MPII at the TREC 2020 Deep Learning Track

Canjia Li
University of Chinese Academy of Sciences
licanjia17@mails.ucas.ac.cn

Andrew Yates
Max Planck Institute for Informatics
ayates@mpi-inf.mpg.de

ABSTRACT

MPII participated in the TREC 2020 Deep Learning track's document ranking task with several variants of our recent PARADE model. PARADE is based on the idea that aggregating passage-level relevance representations is preferable to aggregating relevance scores. We submitted runs using three different PARADE variants that performed well in previous evaluations. The results differ from both those in the PARADE paper and those from the NTCIR-15 WWW-3 track: on this document ranking task, the least complex representation aggregation technique performs best.

1 INTRODUCTION

We participated in the TREC 2020 Deep Learning track's document ranking task in order to supplement previous evaluations of our PARADE ranking model [6]. While pre-trained Transformer architectures like BERT can be directly applied as relevance classifiers to passage ranking tasks [8], their maximum input length limitation must be overcome when applying them to longer documents [7]. Dai and Callan [3] overcame this limitation by applying BERT to each passage independently and then aggregating the passages' *scores* into a document score. PARADE attempts to improve effectiveness by aggregating passage *representations* rather than scores. In this paper we briefly summarize PARADE's aggregation strategies and describe how the model was trained for TREC DL before comparing results for three PARADE variants.

2 METHOD

PARADE uses a pre-trained Transformer-based model, such as BERT [4] or ELECTRA [2], to produce a vector representation of each passage within a document. These vector representations are then aggregated to arrive at a document relevance score using one of several approaches. The model is trained end-to-end, which avoids the need to extend a document's relevance score to apply to all passages within the document (during training). A more detailed description can be found in the original work [6]. We submitted runs corresponding to the aggregation approaches that previous performed well in the original work and in NTCIR WWW-3 [5]: PARADE $_{\text{Max}}$, PARADE $_{\text{Attn}}$, and the full Transformer-based model (denoted "PARADE").

3 EXPERIMENTAL SETUP

3.1 Data

Using a sliding window of 150 tokens with a stride of 100, we split each document into 32 passages. This resulted in 3250 ($100*31+150$) tokens being preserved in each document (excluding overlapping tokens). The maximum sequence length was set to 256 tokens, which includes both query and document terms.

Run Name	Variant	mAP	nDCG@10	MRR
mpii_run1	PARADE	0.4030	0.6017	**0.9000**
mpii_run2	PARADE $_{\text{Max}}$	**0.4205**	**0.6135**	0.8833
mpii_run3	PARADE $_{\text{Attn}}$	0.2587	0.3286	0.6388

Table 1: Deep Learning document task results.

3.2 Training

PARADE was initialized with an ELECTRA-Base model that was first fine-tuned on the MS MARCO passage ranking dataset [1]. PARADE was then trained on the judgments from the TREC 2019 Deep Learning track's document ranking task [9] by reranking the top 100 results from the organizers' baseline run. The model was trained for 3 epochs with a batch size of 32 and a learning rate of 3e-6 with warm-up over the first 10 proportions of training steps. After training, the model was used to rerank the top 100 results from the 2020 Deep Learning track's baseline run. All experiments were conducted on a Google Cloud TPU v3-8.

Our code[1] and the ELECTRA-Base model fine-tuned on MS MARCO[2] are available online. Additionally, Capreolus [10] provides independent implementations of PARADE in both TensorFlow and PyTorch as well as the fine-tuned ELECTRA-base model under the name `electra-base-msmarco`.[3]

