MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

Improving baseline performance on the TRIP model

EECS 595

Wenfei Tang
Juejue Wang

Background

- **Problem**
  - large-scale language models lack verifiable reasoning despite having high accuracy on the end task
- **TRIP dataset**
  - story plausibility classification (end task)
  - includes dense annotations for capturing multi-tiered of reasoning
  - emphasizes language model's verifiable physical commonsense reasoning ability

Project goal

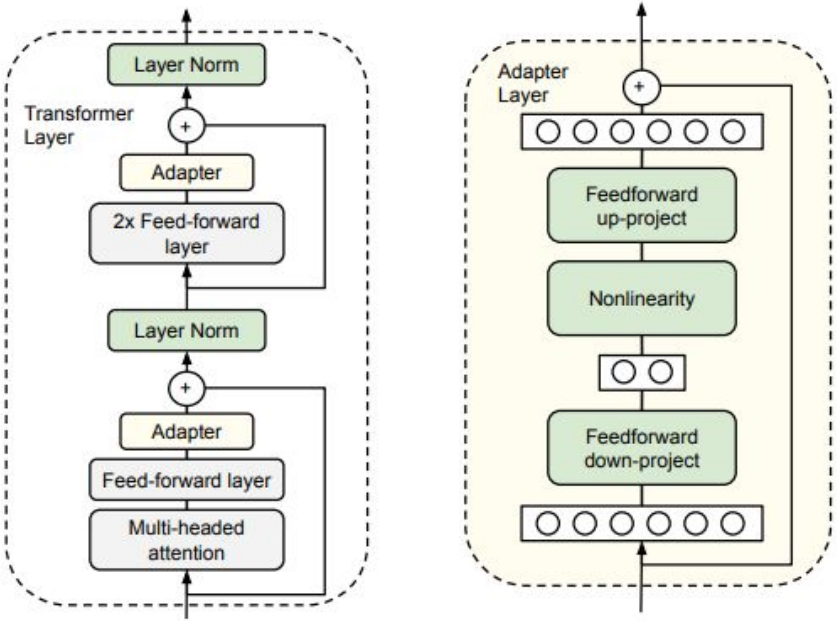
- Perform **architecture modifications** and **optimization schedules** to improve the performance baseline of the TRIP model

Method and novelty

- **Transfer Learning:**
  - Use pre-trained models from relevant dataset and apply them to TRIP
  - Optimizing schedules of the TRIP model
  - Focus on improving the model's tiered-reasoning ability rather than the end task accuracy
- **Adapters:**
  - a set of weights within the layers of a transformer model
  - an alternative to fully fine-tuning the model
  - require small storage space per task!

Source: https://adapterhub.ml/

More about adapters



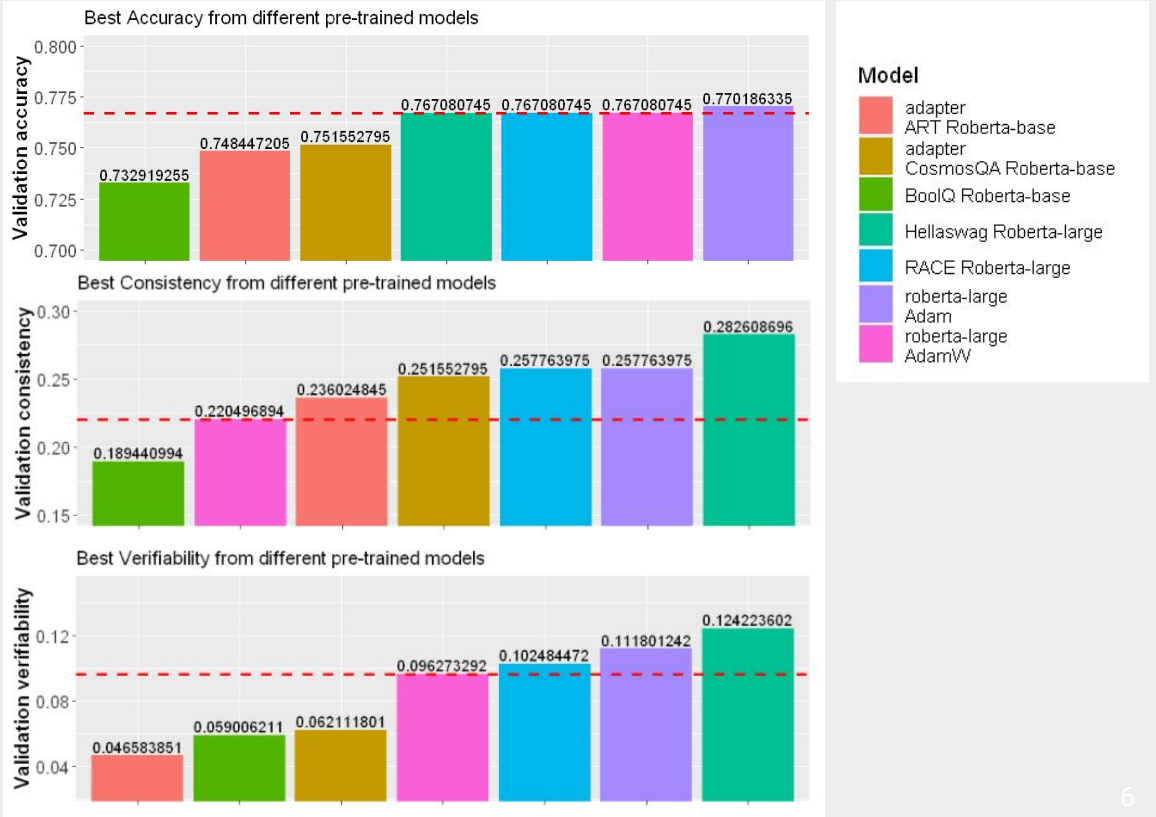
- Architecture of the adapter module and its integration with the Transformer
- During adapter tuning, the green layers are trained on the downstream data

