SynText: Momentum Calibration For Multi-Document Summarization

Anonymous ACL submission

Abstract

Multi-document summarization, a complex task in natural language processing, requires synthesizing information from multiple texts. Despite the focus on pre-training in recent research, the role of fine-tuning has been underexplored. We introduce SynText, a model that builds on the PRIMERA model for multi-document summarization through momentum calibration fine-tuning. Our results show that SynText surpasses the current state-of-the-art on the MultiNews dataset across all major ROUGE metrics. This work highlights the importance of not taking fine-tuning strategies for granted.

1 Introduction

1.1 Background

Multi-document summarization, generating summaries from related documents, has seen significant advancements with pre-trained language models, especially encoder-decoder transformers (Xiao et al., 2022; Beltagy et al., 2020). The current state-of-the-art involves pyramid-based masked sentence pretraining, superior to other transformer models in data diversity (Xiao et al., 2022). This method trains models to identify and aggregate key information across document clusters.

1.2 Research Gap

However, research has focused more on pre-training than on fine-tuning techniques. Vanilla fine-tuning, based on Maximum Likelihood Estimation, faces challenges like exposure bias and loss-evaluation mismatch, affecting model performance during evaluation (Wiseman and Rush, 2016; Ranzato et al., 2015).

1.3 Contributions

This paper introduces momentum calibration fine-tuning, a specialized alternative to enhance multi-document summarization, extending the gains seen in single-document summarization to multi-document contexts (Zhang et al., 2022). Our main contributions include:

1. Exploring advanced fine-tuning methods for performance improvement in multi-document summarization.

2. Presenting SynText, a model using momentum calibration fine-tuning, showing significant performance gains on the MultiNews dataset over the state-of-the-art PRIMERA model.

1.4 Organization

The paper will discuss pre-existing research, our model's architecture, pre-training, fine-tuning strategies, experimental setup, results, limitations, and potential future work and ethical considerations.

2 Relevant work

2.1 Multi-document summarization

Multi-document summarization, a task of synthesizing information from multiple documents, employs various methods categorized into graph-based models, hierarchical models, and pretrained transformer-based models.

2.2 Graph-based summarization

Graph-based models utilize graph neural networks to synthesize inter-document information. However, they depend on external data like discourse structures for graph construction, complicating their use (Liao et al., 2018; Li et al., 2020).

2.3 Hierarchical summarization

Hierarchical models create higher-level document representations before information synthesis. Their downside is the often-required domainspecific knowledge, limiting generalizability (Jin et al., 2020; Su et al., 2021).

2.4 Attention-based summarization

Recent advancements in multi-document summarization have been driven by pretrained transformer models. These models, such as Longformer, efficiently process lengthy sequences and manage long-range dependencies, crucial for multi-document contexts (Beltagy et al., 2020). PRIMERA, a state-of-the-art model, pre-trains a Longformer with an entity-based sentence masking objective, enabling effective cross-document information integration (Xiao et al., 2022).

2.5 Fine-tuning strategies

Despite its success, PRIMERA's vanilla fine-tuning approach misses out on potential performance enhancements from more specialized fine-tuning methods. Various fine-tuning strategies have emerged to maximize pretrained model outputs. These include scheduled sampling and optimization algorithms inspired by reinforcement learning action sequences (Bengio et al., 2015), and contrastive learning approaches for improving text generation in summarization (Pan et al., 2021).

2.6 Ranking-based fine-tuning

Long-standing two-stage re-ranking techniques in text generation for summarization have shown effectiveness. Some methods involve re-ranking output from neural text generation models, demonstrating promising results (Liu et al., 2021).

2.7 Online fine-tuning

In contrast to these, our model employs momentum calibration, a shared-parameter technique. The current state-of-the-art in text generation, momentum calibration aligns candidate sample probabilities with their actual quality, measured by an external metric. This online fine-tuning approach uses a generator model, which is a momentum moving average of the online model, to generate candidate samples for fine-tuning (Zhang et al., 2022).

2.8 Extending momentum calibration

While momentum calibration has proven effective in text generation and summarization, its application to multi-document summarization remains unexplored. Our work investigates the potential of combining momentum calibration with vanilla fine-tuning to achieve further performance gains in multi-document summarization.

3 Model

3.1 Model overview

SynText, a blend of "synthesize" and "text," is based on the PRIMERA model, which currently leads in multi-document summarization (Xiao et al., 2022). Rather than starting from scratch, we build on established models and enhance them with momentum calibration.

3.2 Model architecture

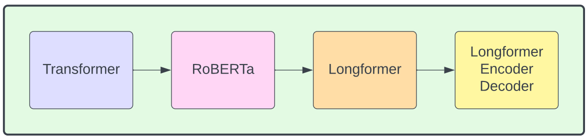

Figure 1: Architecture overview

SynText employs a transformer architecture, widely recognized as the universal framework for various natural language processing tasks (Vaswani et al., 2017; Lin et al., 2021). Specifically, we use the Longformer-Encoder-Decoder variant, adept at handling long text sequences, a key requirement for summarization (Beltagy et al., 2020). Traditional transformers struggle with long sequences due to their quadratic scaling self-attention mechanism. Longformer counters this with a blend of local windowed and global attention, transforming the processing from quadratic to linear complexity. It builds upon RoBERTa's large-scale linguistic and semantic exposure, enabling efficient multi-document handling (Liu et al., 2019).

