kaggle

NBME - Score Clinical Patient Notes Challenge

3rd Place Solution - Raja Biswas

Agenda

Agenda

1. Background
2. Summary
3. Training
4. Inference
5. Important findings

Background

Background

- Academic
  - B Tech in Civil Engineering - IIT Kanpur
  - PhD in Computational Mechanics - NUS
- Professional
  - NLP Data Scientist @ Evonik
- Prior experience in various NLP tasks
  - Classification
  - NER
  - Relation Extraction
  - QA
- Motivation
  - Solve challenging practical challenges
  - Learn and apply semi-supervised learning techniques

Summary

Summary

- Formulation as a multi-label token classification task
  - Predict beginning, end and inside of answer spans
- Making full use of provided unlabelled data
  - Task Adaptation
  - Meta Pseudo Labels
  - Knowledge Distillation

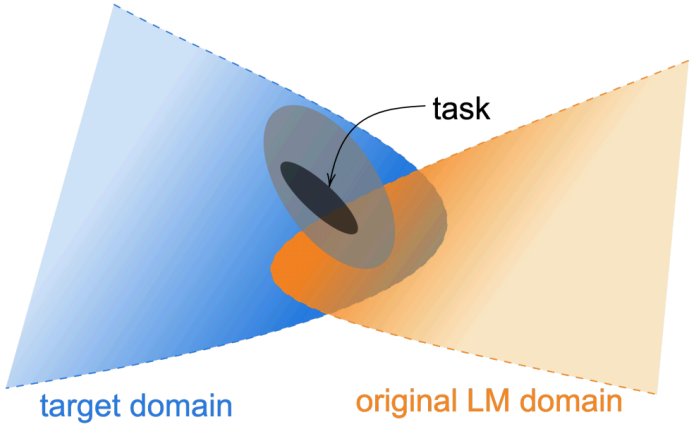
```
==========================================================================================
Prompt:

Auditory hallucination once
==========================================================================================
```

Pt is a 67 yo F presenting with insomnia x 3 weeks since the death of her son. Pt has difficulty falling asleep, tosses and turns while sleeping, wakes up early. Sleeps 4-5 hours/day. Nothing worsens/alleviates sx. Tried Ambien for 5 days but did noto feel like it helped and has stopped taking it. Patient reports feeling drained and tired from lack of sleep and eating more than usual. She reports **truth** seeing her son in the kitchen one day and also thinking she heard a party next door when there was none **truth** . she feels sad and has a loss of normal interests and low neergy, but does not feel any guilt, psychomotor agitation/slowing down, or any suicidal ideation. No fevers, cold sx, wt loss or wt gain, changes to elimination.

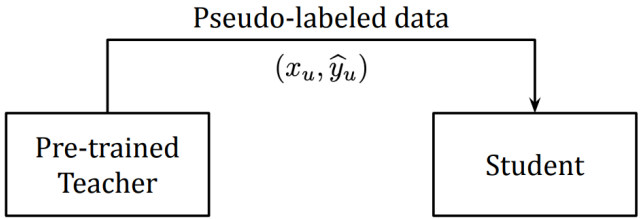
Training

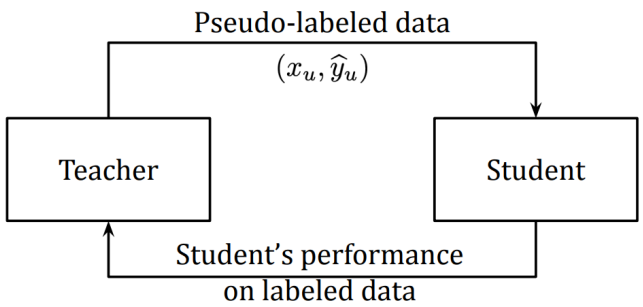
- Task Adaptation
  - Widely used Language Models (LM) are trained with text from a wide variety of sources e.g. books, wikipedia, crawled data
  - Adaption to the domain of a target task (e.g. NBME patient notes) gives performance boost
  - NBME patient notes are leveraged for task adaptation using MLM i.e. Masked Language Modelling



Reference: Gururangan, Suchin, et al. "Don't stop pretraining: adapt language models to domains and tasks." arXiv preprint arXiv:2004.10964 (2020).