Source: https://arxiv.org/pdf/1902.00751.pdf

Preliminary Results

Metrics on the validation set of TRIP

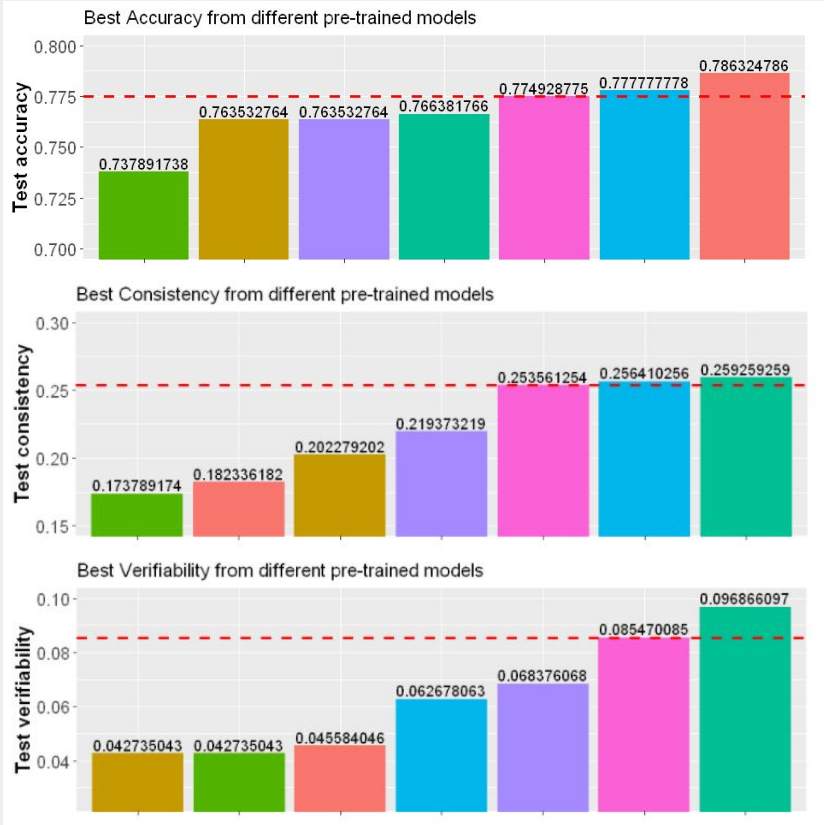
- Higher verifiability and consistency on validation datasets in general



Preliminary Results

Metrics on the test set of TRIP

- Highest verifiability/consistency from Hellaswag
- Highest accuracy from ART adapter



Work so far

- Improved TRIP baseline performance
  - Applied adapters for efficient task transfer (roberta-base)
    - **CosmosQA**, **ART**
  - Used the pre-trained roberta model on state-of-arts relevant datasets
    - PIQA, **Hellaswag**, GPT-neo, **Aristo**, **BoolQ**, **Winogrande**, **Race**
  - Transferred the pre-trained models to TRIP model
  - Experimented on optimizing schedules
    (tested on optimizers: **Adam, AdamW** and **SGD**)
- Refined codebase
  - Made the code compatible to train a gpt and roberta-base model
  - Added code to allow training a model on PIQA, which can later be used on TRIP
  - Fixed some bug in dataset preprocessing code

Future Plans

- Use a pre-trained GPT model to train TRIP (code ready)
- Train a roberta-large model on ART and transfer it to TRIP (given its good performance on roberta-base model)
- Train a roberta-large model on PIQA and transfer it to TRIP (might be limited to computing power)

Reference

Storks, Shane, et al. "Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding." *arXiv preprint arXiv:2109.04947* (2021).

Ruder, Sebastian, et al. "Transfer learning in natural language processing." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials. 2019.

Pruksachatkun, Yada, et al. "Intermediate-task transfer learning with pretrained models for natural language understanding: When and why does it work?." arXiv preprint arXiv:2005.00628 (2020).

Poth, Clifton, et al. "What to Pre-Train on? Efficient Intermediate Task Selection." arXiv preprint arXiv:2104.08247 (2021).

Houlsby, Neil, et al. "Parameter-efficient transfer learning for NLP." International Conference on Machine Learning. PMLR, 2019.

Audience Questions

Q1: Are you pretraining on multiple datasets before fine-tuning or just one?

A: We use pre-trained models from multiple datasets. For each transfer learning process, we apply a pre-trained model from one dataset to the TRIP dataset, except Aristo Roberta-large fine tuned on RACE.

Q2: What is the significance of verifiability and consistency metrics for the model?

A: We use the metrics from the TRIP paper, where consistency indicates whether pairs of conflicting sentences have been correctly identified, and verifiability indicates whether the underlying physical states that contribute to the conflict are correctly identified. Using these two metrics allow us to measure the tiered reasoning ability of the language model.

Audience Questions cont.

Q3: Since your team is using two different datasets for two approaches, how will you compare your model performances be due to the difference in dataset?

The different datasets will only affect the pre-trained model used on TRIP model. The final evaluation will still be performed on the TRIP dataset with accuracy, verifiability and consistency.

Q4: You mentioned that you used pre-trained models in relevant datasets and could you please tell us what datasets did you choose and why did you use them?

We use pre-trained models from the following datasets: cosmosQA, BoolQ, Hellaswag, Aristo, Race, ART, and PIQA. We chose these datasets based on their similarity with TRIP. All of the above models are question-answering datasets and measure the model's reasoning ability.

MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN