Efficient Multi-dimensional Optimization in Abstractive Summarization via Mixture-of-Learnable Prompts Tuning

Fengyu Lu
Peking University, Beijing, Beijing, China

Jiaxin Duan
Peking University, Beijing, Beijing, China

Junfei Liu
Peking University, Beijing, Beijing, China

Abstract

Human beings prefer to read concise summaries that rephrase the exact ideas of a document using novel statements. Consequently, previous works endeavor to coordinate the faithfulness and abstractiveness of automatic summarization, yet this leads to increased computation or data overhead. To address this problem, we propose a novel prompt tuning approach, MoLP, which allocates the optimization of parallel objectives in abstractive summarization to learnable prompts and effectively relieves the cost burden. Inspired by the neural mixture-of-experts model, MoLP learns input-specific expert prompts to optimize saliency awareness, faithfulness, and abstractiveness, respectively, and learns a task-specific router prompt to weigh and polymerize the experts' effects. More importantly, these lightweight prompts are learned from separate tasks, each built upon a heuristic summary of the same document, significantly saving computing costs and improving data utilization. In experiments, we plug MoLP into frozen language models following the classical prompt tuning setting. Extensive evaluations across four benchmark tasks witness the model-generated summaries with simultaneously improved faithfulness and abstractiveness scores. Few-shot learning tests also underscore the advanced generalization of our method.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*