Image-Text Pre-Training for Logo Recognition Supplementary Material

Mark Hubenthal
mhubenth@amazon.com
Amazon Inc.

Suren Kumar
ssurkum@amazon.com
Amazon Inc.

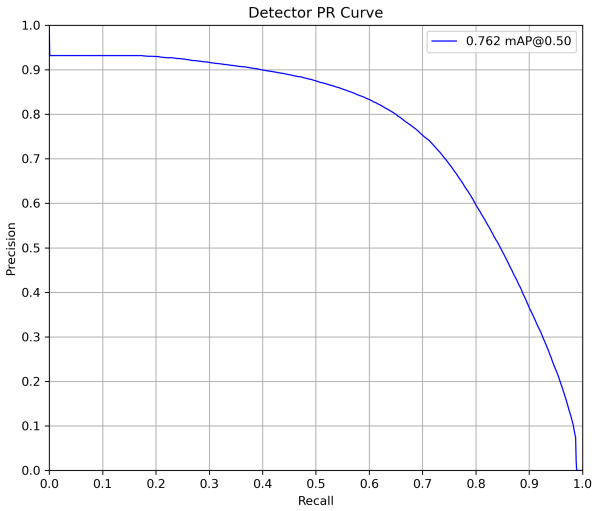

Figure 1. YoloV4 Logo Detector PR Curve

A. Class Agnostic Logo Detector Model

In conjunction with the logo embedding model presented in this work, a class agnostic logo detection model allows one to make end-to-end predictions on an input image. We train a YoloV4 model on the PL2K dataset, consisting of 185247, 46312, and 57970 Amazon product images in the train, validation, and test splits, respectively. Each image is annotated with axis-aligned bounding boxes to localize a logo region. We used an input size of 512 pixels and the medium depth version of the architecture, containing 23.4 million parameters. Average precision at 0.5 IoU is 0.762 and its corresponding precision-recall curve on the test split is shown in Figure 1.

In Figure 2 we also show several predictions on the same image when varying the objectness score threshold. When using a logo embedding model together with the detector in order to identify logos in an input image, we observe that operating the detector at a lower threshold can improve overall recall of the system. However, more extraneous regions are surfaced at such a lower threshold, which places greater importance on the ability of the embedding model to distinguish hard-negative background regions from actual logos.

B. Cleaning of OpenLogoDet3K47 Dataset

After merging the various public logo datasets of which OpenLogoDet3K47 consists, we cleaned the resulting union via the following steps:

1. Force all class names to lower case, replace "-" or " " with "_" and "'" with an empty string.

2. Merge some obvious classes such as lv $\mapsto$ louisvuitton, cocacola $\mapsto$ coca_cola, northface $\mapsto$ the_north_face

3. For certain classes where there are text/symbol child classes from one source and not in another, we carefully reassign labels or merge them to minimize overall label noise.

4. Remove duplicate images within the same class.

After this process, we have 3276 classes with 188244 images and 235738 objects. For our experiments, we filter out any image regions with a minimum side length less than 10 pixels and we remove classes with fewer than 20 instances. This yields 2714 classes with 181552 images and 227176 objects.

C. Fixed Hyperparameters

Table 1 shows hyperparameters that were fixed depending on model architecture for the experiments performed in the main paper.

D. Additional ViT Comparison On OpenLogoDet3K47

The test split of OpenLogoDet3K47 contains a total of 48098 images. Table 2 shows a summary of mistakes by the best ViT embedder pre-trained on ImageNet and the best ViT embedder pre-trained on image-text data. In particular, we see that the image-text pre-trained ViT model has over 3 times as many correct predictions when the ImageNet pre-trained ViT was incorrect as vice versa.

Figure 2. Logo detector predicted bounding boxes on an input image with different applied objectness score thresholds as specified below each image

Table 1. Fixed hyperparameters

	ViT	ResNet50	CLIP RN50
Temperature Scaling σ	0.06	0.06	0.06
Trunk Weight Decay	0.2	0.2	0.2
Last FC Weight Decay	0.001	0.001	0.001
Proxy Weight Decay	0	0	0
Adam $\beta_{1,\{\text{trunk,fc,proxy}\}}$	0.9	0.9	0.9
Adam $\beta_{2,\{\text{trunk,fc}\}}$	0.98	0.999	0.999
Adam $\beta_{2,\text{proxy}}$	0.999	0.999	0.999
Adam $\epsilon_{\{\text{trunk,fc}\}}$	10^{-6}	10^{-8}	10^{-8}
Adam ϵ_{proxy}	1	1	1

In Table 3 we compare performance of the image-text and ImageNet pre-trained ViT embedding models on several public logo datasets. Each model was trained on the train and validation splits of the LogoDet3K datasets and evaluated on the test split. This is the open-set regime where test classes in LogoDet3K are unseen. On the most sizeable of these public datasets, OpenLogo and LogosInTheWild, we see 1% and 3% increases in recall@1 performance when restricted to classes containing "text" in the name. We further note that recall@1 performance on LogoDet3K is over 2% better in the query versus gallery setting. In the few cases where the ImageNet pre-trained model performs better in Table 3, the difference is relatively small.

We noticed among the entire set of 285 query images from the test split where the image-text pre-trained model made an incorrect prediction and the ImageNet1K pre-trained model was correct, many such images were very small in size or blurry to the point of ambiguity. We also noticed some query images that seem to be assigned to the wrong class, such as the top left image in Figure 3. Moreover, there is some label redundancy that lead to mistakes, such as the three armani-related classes: armani, armani_junior, and armani_exchange. Finally, in many cases when the image-text pre-trained model was incorrect, it was still able to match to a logo with similar letters and/or text style. Several of these comparison images are shown in Figure 3.