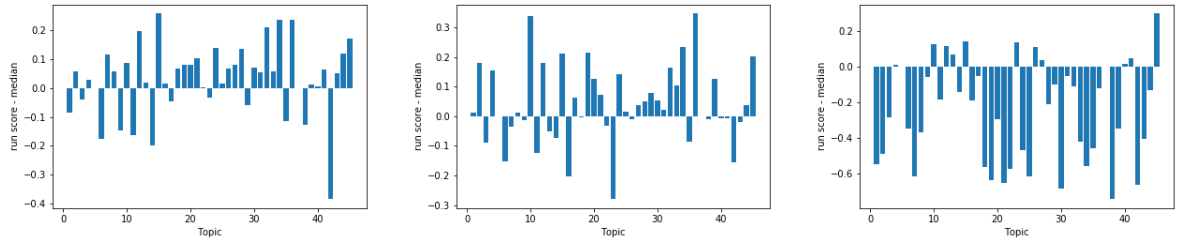
4 RESULTS

We submitted three runs using the PARADE, PARADE $_{\text{Max}}$, and PARADE $_{\text{Attn}}$ variants. The results are shown in Table 1. Unlike the results from WWW-3 and in the original paper's evaluations, PARADE $_{\text{Max}}$ (i.e., max pooling across each dimension of the passages' representations) outperforms the full PARADE model (i.e., Transformer-based aggregation of passage representations). Similarly, PARADE $_{\text{Max}}$ outperforms PARADE $_{\text{Attn}}$, though the latter's low scores suggests that training was not successful. The reason for these approaches' differences in effectiveness on the DL20 document ranking task is not clear and requires further investigation.

Differently from the queries used in the other evaluations (i.e., keyword queries with WWW-3 and topic and description queries with Robust04 and GOV2), the DL20 queries are questions. However, the documents and other aspects of the evaluation also differ; further analysis is needed to determine if the query type or differences in document characteristics have any impact.

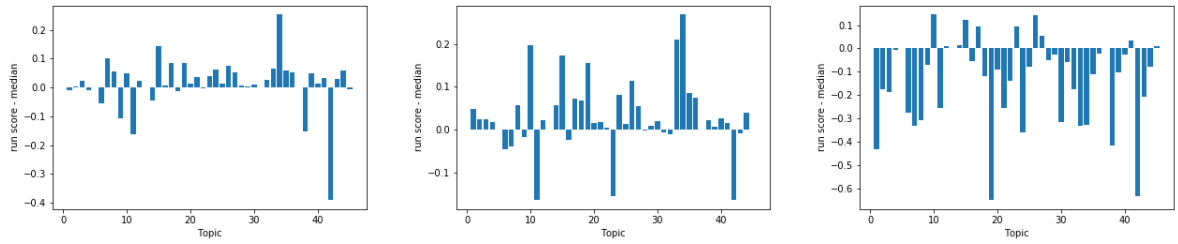
We additionally include comparisons of each variant's metrics against the median run scores in Figures 1-3. The best-performing and worst-performing queries for each variant are shown in Table 2, Table 3, and Table 4.

[1] https://github.com/canjiali/PARADE
[2] https://zenodo.org/record/3974431
[3] https://github.com/capreolus-ir/capreolus



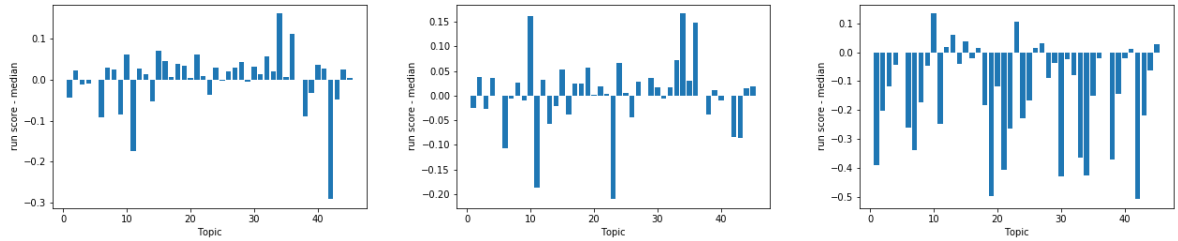
(a) PARADE **(b) PARADE$_{\text{Max}}$** **(c) PARADE$_{\text{Attn}}$**

Figure 1: Per-topic difference from median nDCG@10 for all runs



(a) PARADE **(b) PARADE$_{\text{Max}}$** **(c) PARADE$_{\text{Attn}}$**

Figure 2: Per-topic difference from median mAP for all runs



(a) PARADE **(b) PARADE$_{\text{Max}}$** **(c) PARADE$_{\text{Attn}}$**

Figure 3: Per-topic difference from median nDCG@100 for all runs

ACKNOWLEDGMENTS

This work was supported in part by Google Cloud and by the TensorFlow Research Cloud.