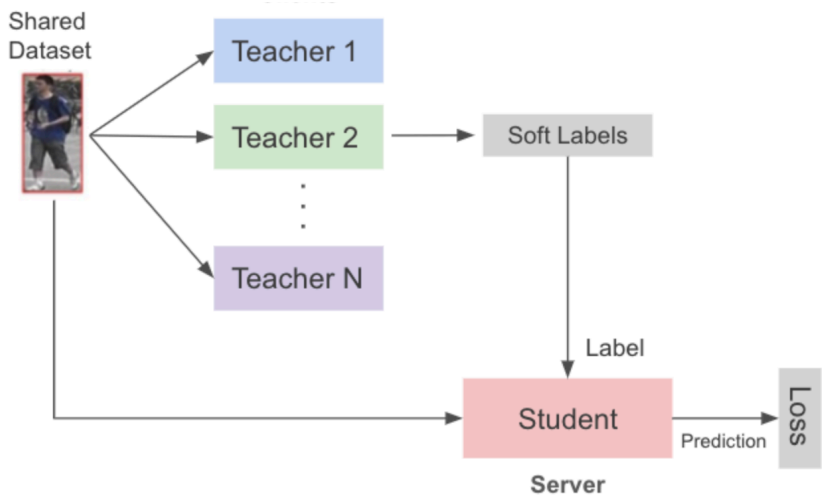
- Meta Pseudo Labels (MPL)
  - Experimented with Pseudo Labels and Meta Pseudo Labels for semi-supervised learning
  - MPL achieves state-of-the-art performance by addressing the limitations of standard PL
    - Mitigation of inaccurate label propagation and confirmation bias
- The teacher and student models are trained in parallel
- Student learns from pseudo labels produced by the teacher
- Teacher learns from labeled data + the performance improvement feedback





Reference: Pham, Hieu, et al. "Meta pseudo labels." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

- Knowledge Distillation
  - An ensemble of two teacher models used to distill their combined knowledge into a student model
  - Teachers are trained using MPL
  - Student can outperforms teachers given sufficient student model capacity and large volume of unlabelled data
  - Adds to model training diversity



Source: https://weimingwill.medium.com/federated-learning-person-re-identification-benchmark-in-depth-analysis-and-performance-f207d3d8791a

Training

- Model Diversity
  - DeBERTa Large
  - DeBERTa XLarge
  - DeBERTa V2 XLarge
  - DeBERTa V3 Large

- Training Diversity
  - Soft Meta Pseudo Labels
  - Hard Meta Pseudo Labels
  - SWA
  - Knowledge Distillation
  - Marked Tokens
    - Prefixing a token e.g. “[QA CASE=0]” to feature text
  - Standard fine-tuning

- Additional Details
  - In MPL both Teacher and Student models need to be loaded into memory
    - Resolving memory overflow issues: mixed precision training, 8-bit Adam optimizer, smaller batch size, freezing of lower layers & gradient checkpointing
  - For token classification head, concatenation of hidden outputs from last 12 transformer layers
  - Re-initialization of top transformer layers
  - LR Scheduler: cosine schedule with warmup

Inference

Inference

- Ensembling
  - Inference from a total of 18 model checkpoints
  - Each model predicts the probability of a character to be included in the answer span given a feature text
  - Simple blending: weighted average of character wise predictions for ensembling
  - Equal weight for all models
  - A threshold of 0.5 is used for all cases
- Post Processing
  - Minor post-processing rules are applied based on error analysis
  - Filtering of certain keywords from predicted spans
    - Feature 309: duration-2-months: filter out spans containing ‘2 weeks’, ‘2weeks’ in it

Important Findings

Important Findings

- The effectiveness of task adaptation and pseudo labelling
  - Reinforces the importance of releasing a large pool of unlabelled task specific data to aid model adaptation through pre-training and leveraging semi-supervised technique
  - Particularly useful in low resource settings where the volume annotated data is limited
- Multi-task learning
  - Design of relevant auxiliary tasks which helps training of the main task
  - Predicting whether a token/character is a part of answer span is the main task, whereas predicting beginning and end of answer spans are the auxiliary tasks
- Incorporation of regularization techniques such as Stochastic Weight Averaging (SWA) leads to better generalization
- Sub-optimal performance of Weighted Box Fusion (WBF) and Non Maximum Suppression (NMS) as ensembling strategy

Question and Answer


kaggle