3.3 Model pre-training

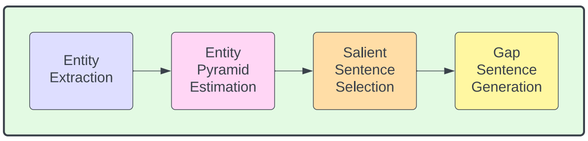

Figure 2: Pre-training overview

PRIMERA extends Longformer-Encoder-Decoder by employing a Gap Sentence Generation objective tailored for multi-document summarization. Task-specific pre-training has been shown to offer performance benefits (Zhang et al., 2020). The Entity Pyramid approach, inspired by the Pyramid Evaluation method, involves generating salient sentences that have been masked with a [sent-mask] token. These sentences are ranked using Cluster ROUGE, based on entity frequency from the document cluster (Xiao et al., 2022).

3.4 Model fine-tuning

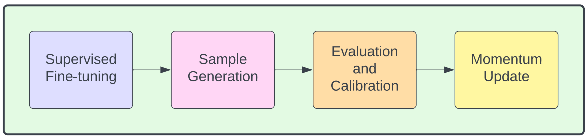

Figure 3: Fine-tuning overview

While PRIMERA's vanilla fine-tuning excels on the MultiNews dataset, we believe specialized fine-tuning strategies can enhance performance. We employ momentum calibration, an online method proven in single-document summarization (Zhang et al., 2022). This method involves two model copies: a generator and an online model. The generator, whose parameters are a moving average of the online model, creates candidate samples. These samples are then evaluated using ROUGE, ranking them to compute a margin-based pairwise ranking loss. This loss, combined with vanilla loss, refines the online model, which undergoes parameter updates and momentum adjustments. We first conduct vanilla fine-tuning before applying momentum calibration, expecting significant performance gains in multi-document summarization.

3.5 Model capabilities

In summary, SynText's strength lies in its combination of a robust transformer-based architecture and a novel fine-tuning approach. By leveraging momentum calibration, we anticipate surpassing the capabilities of existing models in multi-document summarization, as demonstrated in our performance evaluations at each training stage.

4 Experiments

4.1 Model checkpoint

For our experiments, we used the pre-trained PRIMERA model from its official GitHub repository, which outperforms the version on HuggingFace. We selected the last publicly available model checkpoint for our foundation. Our work minimizes dependency on external libraries, relying primarily on PyTorch version 2.1. The libraries and model checkpoints we do use are free, open-source, and publicly-available and these were used as originally intended.

4.2 Dataset

Our results are based on the MultiNews dataset (Fabbri et al., 2019), which consists of numerous news articles and their human-written summaries. These summaries are crafted for fluency rather than mere compression, offering a more realistic model for real-world applications. The dataset varies in the number of documents per example, ranging from 2 to 10, aiding in model generalization. The news articles and the summaries are in English. This dataset is free, open-source, and publicly-available and it was used as originally intended.

The dataset was sourced from HuggingFace, with preprocessing involving the removal of extraneous newlines and splitting training input on the "|||||" symbol. For processing, documents in a cluster are concatenated with a <doc-sep> token, applying a global attention mask, and tokenized with a maximum length of 4096.

4.3 Hyperparameters

Our hyperparameters mirror those used in single-document summarization for momentum calibration (Zhang et al., 2022). This includes 8 epochs, a batch size of 16, Adam optimizer, linear learning rate scheduling with 1 warm-up step, and 225 training steps. For momentum calibration, we set the number of candidate samples to 16, margin coefficient to 0.001, length normalization to 2, vanilla loss weighting to 0.01, and momentum coefficient to 0.995.

4.4 Evaluation metrics

We evaluated model performance using ROUGE metrics (ROUGE-1, ROUGE-2, and ROUGE-L), the industry standard for summarization algorithms (Lin, 2004). These metrics assess unigram, bigram overlap, and longest common subsequence, respectively. These metrics are free, open-source, and publicly-available and these were used as originally intended.

4.5 Performance

Our results on a single run of the MultiNews dataset show that SynText surpasses the current state-of-the-art, PRIMERA, across all major ROUGE metrics. Specifically, SynText achieves a ROUGE-1 score of 54.0, ROUGE-2 score of 23.8, and ROUGE-L score of 30.7, significantly outperforming PRIMERA (Xiao et al., 2022).

