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

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

Abstract

Recent advances in deep learning and computer vision have set new state of the art in logo recognition [2, 9, 36]. Logo recognition has mostly been approached as a closed-set object recognition problem and more recently as an open-set retrieval problem. Current approaches suffer from distinguishing visually similar logos, especially in open-set retrieval for very large-scale applications with thousands of brands. To address the problem, we propose a multi-task learning architecture of deep metric learning and scene text recognition. We use brand names as weak labels and enforce the model to simultaneously extract distinct visual features as well as predict brand name text. To achieve it, we collected a dataset with 3 Million logos cropped from Amazon Product Catalog images across nearly 8K brands, named PL8K. Our experiments show that adding the task of text recognition during training boosts the model's retrieval performance both on our PL8K dataset and on five other public logo datasets.

1. Introduction

Logo recognition is a well-known problem in computer vision with many practical applications, such as product search and recommendations in e-commerce, compliance or authenticity verification or brand presence tracking etc [2, 9, 10]. There are several challenges around detecting logos such as lack of a precise definition of what exactly constitutes a logo, lack of large-scale (i.e. number of brands) well-annotated (i.e. bounding-box for each image) logo datasets, large intra-class variations (i.e. a logo can be stylized text or an abstract figure or a mixture of both), large background, occlusions and illumination changes etc. Nevertheless, recent advances in deep learning and computer vision have set new state of the art results in logo detection either by decoupling detection and recognition [2, 9, 36], or by using weakly-labeled images crawled from the web [30].

Figure 1. Examples of visually very similar logos belonging to different brands. Without text recognition, it's very challenging to differentiate them.

As many logos have considerable amount of text or (stylized) letters, optical character recognition (OCR) or scene-text detection could potentially be used to augment existing logo detection pipelines. However, this requires at least one more model or system to be trained or fine-tuned on logo-specific data, which introduces in practice additional overheads assuming such text-annotated logo data even exists. Thus, it is desired to build a single model that performs the logo-classification task taking the textual content into account while adapting nicely when text is not present.

In logo recognition, there are typically two steps: a) a class-agnostic universal logo detector proposing regions that potentially contain logos, and b) a logo classifier operating on the proposed regions to predict brands. Unlike *closed-set* setting, *open set* logo recognition allows users to detect logos that are *not available* during the training stage. Hence classification is based on nearest neighbor search on the learnt latent space, instead of from a classifier with fixed number of classes. Recent works on object and logo de-

*Work done while at Amazon.

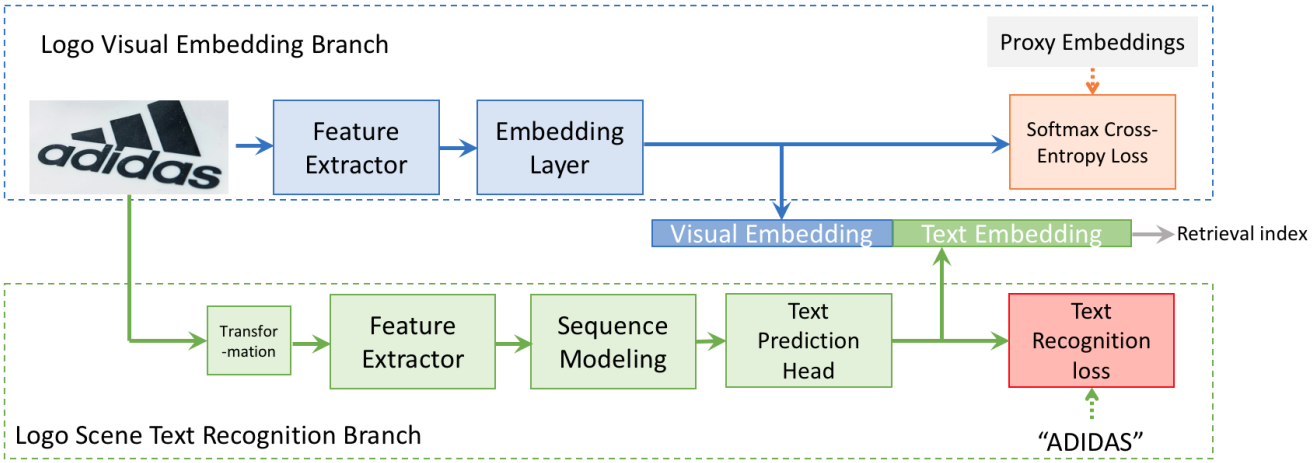

Figure 2. Our proposed multi-modal **SeeTek** model architecture: **Top**: Visual embedding branch and **Bottom**: Scene-text recognition branch. The learnt visual embedding and text embedding are concatenated together to be used as retrieval index.

tection demonstrated that *open set* approaches are on-par or better than dedicated single-model solutions (e.g. using Faster R-CNN or YOLOv3) [9, 18]. In this paper we focus on learning a good latent space that generalizes on novel *open set* brands.

