SeeTek: Very Large-Scale Open-set Logo Recognition with Text-Aware Metric Learning

– Supplementary Material –

Chenge Li[1], István Fehérvári*[2], Xiaonan Zhao[1], Ives Macedo*[2], Srikar Appalaraju[1]

[1]Amazon
[1]{lichenge, xiaonzha, srikara}@amazon.com
[2]istvan@fehervari.org, research@ivesmacedo.com

We present more experiment results and ablation studies in this supplementary material. More qualitative results are shown in figure 1 and figure 2.

1. End-to-end Evaluation on Public datasets

Using the confidence score defined in section 4.3.2 in the paper, we compute the mAP scores across all unseen brands of the 6 datasets' test splits. The result is shown in table 1. Our proposed SeeTek model outperforms [2] by a large margin in all of the datasets. Noticeably on LogoDet-3K[4], mAP increased from 85.48% to 94.45%.

2. Use Text Predictions Explicitly: Rerank after KNN Retrieval

An obvious advantage after we augment the model with text recognition branch is that, we can get text predictions for free. When using the concatenated visual-textual embeddings, we are using the text information implicitly, letting the model decide which modality to trust more. There can be some interesting future research on how to weight different modality differently, how to make multi-modality fusion more interpretable. In this paper, we did an experiment by reranking the predictions using the predicted brand name strings. During retrieval, we first relax the retrieval by using KNN ($K > 1$, e.g. $K = 16$). Among these K retrievals, we then compare the text predictions from the query image and the retrieved anchor images explicitly using edit distance. The K retrievals are reranked based on edit distance and we compute the Recall@1 after the reranking. The relaxed Recall@16 performance is the upper bound for the reranking performance. As show in table 2, for most of the datasets, reranking could bring a small boost of ~0.5% in performance. It doesn't hurt the overall performance, especially when the logos are text-heavy.

Predicting brand names without clean ground truth labels are challenging. There is still room for scene text recognition with weak or noisy text supervision signals, and room for improvement for multi-modal information fusion than simple embedding concatenation, both we will defer to future research.

3. Ablation Study

3.1. Embedder-level Comparisons on Other Public Datasets

Similar with Table 6 in section 4.4 in the paper, we examined the retrieval performance when using the learnt textual embedding from the text branch only, visual embedding from the visual branch only and the concatenated visual-textual embeddings. We reported the comparison on PL8K(our dataset) and on LogoDet-3K in the paper. Here in table 3, we show the comparison on other 4 public datasets. Similarly with observation in PL8K and LogoDet-3K (table 6 in paper), using textual embedding only works for text-heavy logos. It's overall worse than using visual embeddings only. However, when concatenating these two complementary information together, the visual-textual embeddings gives the best performance.

SeeTek model outperforms visual-only model[2] on all datasets. Furthermore, visual embedding with size 512 from the SeeTek model's visual branch outperforms the visual embedding with size 1024 from the single model[2]. This shows that training jointly with text supervision helped the visual branch to attend to logo regions and improved its performance as well.

*Work done while at Amazon.

	PL8K				LogoDet-3K	FlickrLogos-47		LitW	OpenLogo	BelgaLogos	
	Misc	WDT	WT	NT			Text				Text
[2]	94.66%	94.90%	90.35%	91.76%	85.48%	68.39%	83.19%	79.14%	77.79%	93.20%	100%
SeeTek (ours)	**98.06%**	**98.24%**	**93.46%**	**95.23%**	**94.45%**	**69.51%**	**89.16%**	**83.83%**	**84.87%**	**95.98%**	100%

Table 1. mAP comparison on all 6 datasets. SeeTek model is using Attention text prediction head, trained on LogoDet-3K train split.

