ASSESSING SEMANTIC INFORMATION IN CONVOLUTIONAL NEURAL NETWORK REPRESENTATIONS OF IMAGES VIA IMAGE ANNOTATION

Michael B. Mayhew, Barry Chen

Lawrence Livermore National Laboratory
Computational Engineering Division
Livermore, CA USA
mayhew5@llnl.gov, chen52@llnl.gov

Karl S. Ni[*]

In-Q-Tel

Menlo Park, CA USA
kni@iqt.org

ABSTRACT

Image annotation, or prediction of multiple tags for an image, is a challenging task. Most current algorithms are based on large sets of handcrafted features. Deep convolutional neural networks have recently outperformed humans in image classification, and these networks can be used to extract features highly predictive of an image's tags. In this study, we analyze semantic information in features derived from two pre-trained deep network classifiers by evaluating their performance in nearest neighbor-based approaches to tag prediction. We generally exceed performance of the manual features when using the deep features. We also find complementary information in the manual and deep features when used in combination for image annotation.

***Index Terms*—** deep learning, image annotation, feature representations

1. INTRODUCTION

Recent years have witnessed at least two major computational phenomena: a deluge of multimodal data (e.g. images with text tags) [19, 16] and the proliferation of deep neural network models to analyze and annotate these data. As the volume of multimedia data rapidly increases, there is a clear and present need for reliable annotation methods to assign tags to unannotated image collections, to correct or enhance existing annotation, and to enable cross-modal search and retrieval. Deep convolutional neural networks (CNNs), classifiers comprised of multiple hidden layers (including convolutional filter layers to handle spatial structure), have set the standard for supervised image labeling tasks [10, 18, 17]. Systems exploiting deep neural networks to assign multiple text tags to an image are still coming on line [9, 8].

A variety of approaches, particularly nearest neighbor search-based methods, have been successful using manual image features (e.g. SIFT [12]) to automatically assign multiple tags to images. These approaches are based on the assumption that similar images in feature space share contextually similar tags that can be ranked and transferred to the query image [13, 6, 20, 2, 7]. Currently, a set of 15 manual features remains at the heart of high-performance tag prediction [6, 20].

Deep neural networks have been shown to learn feature representations of images that are highly predictive of single image labels [10, 11, 17]. These observations indicate the possibility that the deep features could be used to improve automatic image annotation [14]. In building systems for automatic image annotation, two central questions must be addressed: 1) can deep features replace traditionally used manual features and 2) do the two types of features contain complementary information for image annotation?

In this work, we use nearest neighbor-based image annotation as a platform for assessing the semantic information in image features derived from two deep CNN classifiers: AlexNet [10] and VGG16 [17]. In this way, we draw a direct correspondence between the tag prediction performance of image features and the semantic expressiveness of those features. We demonstrate substantial gains in predictive performance when using deep rather than manual features to train current tag prediction methods. We also assess the performance and relative contributions of both the manual and deep features when used in conjunction for image annotation.

2. RELATED WORK

Methods for automatic image annotation have ranged from generative and discriminative models for image tags [1, 4, 22] to nearest neighbor search-based approaches [13, 6, 20, 7]. The present work follows more closely this second lineage of methods for tag prediction. An important study by Makadia et al. [13] helped establish this line of research, demonstrating state-of-the-art performance at the time with a simple k-nearest neighbor algorithm combined with a greedy approach to prioritization of tags from neighboring images. Subsequent work set the current standard for tag prediction performance using probabilistic models for tag gener-

[*] Author performed work while at LLNL.

