Knowledge Distillation for Language Models

Yuqiao Wen[1] **Freda Shi**[2,*] **Lili Mou**[1,*]
[1]Dept. Computing Science & Alberta Machine Intelligence Institute (Amii)
University of Alberta *Canada CIFAR AI Chair
[2]David R. Cheriton School of Computer Science, University of Waterloo; Vector Institute
yq.when@gmail.com fhs@uwaterloo.ca doublepower.mou@gmail.com

Abstract

Knowledge distillation (KD) aims to transfer the knowledge of a *teacher* (usually a large model) to a *student* (usually a small one). In this tutorial, our goal is to provide participants with a comprehensive understanding of the techniques and applications of KD for language models. After introducing the basic concepts including intermediate-layer matching and prediction matching, we will present advanced techniques such as reinforcement learning-based KD and multi-teacher distillation. For applications, we will focus on KD for large language models (LLMs), covering topics ranging from LLM sequence compression to LLM self-distillation. The target audience is expected to know the basics of machine learning and NLP, but do not have to be familiar with the details of math derivation and neural models.

1 Introduction

Recent advances in deep learning have largely changed the field of natural language processing (NLP). In particular, large language models (LLM) have been the cornerstone of NLP research, and they are now tightly integrated into our daily lives. Despite the success of LLMs in a wide range of applications, they may be cumbersome to use due to their high memory and computational overhead. This calls for an increasing need to make these models more efficient and accessible, so a broader range of users can benefit from LLMs.

Researchers have been working on reducing the computational cost of running LLMs in various ways. For example, *model pruning* is a technique that removes "low-impact" parameters of a network to reduce the memory usage (LeCun et al., 1989; Liu et al., 2018; Fan et al., 2021). Alternatively, *quantization* aims to reduce the number of bits used to represent the parameters without severely deteriorating the performance (Han et al., 2016; Tao et al., 2022).

In this tutorial proposal, we will focus on *knowledge distillation* (Hinton et al., 2015; Kim and Rush, 2016), which aims at transferring knowledge from a teacher (typically a large model) to a student (known as the *student*). It has gained increasing attention in the NLP community, driven by the demands of compressing the ever-growing and high-performing language models.

After an introduction and overview, we will start the tutorial with the basics of KD, mainly falling into the following two categories: intermediate-layer matching and prediction matching. The former refers to the distillation of intermediate layers, including activated features (Sun et al., 2019; Shleifer and Rush, 2020; Yu et al., 2025) and attention weights (Jiao et al., 2020; Wang et al., 2021); we will also discuss relational learning, which distills the relative structures of features (e.g., transformations) instead of the absolute feature values (Wang et al., 2021; Huang et al., 2023b).

For the prediction matching, we will present the classic cross-entropy approach, with an emphasis on its multi-modality issue[1] (Wei et al., 2019; Bao et al., 2020; Khan et al., 2020; Wen et al., 2023a): when the student model's capacity is not large enough, it is unable to learn the multi-modal distribution predicted by the large teacher, oftentimes resulting in severe model collapse and mode issues. We will discuss different divergence-based methods (Kim and Rush, 2016; Wen et al., 2023b) to mitigate this issue.

Then, we will move on to the second part of the tutorial, where we present two selected topics on advanced KD techniques: reinforcement learning (RL)-based KD and multi-teacher KD. Reinforcement learning has been gaining increasing attention in recent years, due to its success in training LLMs, showing great success in aligning the model with

[1]Here, a *mode* refers to a peak of a distribution. It should not be confused with "multi-modality" that refers to multimedia data (e.g., text, image, and video).

Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, pages 25–29

human preference as well as mitigating exposure bias (Ouyang et al., 2022). In here, we will dive into RL in the context of knowledge distillation, where the key challenge is to derive a reward function based on the teacher model (Hao et al., 2022; Li et al., 2024).

We will also discuss multi-teacher KD, where the student model learns from multiple teachers, each having its own expertise. This ties closely to the multi-modality problem that we have posed in the first part of our tutorial, where the knowledge is too diverse for the student to learn. We present a solution to this based on the ensemble-then-distill framework, where an ensemble process is applied before distillation (Shayegh et al., 2024a,b; Wen et al., 2025b). This allows the student to learn high-quality, consolidated knowledge instead of conflicting knowledge from different teachers.

The last part of our tutorial will focus on KD with large language models (LLMs). We start by presenting interesting phenomena observed in LLM distillation, such as the effect of teacher intervention (Saha et al., 2023) and emerging chain-of-thought abilities in small models (Fu et al., 2023). Then, we will showcase how KD can be used to compress the prompts (Wingate et al., 2022; Sun et al., 2023; Chuang et al., 2024; Mu et al., 2023) and the reasoning process (Deng et al., 2024; Cheng and Van Durme, 2024) to speed up inference. We will move on to self-distillation, where LLMs are able to reflect upon its own generations and learn skills such as instruction following (Wang et al., 2023; Sun et al., 2023), reasoning (Huang et al., 2023a) and summarization (Jung et al., 2024). Finally, we will walk through modern distilled systems, including Alpaca (Taori et al., 2023), Vicuna (Chiang et al., 2023), and DeepSeek's distilled models (Guo et al., 2025) to gain a sense of practical use of KD techniques. We conclude the tutorial by showing surprising and interesting applications related to KD, including quantization (Polino et al., 2018; Wen et al., 2025a), speculative decoding (Zhou et al., 2024), and non-autoregressive translation (Zhou et al., 2020).

