Yimeng Wu

Toronto, ON, Canada
+1(438)-926-3522 yimengwu71@gmail.com linkedin.com/in/yimeng-wu-032878126 github.com/yimeng0701

Education

- **McGill University** — Montreal, QC, Canada
  Master of Engineering in Electrical and Computer Engineering; GPA: 3.77/4.0 — *Sept. 2017 - May 2019*
- **Tianjin University** — Tianjin, China
  Bachelor of Engineering in Biomedical Engineering; GPA: 3.48/4.0 — *Sept. 2013 – July 2017*

Work Experience

- **Huawei Noah's Ark Lab Canada** — Toronto, ON
  Researcher: working on 11B T5 Arabic LM — *June 2022 – present*
- **Huawei Noah's Ark Lab Canada** — Toronto, ON
  Associate Researcher: mainly worked on Arabic NLU projects — *June 2021 – Jun 2022*
- **Huawei Noah's Ark Lab Canada** — Montreal, QC
  Support Researcher: researches on knowledge distillation and translation projects — *Dec. 2019 – June 2021*

Research

- **Universal Knowledge Distillation** *Jan. 2021 - June 2021*
  - Proposed a more general intermediate KD method, which is matching intermediate layers of the teacher and the student in the output space via the attention-based layer projection, instead of applying matching in the hidden spaces. (accepted by EMNLP 2021)
  - Applied Universal-KD on 3 scenarios: intermediate layer knowledge distillation, capacity gap problems and cross-architecture distillation (BERT to Gated CNN, BERT to LSTM).
- **Intermediate Layer Knowledge Distillation** *Dec. 2019 - Aug. 2020*
  - Proposed a combinatorial-layer knowledge distillation method to distill from intermediate layers and final predictions. Also, it is the first time that the combination concept is defined for KD approaches. (accepted by EMNLP 2020)
    * Previous intermediate layer KD methods such as PKD only measure the similarity of the hidden outputs of the student layers and chosen teacher layers. However, multiple layers of the teacher might be ignored in the distillation process which can lead to loss of information.
    * In our method, the teacher layers in each bucket are concatenated together and the MSE is calculated between the hidden outputs of each student layer and the combined teacher layer. It shows promising results in multiple translation tasks.
  - Proposed an automatic layer combination technique in intermediate KD which relies on attention. (accepted by AAAI 2020)
    * For any student layer there is a weighted average of all teacher layers to distill into the student layer. The weights assigned to teacher layers are determined by the dot product and the attention mask is spanned over all teacher layers.
    * The performances of both 4-layer and 6-layer BERT are improved on GLUE tasks and outperform other existing students with the same size by at least 0.5 on average.

Projects

- **Arabic NLU** *June 2021 - Present*
  - Pretrained an Arabic BERT from scratch with 115G textual data for product-line use.
  - The model got Rank 1 on a pubic Arabic leaderboard (ALUE), which contains eight NLP/NLU tasks and outperforms the second place by 6% on average score.
  - Applied the model on multiple internal Arabic NLP/NLU tasks, including punctuation detection and question similarity matching. The model outperforms the stat-of-the-art pretrained Arabic models, such as AraBERT and MARBERT, by at least 2%.
  - Now working on the pertaining of the T5-11B Arabic language model.

- **Robust Machine Translation** *Sep. 2020 - Dec. 2020*
  - Trained separate En→Zh and Zh→En on-cloud, as well as bidirectional on-device transformer models that are robust to OCR noise.
  - Analyzed OCR error types inside the training data and decided the noise generation rules. The training set was augmented with extra 80 million noise data, which are generated based on the rules.
  - Got improvement by an average of 4 bleu score and 0.6 bleu score for internal En→Zh and Zh→En OCR test sets, respectively, without degradation on inference speed or performance on Common test sets.

Publications

- Abbas Ghaddar*, **Yimeng Wu***, Sunyam Bagga, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing et al. "Revisiting Pre-trained Language Models and their Evaluation for Arabic Natural Language Understanding." To appear in **EMNLP** 2022.
- **Yimeng Wu**, Mehdi Rezagholizadeh, Abbas Ghaddar, Md Akmal Haidar, and Ali Ghodsi. "Universal-KD: Attention-based Output-Grounded Intermediate Layer Knowledge Distillation." In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (**EMNLP** 2021, **Oral**, acceptance rate: 25.6%)
- Passban Peyman, **Yimeng Wu**, Mehdi Rezagholizadeh, and Qun Liu. "ALP-KD: Attention-Based Layer Projection for Knowledge Distillation." In Proceedings of the AAAI Conference on Artificial Intelligence (**AAAI** 2021, acceptance rate: 21%)
- **Yimeng Wu***, Peyman Passban*, Mehdi Rezagholizadeh, and Qun Liu. "Why Skip If You Can Combine: A Simple Knowledge Distillation Technique for Intermediate Layers." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (**EMNLP** 2020, short, acceptance rate: 16.7%)
- Abbas Ghaddar, **Yimeng Wu**, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang et al. "JABER: Junior Arabic BERt." preprint (2022).

Awards

- Huawei Noah's Ark Lab Commercial Contribution Award, 2022

Skills

- Programming Language: Python, Java, Matlab
- ML Frameworks: Pytorch, Tensorflow, HuggingFace, Scikit-Learn, Pandas, Matplotlib, CNTK
- ML Areas: Natural language processing (NLP), Computer Vision, Deep Learning, Medical Image & Signal Processing