	IAPR-TC12				ESP Game			
	P	**R**	**F1**	**N+**	**P**	**R**	**F1**	**N+**
SD + LEAR15	49	21	30	220	48	17	25	203
σSD + LEAR15	43	30	36	258	39	23	29	230
ML + LEAR15	49	24	32	221	49	17	25	203
σML + LEAR15	45	34	39	263	39	25	30	234
SD + AlexNet-fc7	49	23	31	216	51	20	29	206
σSD + AlexNet-fc7	45	33	38	260	40	**32**	**36**	**244**
SD + VGG-16	50	26	34	226	**52**	20	29	211
σSD + VGG-16	45	36	40	263	40	**32**	**36**	242
SD + LEAR15 + VGG-16	50	22	31	224	48	17	26	203
σSD + LEAR15 + VGG-16	44	32	37	259	39	24	30	231
ML + LEAR15 + VGG-16	**54**	30	38	243	51	21	29	216
σML + LEAR15 + VGG-16	49	**39**	**44**	**271**	40	31	35	243

Table 1. Results computed using TagProp are shown with the indicated variant (SD - single distance-based; ML - metric learning) and feature combination. For all runs, $K = 200$.

ation that accounted for the nearness of an image's neighbors and for the occurrence of rare tags. These studies relied on a set of 15 global and local image features ([6, 20]; http://lear.inrialpes.fr/people/guillaumin/data.php) that have formed the basis of subsequent tag prediction research.

Deep features have found more use in image annotation as alternatives to manual features. Kiros and Szepesvari [9] learned binary codes with real-valued image features derived from an unsupervised, autoencoder-based approach. Here, we opt for features extracted from a pre-trained deep network and specifically targeted for accurate prediction of single image labels. The work most closely related to this study is by Murthy et al. who tested the predictive performance of deep CNN features in multiple image annotation algorithms [14]. The authors' results were inconclusive in using deep features in place of manual features for three different nearest-neighbor algorithms: some algorithms performed worse with the deep features than with the manual features. Similarly to the authors of that study, we investigate predictive performance of deep features in high-performing nearest-neighbor-based image annotation approaches. In addition, we regenerate predictions with the standard manual features to ensure that we have a baseline for comparison with the deep features. Moreover, we extend our analysis to include investigation of relative contributions of the two different classes of features (deep vs. manual) to prediction.

3. DATASETS & APPROACH

3.1. Benchmark Image Datasets

The IAPR-TC12 benchmark dataset consists of natural scene imagery and a vocabulary of 291 words ([5]; http://imageclef.org/photodata). The same training and test split was used as in the preceding literature [6], with the training and test sets consisting of 17665 and 1962 images, respectively. The ESP-Game dataset is labeled from a vocabulary of 269 words by a pair of players who are rewarded for choosing common tags for an image [21]. The training and test sets consist of 18689 and 2081 images, respectively (as in previous work [6]).

3.2. Deep CNN Architecture & Feature Generation

To derive deep features, we use two different neural network classifiers. The first classifier is the CaffeNet implementation of Krizhevsky et al.'s deep CNN (AlexNet; [10]). The second network we consider is the 16-layer VGG network [17], winner of the 2014 ILSVRC competition. Both networks were implemented in Caffe and pre-trained on the $\sim$1.2 million images and 1,000 image labels of the 2012 ImageNET Large-Scale Visual Recognition Challenge [16]. With both networks, we use the 4096-dimensional features from the final fully connected layer (fc7). Particularly for AlexNet, fc7 features performed similarly or better than features derived from other layers. We direct the interested reader to [3] and [17] for more technical details of model training.

3.3. Manual Features, Distance Calculation & Normalization

The 15 manual features (LEAR15) consist of Gist [15] along with 6 color histograms and 8 bag-of-visual-words features. For TagProp, with the exception of Gist and the deep features, all features were L_1-normalized. For distance computation, we use L_2 distance for Gist and deep features, L_1 distance for color histograms, and χ^2 distance for all other features (consistent with previous work [6, 20]). For 2PKNN, preprocessing steps were slightly different to maintain consistency with published results: Gist was L_2-normalized; DenseSiftV3H1, HarrisSiftV3H1, HarrisHueV3H1, Rgb, Hsv, and

	IAPR-TC12				ESP Game			
	P	**R**	**F1**	**N+**	**P**	**R**	**F1**	**N+**
LEAR15	49	30	37	275	42	25	31	249
AlexNet-fc7	53	30	38	269	**57**	23	33	248
VGG-16	**54**	**32**	**40**	**276**	52	**27**	**36**	**250**

Table 2. Results computed using 2PKNN. For both datasets, $K = 2$.