In order to improve recognition performance, especially on text-heavy logos, we propose a deep metric learning model that is trained with two tasks: classification-based metric learning, and scene text recognition. This joint model can learn two decoupled latent spaces: visual space and textual space, which are complementary to each other during retrieval. Further, we use brands' names as a weak labeling mechanism to bridge the lack of fine-grained text annotations. Even with this very noisy brand name labels, we find that the joint model's performance gets boosted especially on text-heavy logos, which are the most challenging ones for existing DML based logo approaches [9].

In order to advance the research in this field we collected a large amount of logo regions cropped from product images for 7,888 brands. We call this dataset PL8K and use it to train and to assess the performance of our proposed models. The contributions of this paper are the following:

- We propose to tackle the logo recognition problem with decoupled latent spaces from two modalities.
- We show that brand names can be used as weak labels to improve logo recognition accuracy, especially on logos that bear minimal styles or text-heavy, while not regressing on logos that have no text content.
- We propose a novel multi-task deep metric learning architecture named **SeeTek**[1] to recognize brand logos in an open-set setting.
- We introduce a new product logo dataset called PL8K containing 3,017,146 logo regions across 7,888 brands.

2. Related Work

In this section we discuss closely related works in the fields of deep learning in computer vision, prior art in logo, text recognition, and metric learning.

2.1. Current Logo Recognition Pipeline

Logo recognition is a very challenging problem in computer vision. In the recent past, quite a few deep-learning based approaches have been published [2–4, 14, 16, 24, 30, 33, 34, 36]. These approaches primarily use a logo-region bounding box localizer followed by the recognition step. Most of the work treat the recognition step as closed-set classification problem, with fixed number of classes, not being able to scale up to thousands of brands easily. What's more, closed-set classification model doesn't generalize to novel brands without retraining.

In [31], a large logo dataset (WebLogo-2M) automatically crawled from Twitter was introduced. The authors used synthetic copy-and-paste logo data to train two complementary object detectors YOLOv2 and Faster R-CNN and utilized the their object scores to future bootstrap each model's training. Though WebLogo-2M doesn't need manual annotation effort, the collected dataset is very noisy with a lot of false positives. The performance comparison was

[1] Putting visual (see) and text recognition (tek) into multi-task learning, hence the codename.

only done for the 194 brands as in closed-set object detection. The trained object detector was not able to detect novel unseen brands, which limits its real scalability in real world.

Our work is mostly related with [2] and [9], where we divide the end-to-end logo recognition pipeline into two steps: firstly a general brand-agnostic logo detector detects candidate logo regions with high recall, and secondly but more importantly, the heavy-lifting of brand classification is done through a logo ROI feature embedder. Both [2][9] and our work tackle the open-set problem, where none of the test brand class is seen in training. In [2], a simple deep metric learning (SDML) framework with hard negative mining was proposed. However, the model was only evaluated on 1500 pairs of images with 248 classes.

2.2. Metric learning

Distance Metric learning (DML)[9,13,38] has a very rich research history. DML has great advantage when doing retrieval in open-set setting for never-seen classes. Its class-number-agnostic feature also makes it ideal for scaling up to more classes easily. In practice, the performance of DML methods depend heavily on data mining strategies as there are exponential number of pairs (mostly negative pairs) that can be generated. Therefore, in this paper we propose to build upon metric learning losses that require no sampling [9,38], where it's shown that metric learning with learnable (proxy) embeddings can achieve state of the art results if the task is formulated as a classification problem.

2.3. Scene Text Recognition (STR)

Recognizing text in natural scenes has attracted a lot of interests and many new works are introduced in recent years [1,6,17,21,22,28,29,37,39]. Most of these papers focus on how to transform or regularize the text region to be better aligned and hence improve the recognition accuracy. In [1], a unified four-stage STR framework was summarized as the standard recognition pipeline that most existing STR models fit into. We adopted this popular architecture as our text recognition branch.

Tying these prior art with our current contributions, we put DML into a multi-task learning framework. We employ text-aware loss along with metric loss and train two branches jointly. With the help of synthetic text dataset[17] and weak class labels, all these contributions lead to state-of-the-art logo detection results as shown in Sec. 4.

