Web Scale Image Annotation: Learning to Rank with Joint Word-Image Embeddings

Jason Weston* Samy Bengio* Nicolas Usunier†

*Google, USA
{jweston,bengio}@google.com

†Université Paris 6, LIP6, France
nicolas.usunier@lip6.fr

Web scale image annotation datasets have tens of millions of images with tens of thousands of possible annotations. We propose a strongly performing method that scales to such datasets by simultaneously learning to optimize precision at k of the ranked list of annotations for a given image *and* learning a low-dimensional joint embedding space for both images and annotations. Our method both outperforms several baseline methods and, in comparison to them, provides a highly scalable architecture in terms of memory consumption and prediction time. We also demonstrate how our method learns an interpretable model, where annotations with alternate spellings or even languages are close in the embedding space. Hence, even when our model does not predict the exact annotation given by a human labeler, it often predicts similar annotations, a fact that we try to quantify by measuring the so-called "sibling" precision metric, where our method also obtains excellent results.

Joint Word-Image Model We jointly learn two mappings, one for images and one for annotations, into the same joint feature space finally learning a ranking function of the form:

$$f_{\hat{y}}(x) = s(\Phi_W(\hat{y}), \Phi_I(x)) \tag{1}$$

where the possible annotations $\hat{y}$ are ranked according to $f_{\hat{y}}(x)$, largest first, $s(\cdot,\cdot)$ is the negative Euclidean distance, and $\Phi_I(x) = Vx$ and $\Phi_W(\hat{y}) = W_{\hat{y}}$ are linear maps from image x and annotation $\hat{y}$. Our goal is, for a given image, to rank the possible annotations such that the highest ranked ones best describe the content of the image. We learn this mapping using a ranking function that optimizes the top of the ranked list.

Weighted Approximate-Rank Pairwise (WARP) Loss A classical approach to learning to rank is to maximize AUC by minimizing the pairwise loss $\sum_x \sum_y \sum_{\bar{y} \neq y} \max(0, 1 + f_{\bar{y}}(x) - f_y(x))$ where y is the true label for x. A scalable version of this cost is based on sampling triplets $(x, y, \bar{y})$ and applying stochastic gradient descent (SGD) on the resulting hinge loss. However this cost considers all pairwise errors similarly while for many applications including image annotation, one is often more interested in optimizing the top of the ranked list. A class of ranking error functions recently defined in (Usunier et al., 2009) defines the loss:

$$err(f(x), y) = L(rank_y(f(x))), \qquad rank_y(f(x)) = \sum_{\bar{y} \neq y} I(f_{\bar{y}}(x) \geq f_y(x)) \tag{2}$$

where $rank_y(f(x))$ is the rank of the true label y given by $f(x)$ and $L(\cdot)$ transforms this rank into a loss: $L(k) = \sum_{j=1}^{k} \alpha_j$, with $\alpha_1 \geq \alpha_2 \geq \cdots \geq 0$. This class allows one to define different choices of $L(\cdot)$ with different minimizers. Results on (small) text retrieval datasets in (Usunier et al., 2009) showed that a choice of $\alpha_j = 1/j$ yields state-of-the-art results in terms of precision@k or mean average precision (MAP).

While it is well known how to optimize AUC online by SGD for large scale tasks, optimizing MAP or precision@k is so far less scalable. In this work we show how to efficiently optimize (2) by SGD for arbitrary differentiable models. The main idea is on each update, instead of measuring all $\bar{y} \neq y$ in (2) which is too expensive, we sample $\bar{y}$ with replacement until $f_{\bar{y}}(x) \geq f_y(x)$. We can then approximate the rank with:

$$rank_y(f(x)) \approx \left\lfloor \frac{Y-1}{N} \right\rfloor$$

where Y is the number of labels, $\lfloor . \rfloor$ is the floor function and N the number of trials in the sampling step. Overall, our method which we call WSABIE (Web Scale Annotation by Image Embedding, pronounced "wasabi") consists of the joint word-image embedding model trained by SGD using the WARP loss.

Experiments We had access to a very large proprietary database of images taken from the web, together with noisy annotation based on anonymized user click information. There was respectively 10M, 3M and 3M training, validation and test images, annotated with 109,444 different labels. Given the size of the label set, many labels can be semantically close to each other. Indeed, two different labels can be synonyms, translations or alternative spellings. Our model tries to capture this structure through the projection in the embedding space. For each label, we had access to a set of so-called "sibling labels", considered semantically near the considered label, which we use for evaluation. Images were represented by a sparse vector of texture and color features, following (Grangier & Bengio, 2008). Results are given in Table 1 compared to k-NN, One-Vs-Rest using (Crammer et al., 2006) and Pamir (Grangier & Bengio, 2008) which optimizes AUC. Example word embeddings learnt by Wsabie are given in Table 2 and some example image annotations are given in Table 3. One can see that the embeddings seem to learn the semantic structure of the annotation space (and images are also embedded in this space) and sibling annotations are close to each other. This explains why the sibling precision of Wsabie is far superior to competing methods, which do not attempt to learn the structure between annotations.

