Tong Zheng

Northeastern University, Heping District, Shenyang 110819
✉: zhengtong12356@gmail.com | ☎: +86 18602457703

Education

Northeastern University (985) **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), Information Processing and Machine Translation (95), Advanced Mathematics (97), Linear A1gebra (95), Numerical Analysis (97), Probability Theory and Mathematical Statistic (97), Discrete Mathematics (92).

Research Interests

My current focuses are as follows:

- **Efficient & Foundation AI**: solving inefficiency in Transformers including 1) parameter-inefficiency, 2) interaction inefficiency in multi-head attention, and 3) inefficiency in single-scale Transformer modeling.
- **Large Language Model and Generative AI**: from Application to Acceleration.
- **Sequence Generation**, especially Machine Translation; **Computer-aided Diagnosis**.

In the future, I plan to do research with following aspects (I am also glad to do some other important topics):

- **Model Pruning & Inference Acceleration**: Investigating techniques to cut down redundant parameters and speed up the inference phase of large models, making real-time applications more feasible.
- **Parameter-Efficient Fine-tuning**: Exploring strategies and methodologies to fine-tune large models more efficiently, ensuring rapid adaptability to new tasks.
- **GCN-driven Prior Knowledge Infusion**: Leveraging Graph Convolutional Networks (GCN) as the primary mechanism to infuse various prior knowledge into large models, aiming to enhance their capabilities.

Publications (* denotes Equal Contribution)

1. Bei Li*, **Tong Zheng***, Yi Jing*, Chengbo Jiao, Tong Xiao, Jingbo Zhu. **Learning Multiscale Transformer Models for Sequence Generation**. ICML. 2022. (first ICML in NEUNLP)

   - Redefined NLP scales to encompass sub-word, word, and phrase; Utilized word boundaries and phrase-level knowledge to enhance sub-word features, culminating in a multiscale Transformer model. Demonstrated effectiveness on both MT and Summarization tasks and improved interpretability of attention maps. **Link:** `https://arxiv.org/abs/2206.09337`.

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 EMNLP, 2023.

   - Proposed an MMT-VQA framework to intensify text-visual interactions using probing signals; Released the benchmarking Multi30k-VQA dataset. Demonstrated effectiveness on Multi30K En-De and En-Fr tasks.

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.

   - Introduced 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. Demonstrated improved classification and alignment with ASD biomarkers. **Link:** `https://www.sciencedirect.com/science/article/abs/pii/S0010482522000312`.

4. **Tong Zheng***, Bei Li*, Huiwen Bao*, Yi Jing, Tong Xiao, JingBo Zhu. **Eit: Enhanced Interactive Transformer**. Preprint on Arxiv. (Submitted to EMNLP2023, Soundness: 3443, Excitment: 4344).

   - Introduced the EIT framework that balances attention head diversity with consensus, using many-to-many mapping and hierarchical interactions. Addressed the traditional attention methods' bias favoring diversity over consensus. Validated EIT's per- formance across five tasks and highlighted its resilience in head pruning. **Link:** `https://arxiv.org/abs/2212.10197`. (This is not the latest version, it will be updated soon)

5. **Tong Zheng***, Huiwen Bao*, Bei Li*, Weiqiao Shan, Tong Xiao, JingBo Zhu. **PartialFormer: Modeling Part Instead of Whole**. Preprint on Arxiv. (Submitted to EMNLP2023, Soundness: 343, Excitment: 333).

   - Introduced PartialFormer, using multiple compact FFNs in multi-head attention for enhanced efficiency, addressing the overlooked importance of the hidden dimension in existing lightweight FFNs; Explored its scala-

bility, devised a unique head scaling approach. Achieved 29.56 BLEU with only 68M parameters on WMT'14 En-De, highlighting advancements in lightweight FFN design. **Link:** `https://arxiv.org/abs/2310.14921`.