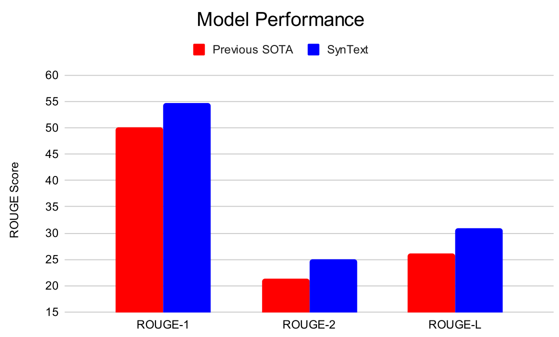

Figure 4: MultiNews results

Model	ROUGE-1	ROUGE-2	ROUGE-L
Base Model	17.3	3.7	10.4
Pyramid Evaluation Gap Sentence Generation Pre-training	42.0	13.6	20.8
Vanilla Fine-tuning	49.9	21.1	25.9
Momentum Calibration Fine-tuning	**54.0**	**23.8**	**30.7**

Table 1: Ablation study results

4.6 Ablation study

Our analysis of the training process reveals that both the vanilla fine-tuning and momentum calibration stages significantly contribute to SynText's performance. Pre-training yields the most substantial improvements, but the specialized fine-tuning strategies also demonstrate substantial gains, validating our approach. These results underscore the effectiveness of momentum calibration in enhancing multi-document summarization performance.

5 Conclusion and Future Work

In this paper, we introduced SynText, a groundbreaking model for multi-document summarization. SynText advances beyond the current state-of-the-art PRIMERA model by integrating momentum calibration fine-tuning, a technique previously successful in single-document summarization. Our results indicate that SynText significantly outperforms the established benchmarks on the Multi-News dataset, achieving superior results across all key ROUGE metrics. This improvement highlights the effectiveness of specialized fine-tuning strategies in enhancing multi-document summarization performance. Future work could apply SynText to diverse datasets, such as WikiSum (Cohen et al., 2021).

6 Limitations

The limitation of our work is that, due to a lack of computational resources, we had to run our experiments on a randomized subset of the dataset. Currently, we are working on gaining access to more computational resources so we can train and evaluate the model on the entire dataset. Nevertheless, the early results are very promising and corroborate our claims.

7 Risks and Reproducibility

The ethical risks include deep learning models like ours being used as part of disinformation campaigns. For reproducibility, we plan to release all of our code in a Google Colaboratory notebook in the future when we have access to more computing resources. Our publicly-released model will be completely open-source and will be trained and evaluated on the full MultiNews dataset and expect to see the same level of state-of-the-art performance as highlighted in our work.

References

[1] Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

[2] Bengio, S., Vinyals, O., Jaitly, N., & Shazeer, N. (2015). Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in neural information processing systems*, 28.

[3] Cohen, N., Kalinsky, O., Ziser, Y., & Moschitti, A. (2021). WikiSum: Coherent summarization dataset for efficient human evaluation.

[4] Fabbri, A. R., Li, I., She, T., Li, S., & Radev, D. R. (2019). Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. *arXiv preprint arXiv:1906.01749*.

[5] Jin, H., Wang, T., & Wan, X. (2020 July). Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 6244-6254).

[6] Li, W., Xiao, X., Liu, J., Wu, H., Wang, H., & Du, J. (2020). Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

[7] Liao, K., Lebanoff, L., & Liu, F. (2018). Abstract meaning representation for multi-document summarization. *arXiv preprint arXiv:1806.05655*.

[8] Lin, C. Y. (2004 July). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out* (pp. 74-81).

[9] Lin, T., Wang, Y., Liu, X., & Qiu, X. (2022). A survey of transformers. *AI Open*.

[10] Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

[11] Liu, Y., & Liu, P. (2021). SimCLS: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.

[12] Nenkova, A., & Passonneau, R. J. (2004). Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Hlt-naacl 2004* (pp. 145-152).

[13] Pan, X., Wang, M., Wu, L., & Li, L. (2021). Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

[14] Qiu, J., Ma, H., Levy, O., Yih, S. W. T., Wang, S., & Tang, J. (2019). Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*.

[15] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21(1), 5485-5551*.

[16] Shen, L., Sarkar, A., & Och, F. J. (2004). Discriminative reranking for machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004* (pp. 177-184).

[17] Su, A., Su, D., Mulvey, J. M., & Poor, H. V. (2021). PoBRL: Optimizing Multi-document Summarization by Blending Reinforcement Learning Policies. *arXiv preprint arXiv:2105.08244*.

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, 30.

[19] Wiseman, S., & Rush, A. M. (2016). Sequence-to-Sequence Learning as Beam-Search Optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing pages 1296–1306, Austin Texas. Association for Computational Linguistics*.

[20] Xiao, W., Beltagy, I., Carenini, G., & Cohan, A. (2022). PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) pages 5245–5263, Dublin Ireland. Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.360*.

[21] Zhang, J., Zhao, Y., Saleh, M., & Liu, P. (2020). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning PMLR 119:11328-11339*.

[22] Zhang, X., Liu, Y., Wang, X., He, P., Yu, Y., Chen, S. Q., ... & Wei, F. (2022). Momentum Calibration for Text Generation. *arXiv preprint arXiv:2212.04257*.

[23] Zou, Y., Zhang, X., Lu, W., Wei, F., & Zhou, M. (2020). Pre-training for abstractive document summarization by reinstating source text. *arXiv preprint arXiv:2004.01853*.