Tong Zheng

Northeastern University, Heping District, Shenyang 110819

✉: zhengtong12356@gmail.com | ☺: kidzheng.github.io | ☎: +86 18602457703

Education

Northeastern University   **Shengyang, CN**
Major: Computer Science and Technology   Aug. 2017 - July. 2021
Cumulative GPA: 4.001/5 (90.001/100)
Relevant Courses:

- Artificial Intelligence (98), Introduction of Machine Learning (95), Intelligent Computing System (100), Field Practice (95), Introduction of Bioinformatics (95), Information Processing and Machine Translation (95).
- C Programming (95), C++ Programming (97).
- Advanced Mathematics (97), Linear A1gebra (95), Numerical Analysis (97), Probability Theory and Mathematical Statistic (97), Discrete Mathematics (92).

Research Interests

- Efficient AI; Large Language Model and Generative AI: from Application to Acceleration; Machine Translation & Multi-modal Translation; Computer-aided Diagnosis.

Hint: Upon a thorough review of your research interests, I am confident that my profound knowledge of foundational architectures could prove particularly valuable in domains like parameter-efficient fine-tuning.

Publications (* denotes Equal Contribution)

1. Bei Li*, **Tong Zheng***, Yi Jing*, Chengbo Jiao, Tong Xiao, Jingbo Zhu. Learning Multiscale Transformer Models for Sequence Generation. International Conference on Machine Learning. 2022. (First ICML in NEU, China.)
   - This work re-defined the concept of scale for NLP, including scales of sub-word, word and phrase. The intention was to leverage the word boundaries and phrase-level prior knowledge to compensate for the sub-word features. Then this study established the relationships among different scales. Ultimately, this study built a multiscale Transformer model via making full use of the relationships.

2. Yuxin Zuo*, Bei Li*, Chuanhao Lv, **Tong Zheng**, Tong Xiao, JingBo Zhu. Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs. Findings of EMNLP2023.
   - This study proposes a MMT-VQA Multi-Task Learning Framework to explicitly model the probing signal in MMT to strengthen the interactions between text and visual modalities. Also, this work releases a Multi30k-VQA dataset. Experimental results on Multi30K En-De and En-Fr demonstrate the effectiveness both in terms of BLEU and specific testsets.

3. Guangqi Wen, Peng Cao, Huiwen Bao, Wenju Yang, **Tong Zheng**, Osmar Zaiane. MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis. Computers in Biology and Medicine. 2022. (IF=7.7)
   - This study introduces MVS-GCN, combining multi-view prior brain structure-guided graph learning and multi-task graph embedding to tackle subject heterogeneity and noise correlations in brain networks. The results show improved classification and alignment with ASD biomarkers.

4. **Tong Zheng***, Bei Li*, Huiwen Bao*, Yi Jing, Tong Xiao, JingBo Zhu. EIT: Enhanced Interactive Transformer. Arxiv. (Submitted to EMNLP2023, Soundness: 3443, Excitment: 4344).
   - This study introduced the EIT framework, which considers both diversity and consensus among attention heads via a many-to-many mapping module and a fine-grained hierarchical interaction module. It addresses the issue in existing attention methods that prioritize diversity among heads at the expense of consensus. Extensive experiments on five tasks confirmed EIT's effectiveness. Also, this study revealed such consensus can enable model to tolerate more ratio when pruning heads.

5. **Tong Zheng***, Huiwen Bao*, Bei Li*, Weiqiao Shan, Tong Xiao, JingBo Zhu. PartialFormer: Modeling Part Instead of Whole for Machine Translation. Arxiv. (Submitted to EMNLP2023, Soundness: 343, Excitment: 333).
   - This study introduced PartialFormer, which deploys smaller FFNs within multi-head attention to reduce parameters while maintaining high performance. It investigated its scalability and introduced a tailored head scaling strategy. Impressively, it achieved 29.56 BLEU points with only 68M parameters on the WMT'14 En-De task, providing valuable insights into designing lightweight FFNs.

6. **Tong Zheng**, Bei Li, Huiwen Bao, Jiale Wang, Can Zhao, Weiqiao Shan, Tong Xiao, JingBo Zhu.Bridging the Gap between NMT and LLM: A Prompting Approach for Integrating NMT Knowledge into LLM.
   - This study introduces a prompting framework that employs a multi-scale alignment format to efficiently integrate NMT prior knowledge into LLMs, improving LLM fidelity. It achieved significant improvements on 8 WMT22 general translation tasks and 4 low-resource translation tasks in both sacreBLEU and COMET-22.

