3 Şubat 2024 Cumartesi

Nazım Anıl Tepe
524123002

CSE7052 NLP Dönem Projesi

Giriş

Dönem projesi kapsamında NLI task'ına ait bir dataset ile T5 mimarisine sahip Türkçe dil desteği olan pre-trained bir model fine-tune edildi. Daha sonrasında fine-tune edilen bu model, Mixture of Experts modelinde expert'lerden biri olarak kullanılması planlanmaktadır. Dokümanın ilerleyen bölümlerinde, yapılan denemelerin ve fine-tune işleminin detayları yer almaktadır.

Kullanılan Dataset

- Hugging Face üzerinde Türkçe destekli NLI dataset'lerine bakıldığında 100k üzeri satır sayısına sahip iki dataset göze çarpıyor.
  - nli_tr: SNLI ve MultiNLI dataset'lerinin Amazon Translate ile Türkçe'ye çevrilmesiyle oluşturulmuştur. Toplamda 983k satır içermektedir.
  - MoritzLaurer/multilingual-NLI-26lang-2mil7: MultiNLI, Fever-NLI, ANLI, LingNLI ve WANLI dataset'lerinin açık kaynak Machine Translation modelleriyle çevrilmesiyle oluşturulmuştur. Türkçe için toplamda 105k satır içermektedir.
- multilingual-NLI-26lang-2mil7 dataset'indeki satırlarda gözle görülür şekilde çeviri daha iyi olduğu için, ayrıca çalışmaya başlandığı sırada nli_tr dataset'ine erişilemediği için (şu an erişilebilmekte) bu çalışmada dataset olarak **MoritzLaurer/multilingual-NLI-26lang-2mil7** kullanıldı.
- Bu dataset 'hypothesis', 'premise', 'label' sütunlarını içermektedir.

Kullanılan Pre-trained Model

- Pre-trained model olarak T5 mimarisine sahip Türkçe destekli bir model kullanılması kararlaştırıldı.
- Multilingual olduğu için google/flan-t5-small ile çalışmalara başlandı. Çalışma devam ederken Boğaziçi Üniversitesi TABILAB tarafından 1.14b parametreli TURNA isimli Türkçe T5 modeli yayınlandı.
- TABILAB tarafından TURNA modelinin nli_tr dataset'iyle fine-tune edildiği **boun-tabi-LMG/turna_nli_nli_tr** modeli pre-trained model olarak kullanıldı.

Yapılan Denemeler

- Denemeler Google Colab Pro üzerinden 40 GB NVIDIA A100 GPU kullanarak yapıldı.
- Evaluation işlemi için 105k satır içeren dataset shuffle yapılıp 5000 satır ayrıldı, seed parametresi sayesinde hep aynı 5000 satırın ayrılacağından emin olundu, bu sayede ayrılan 5000 satırın modeller tarafından hiç görülmeyeceği garanti edildi.
- Dataset'teki satırlar, tokenizer'a verilmeden önce "hipotez: x['hypothesis'] önerme: x['premise']" formatına getirildi. Satırların label'ları integer olarak 0, 1, 2 değerlerini içeriyordu (sırasıyla entailment, neutral, contradiction anlamına gelecek şekilde), sırasıyla "gereklilik", "nötr", "çelişki" değerlerini içerecek şekilde string formatına dönüştürüldü, ardından tokenizer'a verildi.
- Training parametresi batch_size = 4 seçildi, batch_size = 8 seçildiğinde GPU out of memory hatası veriyor, batch_size = 2 seçildiğinde ise training işlemi çok uzun sürüyordu.
- İlk deneme için 20k satır kullanıldı, bunlar %80 train %20 test olarak ayrıldı. Epoch sayısı olarak 5 verildi. Train işlemi 4 saat 45 dakika sürdü, 31.7 GB GPU RAM kullanıldı. Bu model **aniltepe/turna-nlitr-finetuned-20k-5e** isimli modele push edildi.
- İkinci denemede 100k satır kullanıldı, bu satırlar yine %80 train %20 test olarak ayrıldı. Epoch sayısı bu sefer 1 verildi. Train işlemi 4 saat 40 dakika sürdü, 31.7 GB GPU RAM kullanıldı. Bu model **aniltepe/turna-nlitr-finetuned-100k-1e** isimli modele push edildi.
- Denemelerin birinde test amaçlı 50k satır kullanıldı, diğerlerinden farklı olarak batch_size = 8 verildi, Fp16 = True verildi, 3 epoch ile eğitildi. Train işlemi 3 saat 48 dakika sürdü, 33.7 GB GPU RAM kullanıldı. Batch_size = 8 verilmesine rağmen Fp16 = True verildiği için out of memory hatası alınmadı, GPU RAM kullanımı düştü ve işlem süresi kısaldı ancak her bir epoch için Loss ve ROUGE skorları 0 geldi, bottleneck oluştuğu için doğru sonuçlar üretemedi.

Evaluation

- Evaluation işlemi için modellere hiç gösterilmeyen 5000 satır kullanıldı.
- MoritzLaurer/multilingual-NLI-26lang-2mil7 dataset'iyle eğitilmemiş pre-trained model olarak kullanılan **boun-tabi-LMG/turna_nli_nli_tr** evaluate edildiğinde aşağıdaki skorlar elde edildi:
    - 'eval_loss': **0.91394025** 'eval_rouge1': **36.24** 'eval_rouge2': **36.18**

- 20k satır ile 5 epoch eğitim sonrası elde edilen **aniltepe/turna-nlitr-finetuned-20k-5e** modelinin evaluation skorları şu şekildedir:
    - 'eval_loss': **0.75274872** 'eval_rouge1': **36.76** 'eval_rouge2': **36.7**
- 100k satır ile 1 epoch eğitim sonrası elde edilen **aniltepe/turna-nlitr-finetuned-100k-1e** modelinin evaluation skorları şu şekildedir:
    - 'eval_loss': **0.74262571** 'eval_rouge1': **37.14** 'eval_rouge2': **37.14**