3. Our Method

3.1. Sampling-free Classification based Deep Metric Learning

Open-set means that the testing classes cannot be used during training, and number of classes are not fixed. It is a common problem in industry, as it's not economically efficient to retrain a classifier whenever a new logo is added. There are, however, very limited related works on open-set logo retrievals. One good solution to open-set setting is deep metric learning, which tries to learn a mapping of input images to a latent space that preserves the semantic similarity of the samples [9], from which we can use nearest neighbor classifier to make predictions. DML is usually trained with triplet constraints[13], which uses a hinge function to create a fixed margin between the intra-class and inner-class distances. The main issue with such triplet(positive-anchor-negative)-based metric learning losses is that the sampling strategy is crucial. Too easy triplets do not contribute to the loss, while very hard ones could destabilize the train-

Figure 3. 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**: previous non-text aware model's incorrect retrieval (note the visual similarity with query logo) , **right column (ours)**: Despite the large visual dissimilarity, our proposed text-aware model correctly retrieved logos, thanks to the help from text recognition.

Figure 4. Examples of logos with text. There can be other characters other than brand names. Brand names can be only partially visible, or with artsy design fonts which are hard to recognize. Brand name: EVGA, MIZUNO, BEYOUNG, CRXOOX, KFD, HARIBO, KEENSTONE, CALDWELL.

ing procedure which makes adding extra objectives on top of the learned embeddings a very hard task. For example, [2] relies on hard example mining sampling to fetch better triplets during training. The complex training strategy is slower and not easy to scale up to larger number of classes. In [38], a metric learning loss that require no sampling was introduced. It showed that metric learning with learnable (proxy) embeddings can achieve state of the art results if the task is formulated as a classification problem.

With the advantage of not relying on large batch size and more suitable for few-shot learning scenario, in this paper we propose to build upon [9,38], our base DML architecture is then trained with the following loss:

$$L_{dml} = -\log\left(\frac{\exp\left(\frac{x^T p_y}{\sigma}\right)}{\sum_{z \in Z} \exp\left(\frac{x^T p_z}{\sigma}\right)}\right) \quad (1)$$

where x is an L2-normalized embedding corresponding to the output of the last linear layer of our model. y is the class label of x of all possible classes Z, and p_y is its respective proxy embedding. The temperature parameter σ is used to scale the logits to emphasize the difference between classes, thus boosting the gradients [8, 23, 35].

3.2. Brand Name Weak Supervision and Synthetic Text Pretraining

In logo design, there is a commonly accepted visual language (e.g. human interpret blue-color with *trust*, red with *power*) [26], which leads to visually similar designs. A huge amount of logos share similar visual appearance if we ignore the text content (see Figure 1). Thus it brings a great challenge for non-text-aware models to differentiate them. On the other hand, models trained solely with DML loss depend heavily on the diversity of the samples in the batch and the diversity within each brand, which is often not guaranteed. Especially in the few-shot scenario, intra-class distances for visually similar logs are often larger than inter-class distances. Lacking in harder, more diversified training samples causes DML model's performance to saturate. A previous work used word semantics from product descriptions as a weak supervision in the loss functions[40]. However, word semantics are ambiguous in brand names (e.g. Apple is not close to Samsung).

With this motivation, we propose to add text modality and combine deep metric learning with text recognition into a single model to improve logo recognition accuracy. As text information is agnostic to visual designs, text semantics are complementary to visual semantics in the latent space.

To our best knowledge, there is no logo dataset with fine-grained text labels available. As shown in figure 4, the brand name and real texts inside the logo ROI can be very different. For example, General Motors' logo only contains "GM" rather than the full name. Most of the texts have artsy design fonts, the text sequence alignment direction is not always horizontal, some logo texts are only partially visible due to the curvature of product surfaces. Despite all these challenges, observation of a large number of brand logos convinced us that brand names could act as a good proxy for the missing exact labels.

However, introducing new text-based objectives on top of a learnt visual latent space with end-to-end training is a non-trivial task due to the stability and convergence concerns in training deep metric learning models. We found that the text branch is very hard to converge if directly trained with brand names from scratch. To ease training with noisy brand names, we first used the synthetic text dataset MJSynth [17], which contains 9 million images covering 90k English words, to pretrain the text branch. Note that our model is not limited to English logos, any English alphabet based logos will benefit from this text awareness. We further finetune the pretrained text branch using brand names, and evaluated the text recognition performance using string-level accuracy and character-level accuracy. String-level accuracy is based on exact string matches and character-level accuracy Acc_char is defined using edit distance:

$$Acc\_char = \frac{||gt| - d_{edit}(pred, gt)|}{len(gt)} \quad (2)$$

As shown in table 3, the text branch can learn a decent performance of text recognition on logo datasets such as PL8K and LogoDet-3K[36]. Note that the ground truth brand names are noisy labels for evaluation, which is unfair for the model. Nevertheless the performance is still far from ideal and we defer better logo text recognition in future research.

