Final assignment: Cross-encoder re-rankers

SIWEN TU and SHUPEI LI, LIACS, Leiden University, the Netherlands

1 INTRODUCTION

2 TASK 1: EVALUATING CROSS-ENCODERS

Table 1. Effectiveness results of three fine-tuned cross-encoder models

Model	nDCG@10	Recall@100	MAP@1000	#Training Steps
MiniLM	**0.696**	**0.507**	0.450	47569 / 156250
TinyBERT	0.691	0.506	**0.458**	76301 / 156250
distilroberta	0.611	0.463	0.392	9036 / 156250

3 TASK 2: SELECT AND APPLY FIVE ENSEMBLE METHODS

Table 2. Effectiveness results of five ensemble methods

Model	nDCG@10	Recall@100	MAP@1000
Mixed	0.707	0.513	0.465
Reciprocal Rank Fusion (RRF)	0.699	0.507	0.462
BayesFuse	0.697	0.507	0.451
PosFuse	**0.710**	**0.516**	**0.480**
Weighted BordaFuse	0.707	0.511	0.464

4 TASK 3: ANALYZING MOST EFFECTIVE ENSEMBLE METHOD

The PosFuse ensemble method in Task 2 outperforms all other methods. Therefore, we apply PosFuse to all combinations of two models out of three in this task and summarize experimental results in Table 3.

Table 3. Effectiveness results of all combinations

Combination	nDCG@10	Recall@100	MAP@1000
MiniLM + TinyBERT	**0.723**	**0.518**	**0.483**
MiniLM + distilroberta	0.691	0.509	0.463
TinyBERT + distilroberta	0.694	0.507	0.468

MiniLM + TinyBERT ensemble performs best on all metrics. MiniLM + distilroberta performs worst on nDCG@10 and MAP@1000, while TinyBERT + distilroberta performs worst on Recall@100.

IR'23, Master CS, Spring 2023, Leiden, the Netherlands

The result is consistent with our expectations. Because MiniLM and TinyBERT achieve significantly better results than distilroberta in Task 1, we think the performance of the ensemble including distilroberta will be affected by the inferiority of distilroberta. There is no big difference between the performance of MiniLM + distilroberta and that of TinyBERT + distilroberta. However, MiniLM + TinyBERT performs much better than other combinations, which may be caused by removing the distilroberta ranking file. It is worth mentioning that MiniLM + TinyBERT even achieves higher scores than the ensemble of three models in Task 2. This pattern indicates that more models do not mean better performance. A weak model may introduce more noise than information, which impairs the model's performance.

5 TASK 4: MODIFYING THE EVALUATION METRIC IN FINE-TUNING

We modify the CERerankingEvaluator class to change the evaluation metric to nDCG@10. We adopt the implementation in the sklearn package to calculate nDCG@10. Note that the modified class can be easily used by changing the corresponding API in the fine-tuning notebook. No other modification is needed. Our code has been included in the submission.

6 TASK 5: ENSEMBLING THROUGH SCORE INJECTION

In this task, we investigate the effect of the score injection. The score injection is a strategy described in paper [1] that aims at improving the performance of BERT-based re-rankers by injecting BM25 scores. However, we inject the average scores of MiniLM, TinyBERT, and distilroberta, instead of BM25 scores. Besides, we use the microsoft/MiniLM-L12-H384-uncased model as the re-ranker. Our modified version of the fine-tuning notebook first performs inference with three models and sets the average score as the first sentence of the document. Then the pairs of queries and injected documents are inputted in the re-ranker. Note that we convert the type of the average score into the integer following the suggestion in [1]. Due to the GPU quota of Colab, we fine-tune the microsoft/MiniLM-L12-H384-uncased model with and without injection for one hour and run the evaluation notebook. The evaluation results are reported in Table 4. We also include results in [1] for comparison.

Table 4. Effectiveness results of ensembling with and without score injection

Model	nDCG@10	Recall@100	MAP@1000
MiniLM-L12-H384-uncased without injection	0.668	**0.503**	0.450
MiniLM-L12-H384-uncased with injection	0.664	0.495	0.435
$\text{MiniLM}_{\text{CAT}}$[1]	0.704	-	0.452
$\text{MiniLM}_{\text{BM25CAT}}$[1]	**0.711**	-	**0.463**

MiniLM with BM25 injection has the best performance during evaluation. The performance of MiniLM with the injection of three BERT-based models' predictions does not achieve the anticipated effect. We think one reason is that the training process is not sufficient. Table 4 indicates that our version of MiniLM without injection is not at the expected level as that of $\text{MiniLM}_{\text{CAT}}$. Another reason may be the instability of the model performance. We notice that MiniLM with the average score injection is not very robust during the training. The improvement of the model's robustness is a possible future direction for developing score injection strategies.

REFERENCES

[1] Arian Askari, Amin Abolghasemi, Gabriella Pasi, Wessel Kraaij, and Suzan Verberne. 2023. Injecting the BM25 Score as Text Improves BERT-Based Re-rankers. (2023). arXiv:2301.09728 [cs.IR]