Under Review

1. **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.**
   - Introduced a prompting framework that employs a multi-scale alignment format to efficiently integrate NMT prior knowledge into LLMs, improving LLM fidelity. Achieved significant improvements on 8 WMT22 general translation tasks and 4 low-resource translation tasks in both sacreBLEU and COMET-22.
2. Bei Li, Xu Tan, Rui Wang, **Tong Zheng**, Tong Xiao, JingBo Zhu. **Numerical Transformer: Improved Coefficient Learning with Predictor-Corrector Paradigm**. (Submitted to TPAMI)
   - Introduced the Numerical Transformer, leveraging advanced numerical methods to enhance Transformer models. Addressed challenges in coefficient learning for high-order solutions and optimized numerical methods for efficiency. Achieved state-of-the-art BLEU scores on WMT'14 tasks, surpassing previous models in performance and computational efficiency without data augmentation.

Work & Research Experience

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

- Led and participated in several research projects.
  - † **Project 1:** Learning Multiscale Transformer Models for Sequence Generation. (ICML2022)
    - **Purpose:** Tackle text-level semantic neglect in sub-word modeling, improving attention interpretability.
    - **Role & Contribution:**
      - Conceived and designed the general backbone: a) transformation matrices, b) GCN-based multi-scale feature extraction, and c) a unique self-attention approach for fusion.
      - Independently implemented the entire framework via Fairseq; Led all MT and analytical experiments; Contributed to the paper's Method part and crafted all figures via Tikz.
    - **Results**: Boosted MT/Summarization performance with more interpretable attention maps.
  - † **Project 2:** EIT: Enhanced Interactive Transformer. (Arxiv)
    - **Purpose**: Address the issue where the current design of Multi-head self-attention, an instance of multi-view learning, prioritizes complementarity over consensus.
    - **Role & Contribution:**
      - Identified the core issue and pioneered the EIT project; Developed the EIT framework to enhance consensus among attention heads, guided by multi-view learning principles;
      - Implemented the framework using Fairseq; collected data from sources like WMT'14 En-De, WMT'16 En-Ro, CNN-DailyMail, WikiText-103, ABIDE, CONLL14, and spearheaded experiments on them.
      - Authored the entire paper and crafted all figures using Tikz.
    - **Results:** Showcased EIT's excellence across five tasks and its resilience in head pruning.
  - † **Project 3:** PartialFormer: Modeling Part Instead of Whole. (Arxiv)
    - **Purpose:** Address the overlooked importance of hidden dimensions in Lightweight FFNs, introducing a more efficient design.
    - **Role & Contribution:**
      - Pinpointed the above issue and proposed the PartialFormer project;
      - Crafted PartialFormer with multiple smaller FFNs integrated into MHA for modeling small inputs; Introduced residual attention for stability and a unique head scaling strategy for efficiency.
      - Utilized Fairseq for implementation; Collected all data; Conducted all experiments; Authored the entire paper and designed figures with Tikz.
    - **Results:** Scored 29.56 BLEU with 68M parameters on WMT'14 En-De, highlighting FFN innovations.
  - † **Project 4:** Bridging the Gap between NMT and LLM: A Prompting Approach for Integrating NMT Knowledge into LLM. (Under Review)
    - **Propose:** Address the fidelity issue in LM-based translation, evidenced by low BLEU scores.
    - **Role & Contribution:**
      - Identified the core problem, leading to the proposal of this project.

- Developed a prompting framework to extract multi-scale alignment and selectively integrate NMT knowledge into LLMs using Fairseq.
- Gathered data from WMT'22 Benchmark and managed all experiments.
- Authored the entire paper and crafted figures with Tikz.

○ **Results:** Marked improvements on 8 WMT22 general and 4 low-resource translation tasks in both sacreBLEU and COMET-22.

† **Project 5:** Unveiling and Addressing Natural Inefficiencies in Large Language Models. (Work in Progress)

○ **Role & Contribution:**

- Initiated and led the project, identifying the core question regarding computation distribution.
- **Preliminary Results:** 1) Revealed that computation distribution is uneven during inference. 2) Introduced a training-free early exit strategy for 'bigtrans' LLMs, achieving reduced computation and enhanced performance.