3.3. Text Prediction Head

Following [1], we explored two text prediction heads: 1. text prediction trained using CTC[11] loss and 2. text prediction trained using Attention[6, 29]. For CTC head, we use the un-pooled features before the embedding layer and feed them into a two-layer bi-LSTM modules followed by a linear layer to predict the characters. For attention

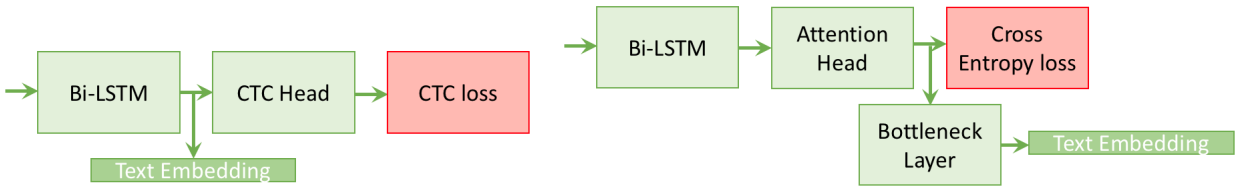

Figure 5. **Left**: Text prediction head using CTC loss, and **Right**: Text prediction head based on Attention.

Dataset	Logos	Images	Annotation	Noisy	Public	Scalability
TopLogo-10 [32]	10	700	Object-Level	✗	✓	Weak
BelgaLogos [19]	37	10,000	Object-Level	✗	✓	Weak
FlickrLogos-32 [27]	32	8,240	Object-Level	✗	✓	Weak
FlickrLogos-47 [27]	47	8,240	Object-Level	✗	✓	Weak
Logo-NET [15]	160	73,414	Object-Level	✗	✓	Weak
WebLogo-2M [31]	194	2,190,757	Image-Level	✓	✓	Medium
QMUL-OpenLogo [33]	352	27,083	Image-Level	✗	✓	Medium
Logos in the wild [34]	871	11,054	Object-Level	✓	✓	Medium
BLAC [2]	2,800	6,200	Object-Level	✗	✗	Medium
Logo Detection 3K [36]	3,000	158,652	Object-Level	✗	✓	Strong
PL8K (Ours)	7,888	3,017,146	Object-Level	✗	planned	Strong

Table 1. Statistics and characteristics of existing logo detection datasets.

	No. of images				No. of brands			
	Misc	WDT	WT	NT	Misc	WDT	WT	NT
Training	1,875,974	36,212	12,281	-	4,944	86	29	-
Validation	467,624	9,790	3,404	7,844	1,211	22	8	21
Testing	569,630	16,316	4,427	13,644	1,507	28	11	21

Table 2. PL8K data splits for train and test sets of miscellaneous logos, word design trademarks, word trademarks, and no-text trademarks, respectively.

		MJSynth	PL8K	LogoDet-3K
CTC Head	string acc	74.64%	20.15%	12.79%
	character acc	93.10%	54.17%	50.31%
Attention Head	string acc	**91.11%**	**33.81%**	**19.24%**
	character acc	**98.18%**	**70.17%**	**58.90%**

Table 3. Text recognition performance. MJSynth dataset has exact character-level ground truth labels. PL8K and LogoDet-3K only have brand names as noisy ground truth labels.

head, we added a fully-connected bottleneck layer after bi-LSTM to get the text embeddings. Figure 2 demonstrates the full SeeTek model architecture. Figure 5 shows the two types of text prediction head in text recognition branch. The text branch was trained either with Connectionist Temporal Classification (CTC) loss or cross entropy loss.

CTC heads predict a whole sequence with characters stuffed with spaces, while the Attention head learns to predict a START token and an EOS token. Real predictions are squeezed between START and EOS and all other characters after the EOS or before START are dropped. From our experiments, text recognition with attention head performs much better than CTC head, as shown in table 3. The text recognition performance also directly influenced the retrieval performance if we only use the learnt text embedding from the text branch as retrieval index. See section 4.4 for detailed ablation studies.

4. Experiments

We evaluate the impact of our proposed methods by comparing to current state-of-the-art on our dataset and 5 other public benchmark datasets. Our experiments cover strong DML based scalable logo recognition baselines and different variants of our method. We report both the end-to-end evaluation results (universal logo detector followed by proposed DML based embedder) and the embedder-level performance to provide more insights.

4.1. Implementation

For visual branch's feature extractors we used ImageNet pretrained ResNet50 [12] with input size of 150x150. The text branch used a custom residual network designed for scene text recognition tasks introduced in [1]. This architecture uses BasicBlocks [12] in a configuration of [1, 2, 5, 3] with twice as many filters per layer (128 256, 512, 512) compared to ResNet50, and double strides along the

height dimension after layer3. The standard input size of this architecture for scene-text tasks is 100x32, but we used 150x150 to provide a more fair comparison and to gain insights on the importance of resolution. We used SGD, momentum of 0.9, and learning rate of 0.01 with exponential decay for 40 epoches.