	PL8K				LogoDet-3K	FlickrLogos-47		LitW	OpenLogo	BelgaLogos	
	Misc	WDT	WT	NT			Text				Text
[2]	95.09%	91.20%	92.78%	94.05%	87.39%	91.88%	91.33%	81.87%	83.14%	96.09%	100.00%
SeeTek (ours) - Recall@1	98.37%	94.80%	95.56%	95.71%	94.90%	91.88%	94.00%	87.72%	89.64%	97.39%	100.00%
SeeTek (ours) - Rerank	98.49%	95.00%	96.11%	95.48%	95.14%	90.31%	94.67%	87.89%	89.27%	97.39%	100.00%
SeeTek (ours) - Recall@16	99.13%	97.00%	98.89%	97.62%	97.77%	98.44%	100.00%	95.04%	96.50%	99.57%	100.00%

Table 2. Rerank 16NN retrieval using text predictions explicitly. When text predictions are more reliable (text-heavy logos), the performance gets further boosted. SeeTek model is using Attention text prediction head.

3.2. Text Recognizer Baseline

We report the retrieval performance using the text recognizer trained from MJSynth dataset [3] as well. This serves as a baseline for the two-branch SeeTek model. As shown in table 4, it is clear that scene text recognition on logo regions are much harder than synthetic dataset, and using text recognizer alone is not sufficient for logo retrieval problems. Also notice that after we finetune the text branch using brand names as noisy labels, the text branch's performance improved a lot from text recognizer baseline. Attention-based text prediction head outperforms CTC text prediction head by a large margin, especially on PL8K-WT(word trademarks) split, from 35.56% to 82.78%. It's interesting though when combined with visual signal, the advantage of Attention head is not very obvious anymore. We suspect that this shows the current multi-task model relies more on visual features and future research is needed for better interpretability in multi-modal information fusion.

3.3. Masking Out Text Regions

In our PL8K dataset, we have word trademarks (WT), word design trademarks (WDT), no-text trademarks (NT) and other mixed miscellaneous logos (Misc). We did another experiment by masking out the text regions in these logo images using an off-the-shelf text detector CRAFT[1]. We first used CRAFT to detect all text bounding boxes, then we filled the bounding box region with the average RGB values of the image. By masking out the text regions, the text branch is practically disabled, but visual branch can still extract useful information from the design if there are any. From table 5, we can see that masking hurts WT logos the most, as it contains the most text. NT (no-text logos) gets affected the least, as they don't contain any text. Textual branch gets affected more than visual branch. Since masking deleted a lot of visual information and introduced artifacts such as harsh edges, it also hurts visual branch. As show in the table, both single model [2] and SeeTek's visual branch is affected.

The performance of 1NN Retrieval using textual embedding drops drastically for WT logos, from 82.78% to 15.56%. As a result, Visual-Textual embedding almost falls back to visual embedding only. This experiment verifies the contribution of the text branch in the reverse way.

3.4. Mirror the Images

We did another experiment by mirroring (horizontal flipping) all the images during testing. When we train the text recognizer and text branch, we didn't add mirroring into the data augmentation scheme. Hence similar to masking out the text regions, this artifact also limits the text branch's performance. From table 5, we can see that performance of SeeTek's textual embedding drops a lot, especially for WT logos, from 82.78% to 60%, while the visual embedding performance almost kept the same, from 95.56% to 92.78%. Overall, SeeTek model is more robust to the mirroring artifacts than single model, with 1.67% performance drop compared with 2.78% drop on WT split. This may show that the text branch also contains visual features complementary to the visual branch, though it was originally designed for text recognition.

4. Conclusion

We showed more ablation studies in this supplementary material to further inspect the improvement of the proposed model and the contribution from text branch and multi-task training. Very large scale logo recognition is a very practical and important problem. Logo text recognition is much harder than other scene text recognition tasks given its highly diverse nature and lack of fine-grained high-quality annotation data. We hope more research work related with deep metric learning, scene text recognition, multi-modality fusion etc. will push the field even further.