† **Project 6:** Incorporating Probing Signals into Multimodal Machine Translation via Visual Question-Answering Pairs. (Findings of EMNLP2023)

○ **Purpose**: Address the inefficient usage of visual modality information in MMT.

○ **Role & Contribution**:

- Collaborated on data analysis, ensuring result reliability.
- Designed a supplementary experiment to critically assess and validate the genuine utilization of the VQA dataset; Proofread hallucinations in the Multi30k-VQA dataset; Refined the manuscript.

○ **Results:** Demonstrated effectiveness on Multi30K En-De and En-Fr tasks.

- Mentored interns in deep learning and prompt engineering for successful completion of **project 4**.
- Assisted faculty with teaching tasks, and contributed to conference-related work including the CCMT2022 report on Neural Network Design and Learning in NLP.
- Contributed to the book 'Natural Language Processing: Representation Learning and Neural Models.'
- Secondary reviewer for conferences/journals: ICML2022, NeurIPS2022, NLPCC2023, TALLIP, TACL, PR.

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*

○ **Purpose:** Enhance brain network representation by leveraging temporal information.

○ **Role & Contribution:**

- Introduced the BrainTGL framework that combines GCN and LSTM for spatio-temporal dynamics in resting-state fMRI data;
- independently managed the project from ideation to paper submission, constituting my undergraduate thesis here, under advisor guidance. Submitted to JBHI, PR, and CIBM.

○ **Results:** Achieved state-of-the-art accuracy on ABIDE and HCP datasets; collaborated with Southern Medical University Hospital, leading to the framework's implementation in their medical system.

† **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*

○ **Purpose:** Tackle subject heterogeneity and noise correlations in brain networks.

○ **Role & Contribution:**

- Proposed the concept of consistency regularization to bolster alignment across different views;
- Developed the codebase and implemented the graph clustering module using Pytorch; Designed and crafted all figures.

○ **Results:** Achieved improved classification and alignment with ASD biomarkers.

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

- Implemented and optimized key algorithms such as logistic regression, K-means, SVM, Bayes, and further explored text hierarchical clustering, addressing complex real-world challenges.
- Delved into deep learning with neural networks, CNN and GAN; Spearheaded a critical project on target detection and recognition, utilizing the MTCNN algorithm to achieve precision and efficiency.

- The internship counted as my **Undergraduate Field Practice course**, where I achieved rank 1.

Skills & Self-Evaluation

- Proficient in academic writing and creating figures using LaTeX.
- Programming Languages: Python, C, C++, Java, LaTeX.
- Tools & Frameworks: PyTorch, Fairseq, Tikz, Numpy, and more.
- English Proficiency: TOEFL iBT 87 (R: 27, L: 19, S: 19, W: 22).
- Driven and passionate with a keen academic inclination; Adaptable team player with strong interpersonal skills.

Projects & Competitions

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

- Designed a unified framework integrating node attention, edge attention, and GCNs.
- Utilized the ni-learn toolkit to gather ABIDE fMRI time series data and implemented the entire pipeline in PyTorch.
- Assessed the framework's efficacy using a 10-fold cross-validation method, ensuring robust performance.

BANGC Programming Practice **Shengyang, CN**
Practice of Intelligent Computing System *Nov. 2020 - Dec. 2020*

- Implemented BANGC power difference operation and integrated it with TensorFlow, boosting computing efficiency.
- Optimized the operation for real-time application through quantization and developed offline inference.
- Achieved the Outstanding Student Award and was distinguished as a Cambrian Best Developer.

Chinese College Student Computer Game Competition **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, optimizing its decision-making process.
- Competed in a national-level championship with 2000+ participants and achieved the Third Award.

Scholarships & Awards

- Northeastern University Excellent Student Scholarship $\times$**4**.
- Northeastern University Chuanglian Industrial Scholarship, 2018-2019.
- The Third Award of 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.
- National Level Good, National Training Programs of Innovation and Entrepreneurship for Undergraduates.

Conferences Attended

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