The output embedding size of visual branch was fixed to 512. All embeddings were L2 normalized. For the text recognition branch we used bi-LSTM modules with a hidden layer size of 256. The output embedding dimension of text branch is also 512, hence the concatenated visual-textual embedding size is 1024. We used a temperature scaling parameter σ of 0.05 in the normalized softmax loss [38].

During testing, we form queries from the first 10 images of each class(logo) and merge the remaining ones into a large gallery set. Note that all testing classes are not seen in training. The model's performance is assessed with the average recall@1 across all queries. Given visually very similar logos most likely belong to very different types of brands due to trademark legislation, we omit calculating recall@k when $k > 1$ since in a practical scenario these are considered as failure cases.

4.2. Datasets

Our experimental evaluation is built upon 6 datasets constructed from different sources and domains. Our main dataset (PL8K) was gathered from product images of online Amazon retail listings and is used to both train our models and to compute performance metrics on in-domain imagery. In addition to PL8K, we use the following as test-sets: the recent LogoDet-3K [36], popular FlickrLogos-47 [27], BelgaLogos [19], Logos-in-the-Wild [34], and the QMUL-OpenLogo [33] datasets, to assess the performance of our models on (and robustness to) out-of-domain images. We omitted the WebLogo-2M [31] dataset from our evaluation since it does not provide bounding boxes which we require to create the logo regions for embedder-level evaluation. Besides PL8K, we have also collected a hard evaluation dataset containing high-confidence false positives from other model variants to stress test the model. We provide some specifics on these datasets in the summaries below, and a quick overview is presented in Table 1.

PL8K (Our dataset) This large logo recognition dataset was built in a semi-automated fashion. We trained a universal logo detector to identify and extract crops of potential logos from product images, with high recall. Once the crop was extracted, it was sent for labeling by human auditors who would assign the crop to the proper brands. Our overall dataset construction followed the methodology proposed in [9], the main differences being the scale in which we collected our dataset and the fact that we further annotated our logos to identify large subsets of word trademarks (WT) and word design trademarks (WDT). We also made sure that we have at least 20 images per brand (discarding those brands with fewer identified crops). This resulted in a dataset with 3,017,146 product logo images covering 7,888 brands, hence the reason we call this dataset PL8K. Table 1 provides a high-level comparison of PL8K with popular logo datasets. To make sure our in-domain evaluations reflect scenarios with new logos, our training/validation/testing sets splits were performed at the brand-level and followed a stratified strategy for each group of WT, WDT, and other miscellaneous logo images: the brands were sorted by number of images and, for every adjacent and disjoint group of 5 brands, one brand was randomly sampled to build the testing set. The remaining brands would be sampled once more to build the validation set with the final remaining brands forming the training set. This procedure created a split were the groups were not overly imbalanced and the final groups contained roughly 64% brands for training, 16% for validation, and 20% for testing. Table 2 provides details of these splits. Notice this dataset does not include non-logo images, since its primary goal being building and evaluating logo recognition models, not logo detection.

LogoDet-3K LogoDet-3K[36] is a recently published dataset with 3K brands and 200K manually annotated logo ROIs on ~160K images images. This is a very challenging dataset where logos are in natural images captured in the wild. The logo ROIs have a wider variety and larger intra-class distance than other public datasets. We split the LogoDet-3K into a training (80%) and testing dataset (20%). Similar to PL8K, we finetune our SeeTek model on the training dataset, with visual branch pretrained on ImageNet[7] dataset and textual branch pretrained on MJSynth[17].

FlickrLogos-47 In order to assess model performance on imagery from a completely different domain without fine-tuning, we test our model on the popular FlickrLogos-47 public dataset [27]. This dataset is provided with bounding-boxes and mask annotations covering 47 logo classes with a mix of symbol logos (32) and text logos(15). We process this dataset to retain only crops of images with logos in them, leveraging the bounding box annotations in each image. This results in a dataset consisting of 5,968 image crops of logos (1,936 coming from the original training set, 573 of which being textual, and 4,032 from the original test set, of which 1,241 are text). All the nearly 6K crops are used to compute our performance metrics in this out-of-domain dataset, with the 1.8K textual ones serving to assess impact on textual trademarks.