Lab features were square-root-transformed before L_1 normalization [20]. Distances were normalized to fall between 0 and 1 with respect to each training or test image.

3.4. TagProp Algorithm

The TagProp algorithm learns rank- or distance-based weights of nearest neighbor images as part of a joint probability model for the occurrence of tags associated with images. The weight of training image j for query image i is $\pi_{i,j} = \frac{e^{-d_\theta(i,j)}}{\sum_{j'} e^{-d_\theta(i,j')}}$ where, in the multi-distance case, θ is a vector of coefficients, $d_\theta(i,j) = \theta' \mathbf{d}_{i,j}$, and $\mathbf{d}_{i,j}$ is the vector of distances between images i and j. In the σ variant of the algorithm, prediction of rare tags (a known impediment to prediction recall) can be enhanced with a logistic discriminant model.

3.5. 2PKNN Algorithm & Weight Learning

In the 2PKNN algorithm, images are first grouped according to their tags. An image's neighbors consist of the K closest images from each semantic or tag group, accounting for label imbalance. The second pass of the algorithm learns feature dimension (v) and metric (w) weights relevant to tag prediction. The distance (D) between two images a and b is: $D(a,b) = \sum_{i=1}^{N} w_i \sum_{f=1}^{F} v_f^i d^i(t_a(f), t_b(f))$. Here, $d^i(t_a(f), t_b(f))$ is the distance metric for a feature type i computed between the fth dimension of the feature vectors (t) of images a and b. For a single feature, w is a scalar.

To learn w, we randomly sampled 10 training ($N = 3000$) and validation ($N = 400$) datasets from the training portions of IAPR-TC12 and ESP-Game. For each of the 10 training-validation sets, we computed distance matrices and evaluated the performance of 2PKNN on the validation set for w's from 1 to 30. We then averaged the performance across all 10 training-validation sets, choosing the w with the highest average F1. Values were as follows: IAPRTC-12 LEAR15, w=18; AlexNet, w=18; VGG-16, w=13; ESP-Game LEAR15, w=17; AlexNet, w=10; VGG-16, w=12.

3.6. Image Annotation Performance Metrics

To compute precision (P), we find the proportion of images predicted to have tag t that truly have tag t in the test set. We then average these keyword-specific precisions over all keywords. For recall (R), we perform a similar computation, averaging over all keywords the proportion of images with each tag t that are predicted to have tag t. We compute the number of recalled words (N+) by tallying all keywords with non-zero recall. Finally, to facilitate the comparison of different feature sets, we compute $F1 = \frac{2PR}{P+R}$. These metrics are computed as in previous work [6, 20]

4. RESULTS

4.1. Deep Features Match or Exceed Manual Feature Performance

To establish a baseline for comparison of the performance of deep and manual features, we first regenerated predictions using the 15 manual features in either TagProp or 2PKNN (shown in the topmost rows of Tables 1 and 2). We find that our results are highly consistent with those previously published, with the exception of those for 2PKNN on ESP-Game. We attribute this result to our 10-fold cross-validation scheme which learned a w with slightly better average validation $F1$ than the w that would've produced the published results. Nevertheless, we note that we attain nearly the same $F1$ as previously published results (31.7) and, as the cross-validation scheme is applied consistently across the deep and manual feature sets, the comparison remains valid.

