin performansını değerlendirir; KL divergence, öğrenci ve öğretmen modelin tahmin dağılımları arasındaki farkı ölçer; cosine loss ise gizil temsillerin benzerliğini korumayı hedefler. Bu şekilde, öğrenci model hem semantik hem de yapısal olarak öğretmen modele benzemeye çalışır.

Sonuç olarak, DistilBERT; düşük kaynaklı ortamlarda yüksek performanslı NLP çözümleri geliştirmek isteyen araştırmacılar ve geliştiriciler için oldukça güçlü ve esnek bir alternatiftir. Sentiment analizi, metin sınıflandırması, adlandırılmış varlık tanıma (NER) ve soru-cevap sistemleri gibi pek çok görevde kullanılabilir. Hugging Face Model Hub üzerinden kolayca erişilebilir ve önceden eğitilmiş sürümleri farklı dillerde mevcuttur.

Kaynak: GeeksForGeeks. (2023). *DistilBERT in Natural Language Processing (NLP)*. https://www.geeksforgeeks.org/nlp/distilbert-in-natural-language-processing/