Logos-in-the-Wild (LitW) One of the largest publicly available logo datasets is Logos-in-the-Wild [34] featuring 11,054 images across 871 brands. Unfortunately, only the image URLs are released from which 4,614 are no longer accessible. This also caused several brands to be filtered

Approach	mAP
LogoDet-3K (YOLOv3, closed-set)[36]	52.28
YOLOv4+SeeTek (open-set)	**70.46**

Table 4. End-to-end performance comparison with closed-set logo recognition on LogoDet-3K.

universal logo detector	embedder	mAP	mAP (soft thresholding)	mAP (hard thresholding)
YOLOv3@0.1	[9]	59.43	24.59	77.65 (OOD 82.52%)
YOLOv3@0.1	SeeTek	**62.21**	**53.14**	**79.45** (OOD 31.38%)

Table 5. End-to-end performance comparison on the hard evaluation dataset.

out due to insufficient number of images. Nevertheless, we used the remaining data (6,440 images across 123 brands) to evaluate our model.

OpenLogo The QMUL-OpenLogos dataset[33] is a medium-sized logo dataset featuring 27,083 images for 352 brands. We merged all splits that contained bounding-box level annotations into a query and a gallery set as explained earlier.

BelgaLogos BelgaLogos dataset [20] is very similar to FlickrLogos47 in terms of its size (10,000 images for 37 brands). Unfortunately, several brands have very few imagery causing a reduction to only 23 brands. The dataset however marks 3 brands as 'text': Adidas, Citroén, and Puma. We create a separate split for these similar to FlickrLogos47.

Hard Evaluation dataset (our dataset) All datasets mentioned above contains true logo images. When evaluating against one brand's performance, all other brands are negatives. This evaluation setting doesn't consider non-logo regions as negatives. In real-world application, the universal detector could return non-logo regions with high confidence and this region could potentially look very similar with a real logo. Therefore in order to reflect the real-world performance, we collected another hard dataset containing 33,410 product images, collected from Amazon Product Catalog as well. We trained a universal logo detector based on YOLOv3[25]. We then cropped all the ROIs with box confidence greater than 0.1. This resulted in a hard evaluation dataset of 53,774 ROI images over 68 brands, which contains a lot of highly confusing false positive regions with no logos.

4.3. End-to-end Evaluation

4.3.1 Comparison with closed-set methods

First, we report the results in end-to-end evaluation, where we compare the DML based open-set logo recognition pipeline with a closed-set object recognition pipeline. For open-set logo recognition, we use a universal logo detector and our proposed SeeTek logo embedder, and KNN 1st nearest neighbour as the predicted classes. We compare our model with a state of the art closed-set approach introduced in the LogoDet-3K [36] dataset paper. This work uses a single YOLOv3[25] model trained on 3000 classes using Focal Loss and optimized anchor boxes and has shown to outperform related works by a large margin. Comparing closed-set and open-set logo recognition poses a few challenges: the former requires to split the dataset across images of each class while the latter one splits across classes (to make sure test classes are unseen during training). Therefore for fair comparison, we split the LogoDet-3K dataset in an 80%/20% train/validation manner with the number of classes being 2400/600. We trained a class-agnostic YOLOv4 [5] logo detector on the train split for 30 epochs with all other settings set to default from the MS-COCO experiments. Similarly, we used the train split only to train our SeeTek embedder. Due to the low number of images per class in LogoDet-3K we used half of the test split as query and the other half as gallery during retrievals. When computing the mAP we used the objectness score from the detector as prediction probability (i.e. the 1st nearest neighbor was considered with 100% confidence). As seen in the table 4, our approach largely outperforms the best performing LogoDet-3K model.

4.3.2 Comparisons on the hard evaluation dataset

We compare SeeTek with Attention head to open-set logo recognition pipeline proposed in [9]. Note that in this setting, we use the same YOLOv3 based universal logo detector model and pass all bounding boxes with confidence larger than 0.1 to the embedder. Instead of treating the 1st nearest neighbour with confidence of 1, we define a naive confidence score using the inverse of the hamming distance between the binarized query feature vector and binarized retrieval feature vector, normalized by the feature dimension. We found that when feature dimension is large (in our case 1024), the performances gap between float-number feature embeddings and binary emeddings are negligible. The information is encoded into signs instead of real values. This observation is also reported in [38]. Here we define the naive confidence score:

$$s = \frac{d_{Hamming}(V_{query}, V_{1NN})}{|V_{query}|} \tag{3}$$

The mAP comparisons are shown in table 5. We also compare the model performance with a precision guard, using a pre-defined out-of-distribution(OOD) distance threshold, in this experiment 170. For soft thresholding, when the distance between the query vector and the top nearest vector is larger than 170, we treat the matching distance to be 1024 (feature vector length). For hard thresholding, we simply throw away all predictions with 1NN distance larger than

Recall@1	Text Head	PL8K				LogoDet-3K
		Misc	WDT	WT	NT	
[9]	Visual-only model	95.09%	91.20%	92.78%	94.05%	87.39%
Visual branch only	—	97.86%	94.00%	95.56%	94.29%	93.52%
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 embedding	CTC	98.13%	94.20%	**97.78%**	94.52%	93.66%
	Attn	**98.37%**	**94.80%**	95.56%	**95.71%**	**94.90%**

Table 6. Recall@1 performance comparison between our model and previous work [9]. Note: WT: word trademarks, WDT: word design trademarks, NT: no-text trademarks, Misc: mixed trademarks.

