Associating Neural Word Embeddings with Deep Image Representations using Fisher Vectors

Benjamin Klein, Guy Lev, Gil Sadeh, Lior Wolf
The Blavatnik School of Computer Science, Tel Aviv University.

In recent years, the problem of associating a sentence with an image has gained a lot of attention. This work continues to push the envelope and makes further progress in the performance of image annotation and image search by a sentence tasks. In this work, we are using the Fisher Vector as a sentence representation by pooling the word2vec embedding of each word in the sentence.

The Fisher Vector [7] is an advanced pooling technique, which provided state-of-the-art results on many different applications in computer vision. The Fisher Vector of a set of local descriptors is obtained as a concatenation of gradients of the log-likelihood of the descriptors in the set with respect to the parameters of a Gaussian Mixture Model that was fitted on a training set in an unsupervised manner. Many different improvements were suggested for the Fisher Vector but all of them are in the context of the Gaussian Mixture Model. In [2], Jia et al. showed empirically that the statistics of gradient based image descriptors, such as SIFT, often follow a heavy-tailed distribution which suggests that a Gaussian distribution does not capture well the descriptors' distribution, and that the Euclidean distance is not a suitable distance. Motivated by their findings, this paper presents and evaluates new variants of Fisher Vectors that are based on the Laplacian distribution.

By using the common assumption in the Fisher Vector that the covariance matrix is a diagonal one, we define the multivariate Laplacian distribution and the Laplacian Mixture Model (LMM). We explain how to fit a LMM by deriving the Expectation-Maximization (EM) equations and supply the Fisher Vector definition for this model. Similar to [7], we approximate the diagonal of the Fisher Information Matrix in order to normalize the dynamic range of the different dimensions in the Fisher Vector variant presented.

In order to gain the benefits of the two distributions in a single model, we define a new distribution, the Hybrid Gaussian-Laplacian distribution, which can be seen as a weighted geometric mean of the Gaussian and Laplacian distributions. As before, we define the Hybrid Gaussian-Laplacian Mixture Model (HGLMM), derive the EM equations for fitting a HGLMM model, derive the Fisher Vector definition and approximate the diagonal of the Fisher Information Matrix.

We employ the new variants of the Fisher Vectors for tasks that match texts with images. In our experiments, the images are represented by the VGG [8] Convolutional Neural Network as a single vector. The text is represented as a set of vectors obtained by the word2vec method. This set is converted to a Fisher Vector based on one of the distributions: GMM, LMM, or HGLMM. Text to image matching is done using the Canonical Correlations Analysis algorithm. This combination of methods proves to be extremely potent as we achieve state-of-the-art results for both the image annotation and the image search by a sentence tasks on four benchmarks: Pascal1K, COCO, Flickr8K, and Flickr30K (Table 1).

[1] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389v2*, 2014.

[2] Yangqing Jia and Trevor Darrell. Heavy-tailed distances for gradient based image descriptors. In *Advances in Neural Information Processing Systems*, pages 397–405, 2011.

[3] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Technical report, Computer Science Department, Stanford University, 2014. URL http://cs.stanford.edu/people/karpathy/deepimagesent/.

[4] Andrej Karpathy, Armand Joulin, and Li Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[6] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.

[7] Florent Perronnin and Christopher Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

[9] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

	Image search					Image annotation					Sentence
	r@1	r@5	r@10	median rank	mean rank	r@1	r@5	r@10	median rank	mean rank	mean rank
DFE [4]	10.3	31.4	44.5	13.0	NA	16.4	40.2	54.7	8.0	NA	NA
BRNN [3]	15.2	37.7	50.5	9.2	NA	22.2	48.2	61.4	4.8	NA	NA
SC-NLM [5]	16.8	42.0	56.5	8.0	NA	23.0	50.7	62.9	5.0	NA	NA
NIC [9]	17.0	NA	57.0	7 .0	NA	17.0	NA	56.0	7.0	NA	NA
m-RNN [6]	12.6	31.2	41.5	16.0	NA	18.4	40.2	50.9	10.0	NA	NA
LRCN [1]	14.0	34.9	47.0	11.0	NA	NA	NA	NA	NA	NA	NA
Mean Vec	20.5	46.3	59.3	6.8	32.4	24.8	52.5	64.3	5.0	27.3	16.3
GMM	23.9	51.6	64.9	5.0	24.8	33.0	60.7	71.9	3.0	19.0	15.1
LMM	23.6	51.2	64.4	5.0	25.2	32.5	59.9	71.5	3.2	19.2	15.6
HGLMM	24.4	52.1	65.6	5.0	24.5	34.4	61.0	72.3	3.0	18.1	14.7
GMM+HGLMM	25.0	52.7	66.0	5.0	23.7	35.0	62.0	73.8	3.0	17.4	14.2

Table 1: Results on the Flickr30K benchmark. Shown are the recall rates at 1,5 and 10 retrieval results (higher is better). Also shown, the mean and median rank of the first ground truth (lower is better). There are three tasks: image annotation, image search, and sentence similarity. We compare the results of [1, 3, 4, 5, 6, 9] to the mean vector baseline and to Fisher Vectors based on GMM, LMM and HGLMM. In addition we report results for the combination of the GMM and HGLMM Fisher Vectors.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.