Minghao Li

☎ +86 13121276176 | @ liminghao1630@buaa.edu.cn | ⦿ GitHub | ⊕ Homepage | ⦿ Beijing, China | 🆔 Chinese CV

EDUCATION

Beihang University   Beijing, China
Ph.D. in Computer Science;   *Sep 2019 − Jun 2024 (Expected)*
Beihang University   Beijing, China
B.S. in Computer Science; **GPA: 3.65/4.00 Top 10%**   *Sep 2015 − Jun 2019*
Lund University   Lund, Sweden
Exchange Student;   *Aug 2017 − Jan 2018*

PUBLICATIONS AND PREPRINTS

LayoutLM: Pre-training of Text and Layout for Document Image Understanding   KDD
Yiheng Xu, **Minghao Li***, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou*   *487 Citations*

TableBank: A Benchmark Dataset for Table Detection and Recognition   LREC
***Minghao Li**, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, Zhoujun Li*   *172 Citations*

DocBank: A Benchmark Dataset for Document Layout Analysis   COLING
Minghao Li, Yiheng Xu*, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, Ming Zhou*   *129 Citations*

TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models   AAAI
***Minghao Li**, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, Furu Wei* *136 Citations*

API-Bank: A Benchmark for Tool-Augmented LLMs   EMNLP
Minghao Li, Yingxiu Zhao*, Bowen Yu, Feifan Song, Haiyang Yu, Zhoujun Li, Fei Huang, Yongbin Li *36 Citations*

Preference Ranking Optimization for Human Alignment   arXiv
*Feifan Song, Bowen Yu, **Minghao Li**, Haiyang Yu, Fei Huang, Yongbin Li, Houfeng Wang*   *10 Citation*

RESEARCH EXPERIENCE

Conversational AI, Alibaba DAMO Academy   Beijing, China
Research Intern   *Feb 2023 − Present*

API-Bank: A Benchmark for Tool-Augmented LLMs   Paper | GitHub
- Propose an evaluation benchmark designed specifically for tool-enhanced LLMs, the API-BANK dataset, containing 753 API calls and 314 annotated conversations, and a runnable evaluation system consisting of 73 API tools, to assess the capability of existing LLMs in planning, retrieving and calling APIs.
- Construct a comprehensive training set, containing 1,888 dialogues using 2,138 APIs across 1,000 different domains. Training on this train set yielded Lynx, a tool-augmented LLM initialized from Alpaca.
- The experimental results show that GPT-3.5 exhibits better tool usage compared to GPT-3, while GPT-4 performs well in terms of planning. However, there is still significant potential for further improvement. In addition, Lynx approaches the effectiveness of GPT-3.5.

Preference Ranking Optimization for Human Alignment   Paper | GitHub
- It is found that Human Alignment can be modeled as an Alignment between LLM's generative probability ranking of N Responses and Human's (Reward Model) ranking of their preferences, and the longer, more diverse, and higher quality the sequence of rankings, the better the Alignment effect.
- PRO outperformed current Human Alignment methods in automated assessment, Reward assessment, GPT-4 assessment and human assessment, and was comparable in quality to ChatGPT and human annotation results.

Works in Progress
- In the setting of unaccessible pre-training data, we investigate how to make large language models (LLMs) maintain prior task knowledge and improve forgetting problems in continual learning using the parameter effective fine-tuning (PEFT) method to achieve knowledge accumulation capability during continual fine-tuning of multiple tasks.

NLC Group, Microsoft Research Asia   Beijing, China
Joint-Ph.D.   *Jul 2018 − Feb 2023*

TableBank: A Benchmark Dataset for Table Detection and Recognition   Paper | GitHub

- Build the TableBank dataset, a new image-based table analysis dataset from online Word and Latex documents, orders of magnitude larger than existing human-labeled datasets.
- The applied weakly supervised automated annotation method can be continued to extend TableBank, as well as in more document tasks.

LayoutLM: Pre-training of Text and Layout for Document Image Understanding Paper | GitHub
- Pre-train the textual and layout information from scanned document images in a single framework, as well as leverage Image features in the downstream tasks.
- Use the masked visual-language model and the multi-label document classification as the training objectives and achieve SOTA in several document image understanding tasks.

DocBank: A Benchmark Dataset for Document Layout Analysis Paper | GitHub
- Present the DocBank dataset, a large-scale dataset built with weak supervision approach, which enables models to integrate both the textual and layout information.
- The dataset adopts a granularity intermediate between pixel and object, proves the effectiveness, and makes a possible fair comparison of models in different modalities.

TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models Paper | GitHub
- Propose TrOCR, an end-to-end Transformer-based OCR model for text recognition, the first work jointly leverages pre-trained image and text Transformers for the text recognition task in OCR.
- Achieve SOTA in printed, handwritten and scene text recognition tasks, which is convolution free and does not rely on any complex pre/post-processing steps.

Awards & Impact

LayoutLM series models are ranked 14th in HuggingFace Hub downloads, after BERT, RoBERTa and other models.

The built TableBank and DocBank datasets are used by Google, Microsoft, Alibaba, Baidu and other organizations, as well as researchers at Stanford University, Tsinghua University, Peking University and other schools

The published papers have been covered by well-known AI media or individuals, such as Synced Global, AI Era, Twitter AK, etc.

Google Scholar Citation: 952

ESI Highly Cited Paper: Top 1% in Computer Science

NOIP: First Prize of Shandong Province