170. In all of the evaluation scenarios, our model outperforms [9] by a large margin. In hard thresholding, 82.52% of [9]'s predictions are out of distribution with the nearest distance larger than 170, which indicates their learned metric space is more scattered than ours.

4.4. Embedder Level Evaluation

Now we dive deep into the embedder performance. We use the ground truth bounding box locations to create the logo crops. We examined the retrieval performance when using the learnt textual embedding only, visual embedding only and the concatenated visual-textual embeddings, as shown in table 6. Since we deliberately train the text branch to learn to recognize texts with text recognition loss, it's not trained to differentiate logo ROIs based on visual appearances. As a result, 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. For most datasets, attention based text prediction head outperforms CTC head. The same trend is observed by [1] as well.

For PL8K and LogoDet-3K datasets, we finetune on the training splits first and test on their test splits to report the performance. For all other smaller datasets as mentioned in section 4.2, we directly test our trained model on the whole sets without finetuning, as shown in table 7. The trained models generalize very well to other public datasets, and are consistently better than visual only model. Our model trained on LogoDet-3K generalizes better than model trained on PL8K. This is expected as PL8K is collected from Amazon Product Catalog, having larger domain gap with other public datasets than LogoDet-3K.

Single model vs Multi-task SeeTek As the concatenated embedding from visual + text has size of 1024, we trained the visual only non-text aware DML model[9] with the same 1024 embedding dimensions for fair comparison. SeeTek model outperforms visual-only model 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. This shows that training jointly with text supervision helped the visual branch to attend to logo regions and improved its performance as well.

Recall@1	Text Head	FlickrLogos-47		LitW	OpenLogo	BelgaLogos	
			Text				Text
[9]	—	91.88%	91.33%	81.87%	83.14%	96.09%	100.00%
SeeTek	CTC	90.62%	92.00%	84.47%	86.32%	94.78%	100.00%
	Attn	91.88%	**94.00%**	**87.72%**	**89.64%**	**97.39%**	**100.00%**

Table 7. Model generalization: Recall@1 performance on public datasets from SeeTek model trained on LogoDet-3K's training split.

PL8K In particular, we see large performance jumps on text-heavy data splits like on word design trademarks (WDT) from 91.20% to 94.80%, on word trademarks (WT) from 92.78% to 97.78% (95.56% for Attention head). These results indicate that the text branch extracted new complementary information the other branch doesn't bring, which helps learning a useful embedding for metric learning even if the labels for that task are noisy.

Out of domain performance Similarly to PL8K, the text-aware models largely outperform the baselines. On some datasets (LitW, OpenLogo) almost ~6% absolute points. Even when the logo dataset does not have that many text-heavy logos, the performance does not drop. This shows that this form of training does not harm the no-text-logo cases and the models are able to generalize to out of domain datasets.

5. Conclusion

In this paper we have proposed to leverage decoupled visual and textual information to improve open-set logo recognition performance. Our approach is highly scalable and doesn't require retraining for unseen test classes. It successfully combines image and text signals into a single loss function that can be optimized end-to-end, leading to considerable performance improvements on Top-1 retrieval tasks. We also introduced a very large-scale dataset we built containing 3M product logo images on 8K brands (PL8K). Moreover, we have observed the performance improvements not only on large in-domain test data but also on challenging out-of-domain imagery of previously unseen brands, both symbol and text-heavy logos, which demonstrates our model's generalization ability.

6. Acknowledgments

The authors want to thank Jason Sun, Shaonan Zhang and Wahaj Chaudhry for their insightful discussions. Thanks to Arun Reddy, Sergio Fernandez, Borja Lafuente and Mario Sotil for their contribution in building the data collection system.

References

[1] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *International Conference on Computer Vision (ICCV)*, 2019. to appear. 3, 4, 5, 8

[2] Muhammet Bastan, Hao-Yu Wu, Tian Yu Cao, Bhargava Urala Kota, and Mehmet Tek. Large scale open-set deep logo detection. *ArXiv*, abs/1911.07440, 2019. 1, 2, 3, 4, 5

[3] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. Logo recognition using CNN features. In *International Conference on Image Analysis and Processing*, pages 438–448. Springer, 2015. 2

[4] Simone Bianco, Marco Buzzelli, Davide Mazzini, and Raimondo Schettini. Deep learning for logo recognition. *Neurocomputing*, 245:23–30, 2017. 2

[5] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv e-prints*, page arXiv:2004.10934, Apr. 2020. 7