Work & Research Experience

Natural Language Processing Laboratory at NEU **Shengyang, CN**
Research Assistant (AI, NLP) *Seq. 2021 - Now*

- I led and participated in several research projects.
  - † **Project 1:** Learning Multiscale Transformer Models for Sequence Generation. (ICML2022)
    - ○ **Designed and implemented transformation matrices** to facilitate cross-scale information flow.
    - ○ **Developed a multi-scale feature extraction and fusion framework** utilizing GCNs and a novel self-attention mechanism, enhancing information utilization and mitigating data loss.
    - ○ **Built the entire framework using Fairseq.**
    - ○ **Managed machine translation data collection and conducted experiments**, including main results and ablation studies.
    - ○ **Contributed to the paper**, particularly the Method section, and created all figures using Tikz.
  - † **Project 2:** EIT: Enhanced Interactive Transformer. (Arxiv, Under Review)
    - ○ **Proposed and Led the project; Identified the problem:** *the current design of Multi-head self-attention, an instance of multi-view learning, prioritizes the complementarity while ignoring the consensus.*
    - ○ **Designed and implemented** a EIT framework, that utilized a many-to-many mapping strategy generate more attention maps and employed a fine-grained hierarchical interaction module to enhance consensus among attention heads, guided by multi-view learning theory.
    - ○ **Built the entire framework using Fairseq.**
    - ○ **Collected data of WMT'14 En-De, WMT'16 En-Ro, CNN-DailyMail, WikiText-103, ABIDE, CONLL14 and conducted all experiments**.
    - ○ **Contributed to all parts of the paper**, and created all figures using Tikz.
  - † **Project 3:** PartialFormer: Modeling Part Instead of Whole for Machine Translation. (Arxiv, Under Review)
    - ○ **Proposed and Led the project; Identified the problem:** *the current design of Lightweight FFNs ignores the importance of hidden dimension.*
    - ○ **Designed and implemented** a PartialFormer: 1) PG-FFNs that consists of multiple smaller FFNs to cut down parameters and computations whiling maintaining large hidden dimension; 2) Inserting them into multi-head attention to maintain high perofrmance; 3) a residual attention calculation method to enable stable optimization.
    - ○ **Investigated** the scalability of PartialFormer and **designed and implemented** a head scaling strategy tailored for PartialFormer: directly adding more heads and widening heads.
    - ○ **Built the entire framework using Fairseq.**
    - ○ **Collected data of WMT'14 En-De, WMT'14 En-Fr, WMT'16 En-Ro, WMT'17 Benchmark, CNN-DailyMail, WikiText-103, and conducted all experiments**.
    - ○ **Contributed to all parts of the paper**, and created all figures using Tikz.
  - † **Project 4:** Bridging the Gap between NMT and LLM: A Prompting Approach for Integrating NMT Knowledge into LLM. (Arxiv, Under Review)
    - ○ **Proposed and Led the project; Recognized the issue:** *existing Language Model-based translation approaches lack high fidelity, as indicated by their low BLEU scores.*
    - ○ **Designed and implemented** a prompting framework that automatically extracts a multi-scale alignment format and selectively integrate NMT prior knowledge into LLMs.
    - ○ **Built the entire framework using Fairseq.**
    - ○ **Collected data of WMT'22 Benchmark, and conducted all experiments**.
    - ○ **Contributed to all parts of the paper**; Designed and produced all the figures using Tikz for the paper.
  - † **Project 5:** Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs. (Findings of EMNLP2023)
    - ○
- Mentored interns in deep learning and prompt engineering for successful completion of **project 4**.
- Assisted teachers to complete the teaching stuffs, and helped with conference-related work including the CCMT2022 report on Neural Network Design and Learning in Natural Language Processing.
- Contributed to editing the book 'Natural Language Processing: Representation Learning and Neural Models.'
- Worked as secondary reviewers of several conferences/journals: ICML 2022, NeurIPS 2022, NLPCC 2023, TALLIP (journal), TACL (journal), Pattern Recognition (journal).

Key Laboratory of Medical Image Intelligent Computing at NEU **Shengyang, CN**
Research Assistant (AI) *Oct. 2020 - Aug. 2021*

- I led and participated in several research projects.
  - † **Project 1:** BrainTGL:Temporal Graph representation learning for brain network by Exploiting Graph Temporal Information. *Oct. 2020 - Aug. 2021*