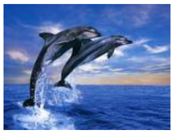
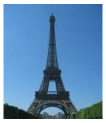
Table 1: **Test Set Results.** Precision at 1 and 10, Sibling Precision at 10, Mean Average Precision (MAP), and time and space complexity to return the top annotation on a test image, not including feature generation. [In brackets, concrete time/memory for a single CPU machine]. Y is the number of classes, n the number of train examples, d the image input dimension, $d_{\bar{\emptyset}}$ the average number of non-zero values per image, and D the size of the embedding space.

Algorithm	p@1	p@10	p$_{sib}$@10	MAP	Time	Space
k-NN	0.30%	0.34%	5.97%	1.52%	$\mathcal{O}(n \cdot d_{\bar{\emptyset}})$ [113s]	$\mathcal{O}(n \cdot d_{\bar{\emptyset}})$ [27GB]
One-vs-Rest	0.52%	0.29%	4.61%	1.45%	$\mathcal{O}(Y \cdot d_{\bar{\emptyset}})$ [0.5s]	$\mathcal{O}(Y \cdot d)$ [8.2GB]
Pamir^{IA}	0.32%	0.16%	2.94%	0.83%	$\mathcal{O}(Y \cdot d_{\bar{\emptyset}})$ [0.5s]	$\mathcal{O}(Y \cdot d)$ [8.2GB]
Wsabie	1.03%	0.44%	9.84%	2.27%	$\mathcal{O}((Y + d_{\bar{\emptyset}}) \cdot D)$ [0.17s]	$\mathcal{O}((Y + d) \cdot D)$ [82MB]

Table 2: **Nearest annotations in the embedding space learnt by** Wsabie**.** Translations and alternative/misspellings and synonyms have close embeddings. Other annotations are from similar visual images.

Annotation	**Neighboring Annotations**
barack obama	*barak obama*, *obama*, barack, barrack obama, bow wow, george w bush, *berlusconi*
david beckham	*beckham*, david beckam, *alessandro del piero*, *del piero*, david becham, *fabio cannavaro*
dolphin	delphin, dauphin, *whale*, delfin, delfini, baleine, *blue whale*, walvis, *bottlenose dolphin*, delphine
cows	*cattle*, shire, *dairy cows*, kuh, *horse*, *cow*, *shire horse*, kone, *holstein*, appaloosa, caballo, vache
mount fuji	mt fuji, fuji, fujisan, fujiyama, *mountain*, zugspitze, fuji mountain, paysage, mount kinabalu
eiffel tower	*eiffel*, *tour eiffel*, la tour eiffel, *big ben*, paris, *blue mosque*, eifel tower, eiffel tour, paris france

Table 3: **Examples of the top 10 annotations of three approaches:** Pamir^{IA} , One-vs-Rest and Wsabie, on the Web dataset. Annotations in **red+bold** are the true labels, and those in *blue+italics* are so-called siblings.

Image	Pamir^{IA}	One-vs-Rest	Wsabie
	bora, free willy, su, orka, worldwide, sunshine coast, bequia, tioman island, universal remote montagna, esperar, *bottlenose dolphin*	surf, bora, belize, sea world, balena, wale, tahiti, delfini, surfing, *mahi mahi*	delfini, *orca*, **dolphin**, mar, delfin, dauphin, *whale*, cancun, *killer whale*, sea world
	air show, st augustine, stade, concrete architecture, streetlight, doha qatar, skydiver, *tokyo tower*, sierra sinn, lazaro cardenas	**eiffel tower**, *tour eiffel*, snowboard, blue sky, *empire state building*, luxor, *eiffel*, *lighthouse*, jump, adventure	**eiffel tower**, *statue*, *eiffel*, mole antoneliana, la tour eiffel, londra, cctv tower, *big ben*, calatrava, *tokyo tower*

References

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., & Singer, Y. (2006). Online passive-aggressive algorithms. *Journal of Machine Learning Research*, *7*, 551–585.

Grangier, D., & Bengio, S. (2008). A discriminative kernel-based model to rank images from text queries. *Transactions on Pattern Analysis and Machine Intelligence*, *30*, 1371–1384.

Usunier, N., Buffoni, D., & Gallinari, P. (2009). Ranking with ordered weighted pairwise classification. *Proceedings of the 26th International Conference on Machine Learning* (pp. 1057–1064). Montreal: Omnipress.