Improving MetaDistil Framework

Abhinav Lugun - st122322

Anh Nguyen - st122910

Mst Maria Rahaman Rumki - st123835

Outline

- Introduction
- Objectives
- Background
- Teacher-Student Architecture Choice
- Methodology
- Experiment
- Result

Introduction

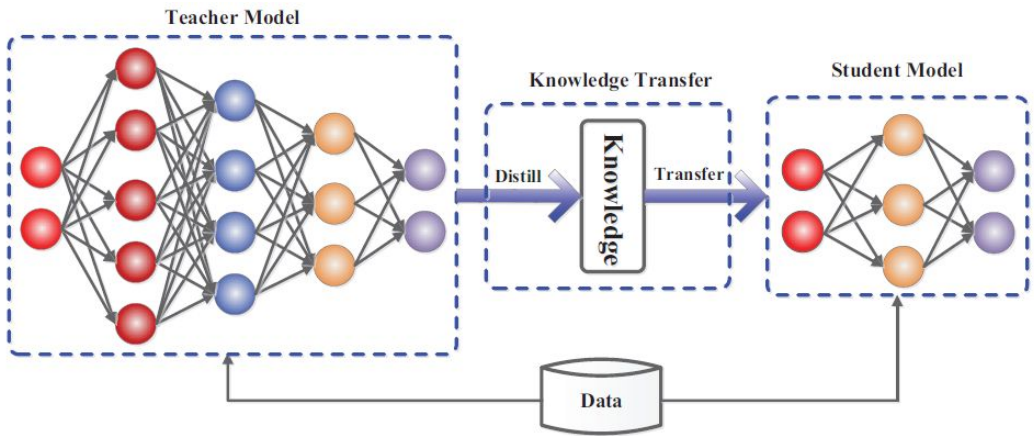
- Pre-trained models (PLMS), such as Bert and Roberta, have gained success in many NLP tasks
- However, it is difficult to deploy them due to their huge model size and high computations
- To address this, model compression is gaining prominence
- Among them , knowledge distillation is attention due to its proven effectiveness in both computer vision and natural language processing tasks

Objectives

- The goal of the work is to **improve** the MetaDistil framework for knowledge distillation in NLP.
- The improvement involves
  - **training the student model partially** before the distillation process
  - **Incorporating more teacher**
  - and **injecting noise**
- Partially training the student model beforehand helps it better learn the knowledge it will be learning.
- Introducing noise help model be more robust to dataset noises.
- These improvements are expected to enhance the performance of the student model in NLP tasks, specifically summarization.
- The related work section explores different ways researchers have attempted to improve knowledge distillation in NLP and enhance model generalization and adaptation to new tasks using meta-learning.

Quick background: Knowledge Distillation

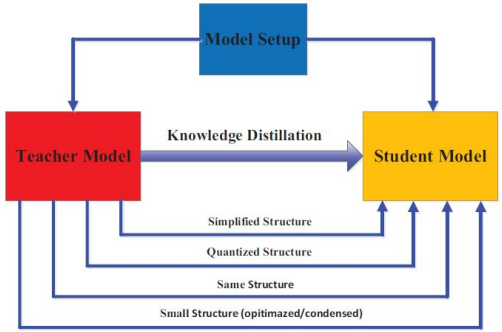
- Refers to the process to transfer knowledge from a teacher (large model) to a student (small model)
- Achieved by training student model to mimic teacher model's soft labels
  - Idea: soft labels contain more information
- Later works explored using intermediate layers and relations as sources of knowledge



Source:
https://arxiv.org/abs/2006.05525

Teacher-Student Architecture Choice

- Designing Teacher-student architecture crucial for efficient knowledge transfer
- Large model with highest accuracy may not be the best teacher for a student model
- Common architectures of the student model
  - Shallower version of the teacher model (fewer layers and neurons per layer)
  - Quantized version of the teacher model
  - Smaller network with efficient basic operations
  - Smaller networks with the optimized global network architecture
  - Same model as the teacher



Source:
https://arxiv.org/abs/2006.05525

Addressed Challenges

- Addressed drawback
  - Teacher is unaware of student's capacities
    - Normally student accepts knowledge passively
  - Teacher not optimized for distillation
    - Trained to optimizer on its own inferences
    - Knowledge transfer done sub-optimally
    - PhD analogy
- Proposed framework → MetaDistill
  - Knowledge distillation with meta learning
  - Mainly addressed by enabling teacher model to receive feedback on student model's learning progress
  - This would allow the teacher model to improve its "teaching skills"

