Evaluating Transferability in Retrieval Tasks: An Approach Using MMD and Kernel Methods

Supplementary Material

1. Experimental Details

1.1. Fine-Tuning Objective

Given a dataset's query-document pair denoted by (q, p) and the encoder f_θ, the relevancy score between each query and document is calculated based on their cosine similarity as follows:

$$s(q,p) = \frac{< f_\theta(q), f_\theta(p) >}{\|f_\theta(q)\|\|f_\theta(p)\|} \tag{1}$$

The final representation of a document is gained by averaging the final hidden representation of the transformer model. In order to train such models for retrieval tasks we use contrastive learning [1]. Given a query q, positive document p^+ (relevant document) and collection of negative documents $\{p_i^- | i \in (1, N)\}$ the contrastive loss function is computed as:

$$L(q,p^+,p^-) = -\log \frac{\exp(s(q,p^+)/\tau)}{\exp(s(q,p^+)/\tau) + \sum_{i=1}^{N} \exp(s(q,p_i^-)/\tau)} \tag{2}$$

1.2. Image

The optimization process leverages Adam optimizer [2] coupled with a cosine learning rate scheduler. We grid search learning rate from multiple choices in-between $(1e-5, 1e-2)$ for each model architecture on each dataset and report the best result trained up to 100 epochs. We present the fine-tuning results in terms of R@1 on each model and dataset combination as well as evaluations using RetMMD-S in Table 1. For each model and dataset, we randomly select 500 queries and run experiments 5 times. In Table 1 shows the number of significant p-values with the corresponding kernel choice in calculation of RetMMD-S. For each run 500 queries were randomly picked.

1.3. Text

The models were fine-tuned using a grid search approach, exploring learning rates of $1e-3, 1e-4, 1e-5, 2e-5$, and $1e-6$, schedulers warmuplinear and warmupcosine as well as warmup steps of $100, 500$, and 1000. Optimizer used was AdamW. After encoder model, mean pooling was done to generate same size embedding. Batch size was set to 50 and maximum sequence length to 300. The reported results for each model on each dataset are taken from the best performing model trained for up to 20 epochs. Fine-tuning results on FiQA-2018, FEVER and NFCorpus dataset are presented in Table 2, Table 3 and Table 4 respectively.

References

[1] Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*, 2017. 1

[2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014. 1

[3] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, 2013. 2

[4] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[5] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 2

Table 1. RetMMD-S evaluation on image datasets.

	CUB200 [5]			CARS196 [3]			SOP [4]		
Model	RetMMD-S	Kernel	R@1	RetMMD-S	Kernel	R@1	RetMMD-S	Kernel	R@1
ViT-Tiny	121.6 ± 6.37	Cosine	0.67	50.2 ± 5.19	Linear	0.47	419.4 ± 4.59	Linear	0.74
ViT-S	182.0 ± 6.36	Linear	0.76	85.4 ± 9.09	Linear	0.60	451.0 ± 8.27	Cosine	0.78
ViT-S-DINO	78.4 ± 7.31	Linear	0.74	41.2 ± 4.26	Sigmoid	0.49	352.2 ± 4.71	Linear	0.78
ViT-S-IN21k	143.4 ± 9.44	Linear	0.75	63.8 ± 8.16	Linear	0.58	440.4 ± 2.33	Cosine	0.78
ViT-B	200.0 ± 8.56	Linear	0.78	66.0 ± 4.56	Linear	0.67	473.2 ± 4.26	Linear	0.81

Table 2. Fine-tuning performance on the FiQA-2018 dataset

Model	MAP@1	MAP@5	MAP@10	Recall@1	Recall@5	Recall@10
distilbert-base-uncased	0.378	0.256	0.211	0.040	0.108	0.142
all-distilroberta-v1	0.396	0.268	0.216	0.041	0.112	0.153
nq-distilbert-base-v1	0.402	0.270	0.223	0.041	0.111	0.145
msmarco-distilbert-dot-v5	0.433	0.278	0.228	0.047	0.116	0.162
all-MiniLM-L6-v1	0.440	0.304	0.251	0.049	0.126	0.165
all-MiniLM-L12-v1	0.458	0.307	0.258	0.047	0.114	0.174

Table 3. Fine-tuning performance on the FEVER dataset

Model	MAP@1	MAP@5	MAP@10	Recall@1	Recall@5	Recall@10
distilbert-base-uncased	0.432	0.507	0.517	0.404	0.668	0.736
all-distilroberta-v1	0.503	0.580	0.589	0.470	0.748	0.812
nq-distilbert-base-v1	0.393	0.469	0.480	0.367	0.628	0.707
msmarco-distilbert-dot-v5	0.487	0.562	0.572	0.456	0.721	0.792
all-MiniLM-L6-v1	0.410	0.489	0.500	0.383	0.657	0.739
all-MiniLM-L12-v1	0.413	0.496	0.508	0.387	0.665	0.748

Table 4. Fine-tuning performance on the NFCorpus dataset

Model	MAP@1	MAP@5	MAP@10	Recall@1	Recall@5	Recall@10
distilbert-base-uncased	0.393	0.258	0.211	0.043	0.102	0.143
all-distilroberta-v1	0.430	0.275	0.227	0.049	0.113	0.154
nq-distilbert-base-v1	0.418	0.293	0.235	0.049	0.113	0.149
msmarco-distilbert-dot-v5	0.424	0.270	0.217	0.049	0.115	0.149
all-MiniLM-L6-v1	0.471	0.309	0.256	0.055	0.128	0.170
all-MiniLM-L12-v1	0.486	0.314	0.268	0.051	0.125	0.182