Recall@1		Text Pred Head	FlickrLogos-47		LitW	OpenLogo	BelgaLogos	
				Text				Text
[2]	Visual only model	—	91.88%	91.33%	81.87%	83.14%	96.09%	100.00%
SeeTek (ours)	Visual branch only	—	**93.13%**	92.67%	87.48%	89.05%	96.52%	100.00%
	Textual branch only	CTC	55.94%	81.33%	61.63%	61.23%	79.57%	100.00%
	Textual branch only	Attention	48.75%	82.00%	63.98%	62.86%	69.57%	93.33%
	Visual-Textual embedding	CTC	90.62%	92.00%	84.47%	86.32%	94.78%	100.00%
		Attention	91.88%	**94.00%**	**87.72%**	**89.64%**	**97.39%**	**100.00%**

Table 3. Model generalization and ablation study: Recall@1 performance on public datasets from SeeTek model trained on LogoDet-3K[4] train split.

Recall@1		Text Head	PL8K				LogoDet-3K
			Misc	WDT	WT	NT	
Text Recognizer trained from MJSynth[3]	-	CTC	54.82%	**52.60%**	45.56%	31.90%	65.98%
	-	Attn	**57.03%**	51.40%	**61.11%**	**35.24%**	**68.45%**
SeeTek(ours)	Textual branch only	CTC	55.53%	50.20%	35.56%	51.19%	77.27%
	Textual branch only	Attn	**86.48%**	**78.60%**	**82.78%**	**60.71%**	**81.84%**
	Visual-Textual	CTC	98.13%	94.20%	**97.78%**	94.52%	93.66%
	Visual-Textual	Attn	**98.37%**	**94.80%**	95.56%	**95.71%**	**94.90%**

Table 4. Ablation study: Recall@1 on PL8K(ours) and LogoDet-3K[4] using text recognizer, or using different embeddings and text prediction heads. Attention-based text prediction head outperforms CTC text prediction head by a large margin, especially on PL8K-WT(word trademarks) split with textual branch embedding only.

Recall@1		PL8K			
		Misc	WDT	WT	NT
Original images	[2]	95.09%	91.20%	92.78%	94.05%
	SeeTek	**98.37%**	**94.80%**	**95.56%**	**95.71%**
	SeeTek (Visual only)	97.86%	94.00%	95.56%	94.29%
	SeeTek (Textual only)	86.48%	78.60%	82.78%	60.71%
Masking out text regions	[2]	95.12%(not masked)	68.40%(-22.80%)	53.89%(-38.89%)	85.95% (-8.10%)
	SeeTek	98.35%(not masked)	65.80%(-29.00%)	50.00%(-45.56%)	87.38%(-8.33%)
	SeeTek (Visual only)	97.84%(not masked)	67.20%(-26.80%)	50.00%(-45.56%)	85.95%(-8.34%)
	SeeTek (Textual only)	86.46%(not masked)	25.20%(-53.40%)	15.56%(-67.22%)	50.48% (-10.23%)
Mirroring all images	[2]	94.98%(-0.11%)	90.20%(-1.0%)	89.44% (-3.34%)	92.62%(-1.43%)
	SeeTek	97.91%(-0.46%)	94.20%(-0.60%)	93.89%(-1.67%)	95.00%(-0.71%)
	SeeTek (Visual only)	97.64%(-0.22%)	94.00% (-0.0%)	92.78% (-2.78%)	95.24% (+0.95%)
	SeeTek (Textual only)	76.88% (-9.58%)	71.80% (-6.80%)	60.00% (-22.78%)	56.67% (-4.04%)

Table 5. Recall@1 on PL8K dataset with different data artifacts. All SeeTek model variants are using Attention text prediction head trained on PL8K train split.

References

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9365–9374, 2019. 2

[2] I. Fehérvári and S. Appalaraju. Scalable Logo Recognition Using Proxies. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 715–725, 2019. 1, 2, 3

[3] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. In *Workshop on Deep Learning, NIPS*, 2014. 2, 3

[4] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image

Figure 1. Sample images from the PL8K dataset showing that our text-aware logo recognition approach accurately detects the right logo for a given query-logo. Left column: query image, middle column: vision-only model incorrect top1 retrieval, right column (ours): text-aware correct top1 retrieval

dataset for logo detection. *arXiv preprint arXiv:2008.05359*, 2020. 1, 3

Figure 2. Sample images continued.