Methodology

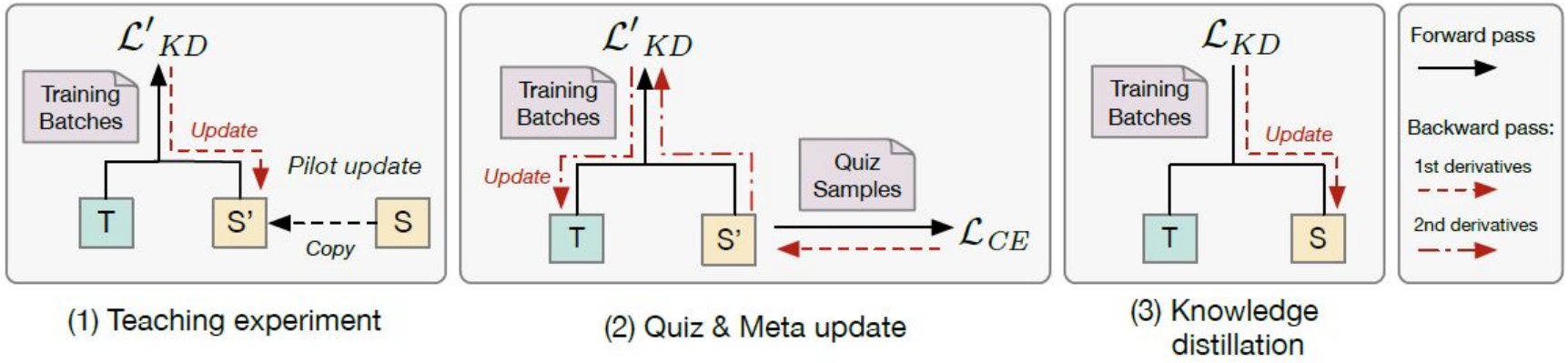


Fig: Working of MetaDistil Framework

Methodology

Repeat:

1: Sample batch of training data
2: Create student backup // Time Machine
3: Perform knowledge distillation to update student model
4: Sample batches of quiz data
5: Calculate loss of student models on quiz data
6: Update teacher model according to the student loss
7: Restore initial parameters of student model from backup
8: Perform knowledge distillation to update student model

Until: Convergence or end of epoch

Fig: Sudo Code

Methodology

Additional Teacher

- In addition to learning from on teacher, student model will from another teacher at the same time
- Multiple teachers offer different perspectives
- Incorporate multiple teacher predictions by averaging their output logits

1 Epoch Fine-Tuning

- Student will absorb information before if he/she goes through the class material before lecture

Noise Injection

- Noise Injection is a technique used in Natural Language Processing to introduce noise into sentences or text.
- The purpose of Noise Injection is to generate more diverse data and help machine learning models understand variations in real-world language.
- Example: add_noise("This is a sentence.", p=0.2)
- Result: "This is a noise_word."
- In this example, the word "sentence" is replaced with "noise_word" because the randomly generated value is less than the probability p=0.2.

Experiment

- **Objective:** Test performance on sentiment analysis task
- **Dataset:** SST-2
- **Metric:** accuracy
- **Teacher Models:** BERT + RoBERTa
- **Student Models:** DistilBERT + TinyBERT
- **Epoch:** 15
- **Baselines:**
  - DistilBERT and TinyBERT which are directly fine-tuned
  - Normal knowledge distillation

Experiment

Knowledge Distillation and MetaDistil Framework

- **Train -** 30,000
- **Quiz -** 100
- **Validation -** 872
- **Test -** 2,000

Mult-Teacher

- **Teacher - (**Bert + RoBERTa) and (BERT + BERT)
- **Student -** DistilBERT

Result

Methods	Accuracy	No. of Parameters	Times Smaller
Teacher			
Bert	93.8	109483778	
Roberta	92.95	124647170	
Fine-Tuned Models			
Distilbert	92.65	66955010	1.7
Tinybert	89.95	14350874	8.2
Student Model with KD			
Bert Distilbert	93.1	66955010	1.7
Bert Tinybert	91.65	14350874	8.2
Roberta Distilbert	81.5	66955010	1.7
Student Model with MetaDistil			
Bert Distilbert	93.1	66955010	1.7
Bert Tinybert	90.3	14350874	8.2
Roberta Distilbert	89.3	66955010	1.7
Student Model with MetaDistil + Additional Teacher			
Bert + Roberta Distilbert	92.85	66955010	1.7
Bert + Bert Distilbert	93.55	66955010	1.7

Table 1: Results Baseline and Additional Teacher Comparison

Result

Methods	Accuracy	No. of Parameters	Times Smaller
Bert Distilbert	93.1	66955010	1.7
Bert Distilbert (ft for 1 epoch)	92.45	66955010	1.7

Table 2: MetaDistil and MetaDistil with 1 epoch fine-tuning beforehand

Result

Methods	Accuracy	No. of Parameters	Times Smaller
Bert Tinybert	90.3	14350874	8.2
Roberta Distilbert	89.3	66955010	1.7
Multi-Teacher Bert + Bert Meta Distil + noise	88.55	66955010	1.7
Multi-Teacher Bert + Roberta Meta Distil + noise	88.5	66955010	1.7

Table 3: Comparison of two model with noise and and vice versa.

Discussion

- MetaDistil did not do better than KD as hoped in some areas
  - there was huge improvement over using teacher as RoBERTA
  - Performed worse for student model TinyBERT
- DistilBERT trained with BERT + BERT teacher performed the best
- Adding noise and fine-tuning student model before distillation process did not improve as hoped

Conclusion

- Having mult-teacher has great potential to improve student accuracy
- Teacher-student choice matters
- Student capacity can affect knowledge transfer efficacy

Future Work

- Try teacher assistants
- Mutual student learning
- Incorporate pruning
- Add knowledge from intermediate layers
- Test on more NLP tasks

Demo

Improving MetaDistil Framework Demo

Entert Text

NLP is fun.

PREDICT SENTIMENT

Model	Prediction
Bert Model	Positive
BERT DistilBERT MetaDistil	Positive
BERT + BERT DistilBERT MetaDistil	Positive

Thank you