[6] Zhanzhan Cheng, Fan Bai, Yunlu Xu, Gang Zheng, Shiliang Pu, and Shuigeng Zhou. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pages 5076–5084, 2017. 3, 4

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 6

[8] István Fehérvári, Avinash Ravichandran, and Srikar Appalaraju. Unbiased evaluation of deep metric learning algorithms. *ArXiv*, abs/1911.12528, 2019. 4

[9] 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, 4, 6, 7, 8

[10] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable detection of offensive and non-compliant content / logo in product images. In *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2020. 1

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA, 2006. Association for Computing Machinery. 4

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 5

[13] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In Aasa Feragen, Marcello Pelillo, and Marco Loog, editors, *SIMBAD*, volume 9370 of *Lecture Notes in Computer Science*, pages 84–92. Springer, 2015. 3

[14] Steven CH Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *arXiv preprint arXiv:1511.02462*, 2015. 2

[15] Steven C. H. Hoi, Xiongwei Wu, Hantang Liu, Yue Wu, Huiqiong Wang, Hui Xue, and Qiang Wu. Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks. *CoRR*, abs/1511.02462, 2015. 5

[16] Forrest N Iandola, Anting Shen, Peter Gao, and Kurt Keutzer. Deeplogo: Hitting logo recognition with the deep neural network hammer. *arXiv preprint arXiv:1510.02131*, 2015. 2

[17] 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. 3, 4, 6

[18] S. Jiang, S. Liang, C. Chen, Y. Zhu, and X. Li. Class agnostic image common object detection. *IEEE Transactions on Image Processing*, 28(6):2836–2846, 2019. 2

[19] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 581–584, New York, NY, USA, 2009. ACM. 5, 6

[20] Alexis Joly and Olivier Buisson. Logo retrieval with a contrario visual query expansion. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 581–584. ACM, 2009. 7

[21] Chen-Yu Lee and Simon Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2231–2239, 2016. 3

[22] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8714–8721, 2019. 3

[23] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. SphereFace: Deep hypersphere embedding for face recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6738–6746, 2017. 4

[24] Gonçalo Oliveira, Xavier Frazão, André Pimentel, and Bernardete Ribeiro. Automatic graphic logo detection via fast region-based convolutional networks. In *Neural Networks (IJCNN), 2016 International Joint Conference on*, pages 985–991. IEEE, 2016. 2

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7

[26] Rose M Rider. Color psychology and graphic design applications. 2010. 4

[27] Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof van Zwol. Scalable logo recognition in real-world

images. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 25:1–25:8, New York, NY, USA, 2011. ACM. 5, 6

[28] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11):2298–2304, 2016. 3

[29] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4168–4176, 2016. 3, 4

[30] Hang Su, Shaogang Gong, and Xiatian Zhu. Scalable logo detection by self co-learning. *Pattern Recognition*, 97:107003, 2020. 1, 2

[31] Hang Su, Shaogang Gong, Xiatian Zhu, et al. Weblogo-2m: Scalable logo detection by deep learning from the web. 2018. 2, 5, 6

[32] Hang Su, Xiatian Zhu, and Shaogang Gong. Deep learning logo detection with data expansion by synthesising context. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 530–539, 2017. 5

[33] Hang Su, Xiatian Zhu, and Shaogang Gong. Open logo detection challenge. *arXiv preprint arXiv:1807.01964*, 2018. 2, 5, 6, 7

[34] Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. Open Set Logo Detection and Retrieval. In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications: VISAPP*, 2018. 2, 5, 6

[35] Hai Jun Wang, Yitong Wang, Zuo-Feng Zhou, Xing Ji, Zhifeng Li, Dihong Gong, Jingchao Zhou, and Wenyu Liu. CosFace: Large margin cosine loss for deep face recognition. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 4

[36] Jing Wang, Weiqing Min, Sujuan Hou, Shengnan Ma, Yuanjie Zheng, and Shuqiang Jiang. Logodet-3k: A large-scale image dataset for logo detection. *arXiv preprint arXiv:2008.05359*, 2020. 1, 2, 4, 5, 6, 7

[37] Xiao Yang, Dafang He, Zihan Zhou, Daniel Kifer, and C Lee Giles. Learning to read irregular text with attention mechanisms. In *IJCAI*, volume 1, page 3, 2017. 3

[38] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. In *Proceedings of the British Machine Vision Conference (BMVC'19)*, 2019. 3, 4, 6, 7

[39] Fangneng Zhan and Shijian Lu. Esir: End-to-end scene text recognition via iterative image rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2059–2068, 2019. 3

[40] Xiaonan Zhao, Huan Qi, Rui Luo, and Larry Davis. A weakly supervised adaptive triplet loss for deep metric learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 4

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.