REFERENCES

[1] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv preprint arXiv:1611.09268v3* (2018).

[2] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations (ICLR 2020)*.

[3] Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2019)*. 985–988.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*. Association for Computational Linguistics, 4171–4186.

[5] Canjia Li and Andrew Yates. 2020. MPII at the NTCIR-15 WWW-3 Task. In *Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies*.

[6] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage Representation Aggregation for Document Reranking. *CoRR* abs/2008.09093 (2020). arXiv:2008.09093 https://arxiv.org/abs/2008.09093

[7] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond. *arXiv preprint arXiv:2010.06467* (2020).

[8] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[9] Ming Yan, Chenliang Li, Chen Wu, Bin Bi, Wei Wang, Jiangnan Xia, and Luo Si. 2019. IDST at TREC 2019 Deep Learning Track: Deep Cascade Ranking with Generation-based Document Expansion and Pre-trained Language Modeling. In *TREC*.

[10] Andrew Yates, Kevin Martin Jose, Xinyu Zhang, and Jimmy Lin. 2020. Flexible IR Pipelines with Capreolus. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 3181–3188.

Type	QID	Score - t	Query
t = median	1116380	0.2557	what is a nonconformity? earth science
	42255	0.2345	average salary for dental hygienist in nebraska
	67316	0.2340	can fever cause miscarriage early pregnancy
	911232	-0.3831	what type of conflict does della face in o, henry the gift of the magi
	1115210	-0.1983	what is chaff and flare
	1056416	-0.1771	who was the highest career passer rating in the nfl
t = 0	1116380	0.9716	what is a nonconformity? earth science
	1113256	0.9401	what is reba mcentire's net worth
	1064670	0.9235	why do hunters pattern their shotguns?
	673670	0.0000	what is a alm
	258062	0.1526	how long does it take to remove wisdom tooth
	1105792	0.2120	define: geon

Table 2: Queries solved best/worst by PARADE according to nDCG@10.

Type	QID	Score - t	Query
t = median	67316	0.3460	can fever cause miscarriage early pregnancy
	1105792	0.3380	define: geon
	42255	0.2345	average salary for dental hygienist in nebraska
	1136769	-0.2797	why does lacquered brass tarnish
	1119543	-0.2030	what does a psychological screening consist of for egg donors
	911232	-0.1554	what type of conflict does della face in o, henry the gift of the magi
t = 0	1116380	0.9266	what is a nonconformity? earth science
	1131069	0.9149	how many sons robert kraft has
	1136962	0.9071	why did the ancient egyptians call their land kemet, or black land?
	673670	0.0000	what is a alm
	1056416	0.2698	who was the highest career passer rating in the nfl
	258062	0.2914	how long does it take to remove wisdom tooth

Table 3: Queries solved best/worst by PARADE$_{\text{Max}}$ according to nDCG@10.

Type	QID	Score - t	Query
t = median	997622	0.2985	where is the show shameless filmed
	1116380	0.1413	what is a nonconformity? earth science
	1136769	0.1369	why does lacquered brass tarnish
	701453	-0.7443	what is a statutory deed
	324585	-0.6847	how much money do motivational speakers make
	911232	-0.6641	what type of conflict does della face in o, henry the gift of the magi
t = 0	1113256	0.9881	what is reba mcentire's net worth
	997622	0.9048	where is the show shameless filmed
	169208	0.8746	does mississippi have an income tax
	1136043	0.0000	difference between a hotel and motel
	324585	0.0000	how much money do motivational speakers make
	336901	0.0000	how old is vanessa redgrave

Table 4: Queries solved best/worst by PARADE$_{\text{Attn}}$ according to nDCG@10.