Overall, this tutorial will lay a solid foundation of knowledge distillation for language models, with highlights of both machine learning challenges and cutting-edge applications.

2 Target Audience

The tutorial targets a diverse audience, including machine learning and NLP researchers, as well as practitioners.

We expect the audience to have a brief knowledge of deep learning (e.g., cross-entropy loss and back-propagation training) and NLP (e.g., autoregressive text generation and large language models). However, the audience do **not** have to be familiar with the details (e.g., derivative calculations, transformer attention formulas); only a general impression would suffice.

The audience does **not** have to have heard of knowledge distillation. We will teach the foundations before moving on to cutting-edge algorithms and applications.

3 Outline

PART I: Introduction [10min]

- KD definition
- Motivation
- Overview of this tutorial

PART II: KD Basics [45min]

- Overview
- Intermediate-layer matching
  - Matching loss
  - Layer selection
- Prediction-matching KD
  - Classic cross-entropy matching
  - f-divergence matching
  - Ranking-based matching

BREAK [10min]

PART III: Selected Advanced KD Techniques [45min]

- Reinforcement learning for KD
  - Motivation and challenges
  - Reward induction from teacher
- Multi-teacher KD
  - Motivation and challenges
  - Ensemble-then-distill framework

BREAK [10min]

PART IV: KD Applications for LLMs [45min]

- Empirical findings in LLM distillation
- LLM sequence compression
- LLM self-distillation for performance improvement
- SOTA distilled systems
- Other interesting KD applications

PART V: Conclusion, Future Directions, and QA [15min]

4 Presenters

Yuqiao Wen is currently a third-year PhD student at the Department of Computing Science, University of Alberta, after having his MSc in 2022 and BSc in 2020. Yuqiao's research lies in developing efficient methods for large language models and making them more accessible for everyone; he has a focus on machine learning problems in knowledge distillation such as label bias and exposure bias. He has published a number of papers at top-tier venues such as AAAI, ACL, and ICLR, including one winning an Area Chair's Award. He was a co-presenter of a three-hour tutorial at the Amii Upper Bound Conference, which attracted several thousand attendees.

Freda Shi is a first-year Assistant Professor in the David R. Cheriton School of Computer Science at the University of Waterloo and a Faculty Member at the Vector Institute. Her research interests are in computational linguistics, natural language processing, and cognitive sciences. She has been working on knowledge distillation for syntactic analysis and multilingualism, with relevant papers published at ACL and ICLR. Her work has been recognized with a Google PhD Fellowship, a Facebook Fellowship Finalist Award, and Best Paper Nominations at ACL 2019, 2021, and 2024. She has served as an Area Chair for conferences such as ACL, EMNLP, and COLM, and as a program committee member or a reviewer for leading journals and conferences in computational linguistics and machine learning, including TACL, TPAMI, ACL, COLING, EMNLP, NAACL, ICLR, ICML, and NeurIPS.

Lili Mou is a sixth-year Assistant Professor at the Department of Computing Science, University of Alberta. His main research interest lies in developing novel machine learning methods for NLP tasks; successful examples include tree-based convolutional neural networks, edit-based unsupervised text generation, and an ensemble-then-distill framework for multi-teacher KD. He regularly serves as a Senior Program Committee Member or an Area Chair for AI and NLP conferences, and is an Action Editor for ACL Rolling Review. He is an Amii Fellow and a Canada CIFAR AI Chair, and has received a AAAI New Faculty Highlight Award; he also received an ACL Best Paper Nomination (2019) and ACL Area Chair's Award (2024). Lili has been a co-organizer of the Workshop on Efficient Speech and Natural Language Processing, co-located with NeurIPS during 2021–2023. He presented two conference tutorials at EMNLP-IJCNLP 2019 and ACL 2020.

Acknowledgments

The research is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Amii Fellow Program, the Canada CIFAR AI Chair Program, an Alberta Innovates Program, a donation from DeepMind, and the Digital Research Alliance of Canada (alliancecan.ca).

References

Siqi Bao, Huang He, Fan Wang, Hua Wu, and Haifeng Wang. 2020. PLATO: Pre-trained dialogue generation model with discrete latent variable. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 85–96.

Jeffrey Cheng and Benjamin Van Durme. 2024. Compressed chain of thought: Efficient reasoning through dense representations. *arXiv preprint arXiv:2412.13171*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. *LMSYS Blog*.