Table 2. Counts of correct and incorrect predictions (using the closest neighbor) by two ViT embedding models. $|TP \cap FP_{\text{other}}|$ indicates the size of the set of true positives from the given model intersected with the set of false positives from the other model. $FP_1 \cap FP_2$ indicates the set of query images for which both models predicted incorrectly

Model	Pre-Training	$\lvert TP \rvert$	$\lvert FP \rvert$	$\lvert TP \cap FP_{\text{other}} \rvert$	$\lvert FP_1 \cap FP_2 \rvert$	Precision
ViT	OpenAI IT	47461	637	**826**	352	**0.9867**
ViT	ImageNet	46920	1178	285	352	0.9755

Table 3. Recall@1 performance for best ViT model pre-trained on image-text data vs best ViT model pre-trained on ImageNet data. Both models trained on LogoDet3K train and val splits with test set held out.

Model	Pre-Training	LogoDet3K Test		OpenLogo			BelgaLogo		FlickrLogos-47		LiTW	
		QvG	All	QvG	All	Text	All	Text	All	Text	All	Text
ViT	OpenAI IT	**0.9836**	**0.9886**	**0.9371**	0.9629	**0.9568**	0.9797	**0.9784**	0.9834	**0.9778**	0.9391	**0.9456**
ViT	ImageNet	0.9622	0.9740	0.9305	**0.9675**	0.9463	**0.9809**	0.9753	**0.9879**	0.9759	**0.9394**	0.9169

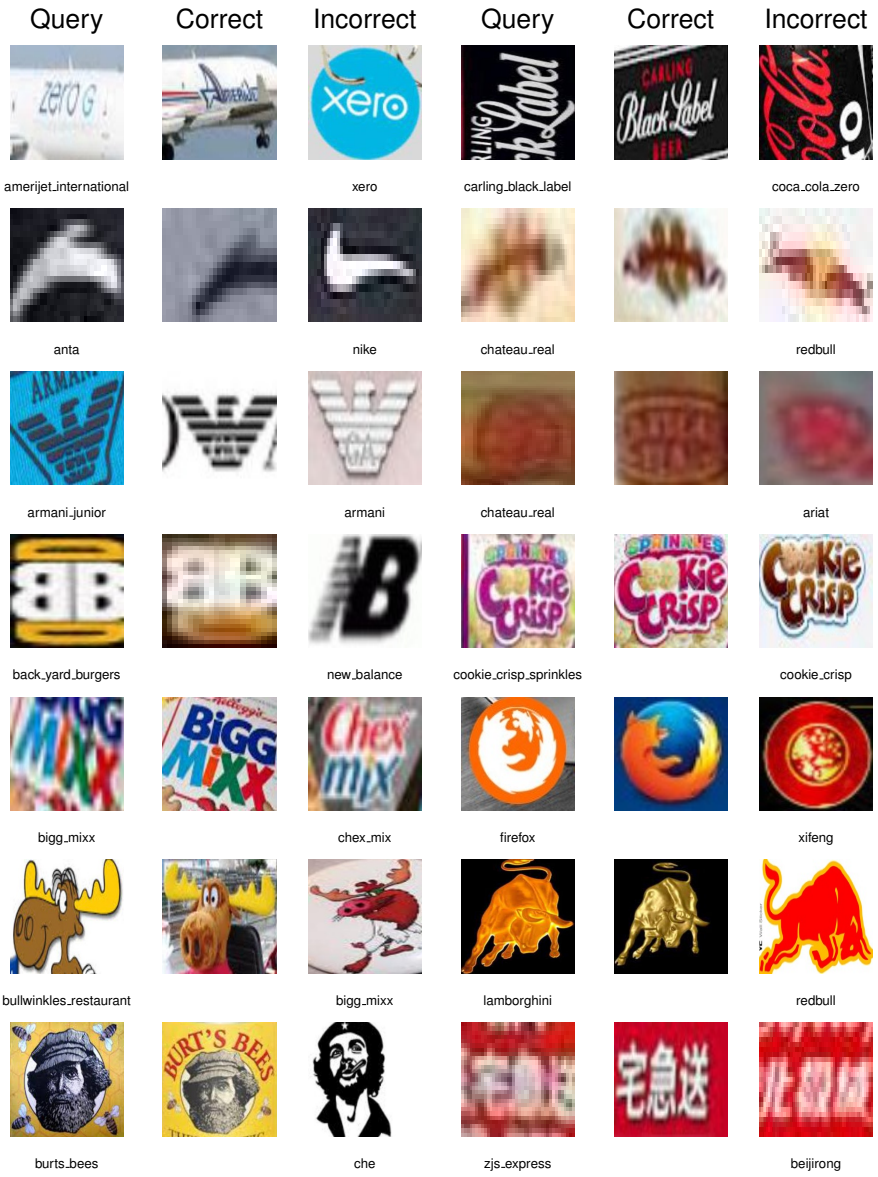

Figure 3. Sample images from the OpenLogoDet3K47 dataset test split where the image-text pre-trained ViT logo embedder predicts incorrectly while the ImageNet pre-trained ViT predicts correctly. Columns 1+4: query image, columns 2+5: ImageNet-pre-trained ViT model's correct logo retrieval, columns 3+6: Image-text pre-trained ViT model's incorrect logo retrieval