

Nguyen Minh Tam

+84 973172740 | nguyenminhtam9520@gmail.com | <https://github.com/Minhtam2010>

RESEARCH INTEREST

My research focuses on developing a probabilistic framework for self-attentions in transformers. Studying the probabilistic perspective of transformers allows me to explain and reduce the model redundancy, providing a principled approach to designing more efficient transformers.

EDUCATION

University of Business and Economics - VNU

Bachelor of International Business and economics

GPA: 3.68/4.00

Ha Noi

Sep. 2013 - Jan. 2018

EXPERIENCE

AI Engineer

Sun Asterisk Inc.

Apr 2020 – Apr 2021

Ha Noi

- CVPR AI city challenge track 5: Natural Language-Based Vehicle Retrieval
 - * Achieve 2nd ranking on public test dataset and 4th ranking on private test dataset.
- SemEval 2021 Task 5: Toxic Span Detection
 - * Achieve 2nd ranking.
 - * Report Paper: S-NLP at SemEval-2021 Task 5: An Analysis of Dual Networks for Sequence Tagging
- VLSP 2020 Relation Extraction
 - * Achieve 3rd ranking.

AI Resident

FPT Software

May 2021 – Present

Ha Noi

- Paper: Improving Transformers with Probabilistic Attention Keys - accepted at ICML 2022
- Paper: FiSHformer: Transformer with a Finite Admixture of Shared Heads - submitted as Neural IPs 2022
- Paper: A Probabilistic Framework for Pruning Transformers via a Finite Admixture of Keys - submitted at ECCV 2022

PROJECTS

Improving Transformers with Probabilistic Attention Keys

Research

Jun. 2021 – Nov. 2021

- We propose Transformer with a Mixture of Gaussian Keys (Transformer-MGK), a novel transformer architecture to address the head-redundancy problem in Transformer.
- Our method allows each attention head to focus on different parts of the input sequence efficiently by replacing Softmax attention with mixture of keys at each head.
- Transformer-MGK accelerates training and inference, has fewer parameters, and requires fewer FLOPs to compute while achieving comparable or better accuracy across tasks.

FiSHformer: Transformer with a Finite Admixture of Shared Heads

Research

Jan. 2022 – Feb. 2022

- We construct an admixture model for shared attention matrices between heads and propose FiSHformer, a novel class of transformers that take advantage of this admixture model to efficiently compute multi-head attention.
- We introduce a nonlinearity mapping from global heads to local heads into FiSH and propose the Generalized FiSHformer (GFISHformer).
- We empirically verify that FiSHformer and GFISHformer achieve similar or even better accuracy but with much less computational cost in terms of FLOPs and smaller model complexity measured by the number of parameters.

A Probabilistic Framework for Pruning Transformers via a Finite Admixture of Keys

Research

Nov. 2021 – Mar. 2022

- We develop FiAK, a new finite admixture of keys for self-attention that allows key sharing to diversify attention patterns while guaranteeing the efficiency of the model.

- We design a probabilistic framework for pruning transformers that employs the prior distributions of keys in FiAK to remove redundant attention scores and keys.
- We demonstrate the advantages of our FiAK-based pruning protocols on Imagenet object recognition, COCO object detection, and WikiText-103 language modeling tasks.

Multi-modal with Natural Language-Based Vehicle Retrieval

CVPR Challenge

Apr. 2021

- Applying contrastive learning for Natural Language-Based Vehicle Retrieval. InfoNCE and Marginal Triplet Loss have been used.
- Customize Self-training technique for sequence tagging problem, results in a marginal improvement over the strong baseline.

Self-training for Toxic Span Detection

SemEval Challenge

Jan. 2021

- Pre-train on Mask Language Model task with a large amount of in-domain data to improve the model's adaptability.
- Utilizing pretrained sentence-level embeddings for Hard Negative Mining.

Multi-task learning for Relation Extraction

VLSP Challenge

Oct. 2020

- Applying Biaffine Attention layer to enhance the prediction of directional relations.
- Multi-task learning: training Name-Entity Recognition and Relation Extraction with table filling output format.

TECHNICAL SKILLS

Languages: IELTS 7.5

Frameworks: Pytorch

Programing Language: Python

Libraries: Pandas, Numpy, Hugging Face, Flair, Spacy, Sckit Learn

INTERESTS

Behavioral Economics: Books: Thinking fast and slow, Noise, Nudge, Freakonomics (Series), etc.