Yu-Neng Chuang, Tianwei Xing, Chia-Yuan Chang, Zirui Liu, Xun Chen, and Xia Hu. 2024. Learning to compress prompt in natural language formats. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7756–7767.

Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit COT to implicit COT: Learning to internalize COT step by step. *arXiv preprint arXiv:2405.14838*.

Chun Fan, Jiwei Li, Tianwei Zhang, Xiang Ao, Fei Wu, Yuxian Meng, and Xiaofei Sun. 2021. Layer-wise model pruning based on mutual information. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3079–3090.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *Proceedings of the International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10421–10430.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Song Han, Huizi Mao, and William J. Dally. 2016. Deep compression: Compressing deep neural network with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*.

Yongchang Hao, Yuxin Liu, and Lili Mou. 2022. Teacher forcing recovers reward functions for text generation. In *Advances in Neural Information Processing Systems*, pages 12594–12607.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2023a. Large language models can self-improve. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068.

Kun Huang, Xin Guo, and Meng Wang. 2023b. Towards efficient pre-trained language model via feature correlation distillation. In *Advances in Neural Information Processing Systems*, pages 16114–16128.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 4163–4174.

Jaehun Jung, Peter West, Liwei Jiang, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. 2024. Impossible distillation for paraphrasing and summarization: How to make high-quality lemonade out of small, low-quality model. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4454.

Kashif Khan, Gaurav Sahu, Vikash Balasubramanian, Lili Mou, and Olga Vechtomova. 2020. Adversarial learning on the latent space for diverse dialog generation. In *Proceedings of the International Conference on Computational Linguistics*, pages 5026–5034.

Yoon Kim and Alexander M. Rush. 2016. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327.

Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, pages 598–605.

Dongheng Li, Yongchang Hao, and Lili Mou. 2024. LLMR: Knowledge distillation with a large language model-induced reward. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 10657–10664.

Liyuan Liu, Xiang Ren, Jingbo Shang, Xiaotao Gu, Jian Peng, and Jiawei Han. 2018. Efficient contextualized representation: Language model pruning for sequence labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1215–1225.

Jesse Mu, Xiang Li, and Noah Goodman. 2023. Learning to compress prompts with gist tokens. In *Advances in Neural Information Processing Systems*, pages 19327–19352.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *OpenAI Blog*.

Antonio Polino, Razvan Pascanu, and Dan Alistarh. 2018. Model compression via distillation and quantization. In *International Conference on Learning Representations*.

Swarnadeep Saha, Peter Hase, and Mohit Bansal. 2023. Can language models teach weaker agents? Teacher explanations improve students via personalization. *arXiv preprint arXiv:2306.09299*.

Behzad Shayegh, Yanshuai Cao, Xiaodan Zhu, Jackie CK Cheung, and Lili Mou. 2024a. Ensemble distillation for unsupervised constituency parsing. In *International Conference on Learning Representations*.

Behzad Shayegh, Yuqiao Wen, and Lili Mou. 2024b. Tree-averaging algorithms for ensemble-based unsupervised discontinuous constituency parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 15135–15156.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for BERT model compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4323–4332.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *Advances in Neural Information Processing Systems*, pages 2511–2565.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4821–4836.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. *Standfard Blog*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2021. MiniLMv2: Multi-head self-attention relation distillation for compressing pre-trained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, pages 2140–2151.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 13484–13508.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2019. Why do neural dialog systems generate short and meaningless replies? A comparison between dialog and translation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 7290–7294.

Yuqiao Wen, Yanshuai Cao, and Lili Mou. 2025a. Exploring model invariance with discrete search for ultra-low-bit quantization. *arXiv preprint arXiv:2502.06844*.

Yuqiao Wen, Yongchang Hao, Yanshuai Cao, and Lili Mou. 2023a. An equal-size hard EM algorithm for diverse dialogue generation. In *International Conference on Learning Representations*.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023b. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 10817–10834.

Yuqiao Wen, Behzad Shayegh, Chenyang Huang, Yanshuai Cao, and Lili Mou. 2025b. EBBS: An ensemble with bi-level beam search for zero-shot machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

David Wingate, Mohammad Shoeybi, and Taylor Sorensen. 2022. Prompt compression and contrastive conditioning for controllability and toxicity reduction in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 5621–5634.

Zony Yu, Yuqiao Wen, and Lili Mou. 2025. Revisiting intermediate-layer matching in knowledge distillation: Layer-selection strategy doesn't matter (much). *arXiv preprint arXiv:2502.04499*.

Chunting Zhou, Jiatao Gu, and Graham Neubig. 2020. Understanding knowledge distillation in non-autoregressive machine translation. In *International Conference on Learning Representations*.

Yongchao Zhou, Kaifeng Lyu, Ankit Singh Rawat, Aditya Krishna Menon, Afshin Rostamizadeh, Sanjiv Kumar, Jean-François Kagy, and Rishabh Agarwal. 2024. DistillSpec: Improving speculative decoding via knowledge distillation. In *International Conference on Learning Representations*.