- This project, in partnership with Southern Medical University Hospital, is now used in the medical system.
- This study introduces the BrainTGL framework, which integrates GCN and LSTM to address spatio-temporal dynamic pattern challenges in resting-state fMRI data. It achieves state-of-the-art accuracy on the ABIDE and HCP datasets, provides consistent clustering results with ASD biomarker evidence.
- **My Contributions**: 1) I led this project. I found this problem and came up the solution. 2) I wrote all the codes in this project. 3) I conducted all the experiments in this paper. 4) I wrote the paper and drew all Figures in this paper. (submitted to JBHI, PR and CIBM. The manuscript can be found here.)
- This project also served as **my undergraduate thesis**.
  - † **Project 2:** MVS-GCN: A prior brain structure learning-guided multi-view graph convolution network for autism spectrum disorder diagnosis (CIBM2022)   *Apr. 2021 - Aug. 2021*
    - Developed codebase: Adapted and extended code from the BrainTGL project.
    - Implemented graph clustering module: Utilized PyTorch for coding the graph clustering component.
    - Proposed the idea of consistency regularization to enhance alignment among different views.
    - Designed and produced all the figures for the paper.

Shenyang Sulianda Technology Co., Ltd.   **Shengyang, CN**
AI Intern   *Jun. 29, 2020 - Jul. 29, 2020*

- Completed the implementation of basic machine learning algorithms (logistic regression, K-means, SVM algorithm, neural network, CNN, GAN, Bayes algorithm, text hierarchical clustering algorithm) and solved some practical problems perfectly.
- Accomplished the work of target detection and recognition based on MTCNN algorithm.
- Served as my **Undergraduate Field Practice course** (Rank: 1)

Skills & Self-Evaluation

- Good at **paper writing** and **figure drawing** in **Latex**.
- **Programming**: Python, LaTeX, C, C++, Java
- **Tools**: PyTorch, Fairseq, Tikz, Numpy, etal.
- **English Proficiency**: TOEFL iBT total 87 (*Reading:* 27, *Listening:* 19, *Speaking:* 19, *Writing:* 22)
- **Persistent**, **passionate**, **adaptable**, and have a certain taste for academic pursuit; Good at **team working** and **getting along well** with others.

Projects & Competitions

Assignment for Machine Learning Course (Rank: 1 in this Course)   **Shengyang, CN**
Assignment for Introduction of Machine Learning   *Nov. 2020 - Dec. 2020*

- Chose computer-aided autism diagnosis as the topic.
- Designed and implemented a novel unified framework integrating node attention, edge attention, and GCNs.
- Collected ABIDE fMRI time series data using the ni-learn toolkit.
- Implemented the pipeline using PyTorch and re-implemented some baselines with PyTorch.
- Employed a 10-fold cross-validation to validate the effectiveness of the proposed framework.

BANGC Programming Practice (Rank: 1, outstanding student award)   **Shengyang, CN**
Practice of Intelligent Computing System   *Nov. 2020 - Dec. 2020*

- Implementing BANGC power difference operation and integrating it with TensorFlow; BANGC operation quantization and offline inference.

Chinese College Student Computer Game Competition (3rd Prize)   **Chongqin, CN**
Hosted by Chinese Society for Artificial Intelligence   *Jul. 2020 - Aug. 2020*

- Modified an open-source Surakat Chess AI algorithm based on MCTS to use the PVS algorithm.

National Training Programs of Innovation and Entrepreneurship for Undergraduates (National Level)
Shengyang, CN
Undergrad Competition   *Aug. 2021 - Jul. 2022*

- This project utilized convolutional neural networks and U-Net algorithms, for intelligent chest CT image analysis in COVID-19 patients, aiding in infection detection and nucleic acid testing.
- Provided technical guidance for this project: debugging codes and tuning the hyper-parameters.

Scholarships & Awards

- The Second Award Scholarship of Northeastern University Excellent Student, 2017-2018.
- Northeastern University Chuanglian Industrial Scholarship, 2018-2019.
- The Third Award Scholarship of Northeastern University Excellent Student, 2018-2019.
- The Third Award Scholarship of Northeastern University Excellent Student, 2019-2020.
- The Second Award Scholarship of Northeastern University Excellent Student, 2020-2021.
- The Third Award of 2020 "Competitive World Cup" Chinese College Student Computer Gaming Competition and

the 14th China Computer Gaming Championship, Chinese Society of Artificial Intelligence and the Ministry of Education higher education computer teaching Steering Committee, in 2020.
- Outstanding student award of 2020 Intelligent Computing Systems Course, Institute of Computing Technology, Chinese Academy of Sciences, 2021.
- Cambrian best developers, Cambrian Industry, 2021.

Extracurricular Activities

- School Sports Meeting Opening Dance Participant.
- School Sports Meeting Tug of War Participant.
- Participated in College Soccer Competition.
- Returning to Promote Northeastern University at My Alma Mater.

Conferences Attended

- The Thirty-ninth International Conference on Machine Learning. (remote)Baltimore. 2022.
- CIPS ATT Issue 37&38. Beijing. 2023.