We then trained the two different image annotation algorithms (TagProp and 2PKNN) with features derived from the two CNN architectures (AlexNet and VGG-16). Our results shown in Tables 1 and 2 demonstrate that both deep features not only match but generally exceed performance of the 15 manual features. We observe the sharpest gains in performance (16%-2PKNN, 20%-TagProp) for the ESP-Game dataset when training with the deep versus the manual features. We also note that our 2PKNN results with VGG-16 are considerably different from those reported in [14]. Upon investigation, we attribute this difference to our learning of 2PKNN's weight parameter w by cross-validation (at 2PKNN's default $w = 1$; IAPRTC-12 VGG-16 $P = 38, R = 23, F1 = 29, N+ = 267$; ESP-Game VGG-16 $P = 38, R = 23, F1 = 29, N+ = 252$). In addition, it is noteworthy that the VGG-16 features perform slightly better than the AlexNet features as the VGG-16 network classifier has demonstrated better single tag prediction performance than AlexNet on the ILSVRC-2012 dataset. This result motivates further investigation of other CNN architectures targeted to single labels for application to multi-tag image annotation. Overall, these

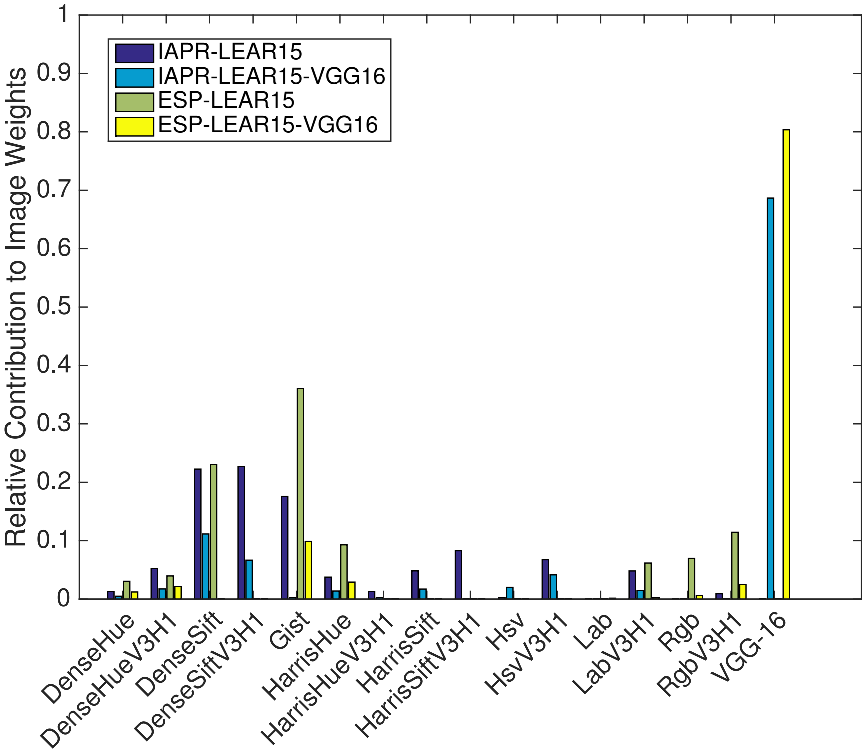

Fig. 1. Coefficients (θ) of feature distance types learned by TagProp (ML), normalized to sum to 1. Similar patterns observed for AlexNet-fc7.

results support the use of deep features as alternatives to the manual features in image annotation systems.

4.2. Complementary Information in Deep and Manual Features

To assess whether the manual and deep features both contribute to image annotation performance, we learn coefficients (θ) for the different feature distance types with the metric learning (ML) variant of TagProp. While the deep features are weighted more highly than the manual features, some of the manual features are still important for prediction (e.g. DenseSift in IAPR-TC12; Figure 1). Notably, we achieve even better predictive performance for IAPR-TC12 when we use metric learning with the manual and deep features in combination (Table 1). However, ESP-Game performance does not improve, potentially owing to the relative importance of the deep features for that dataset. We also do not observe performance gains when the 16 feature distances are weighted equally (SD variant results; bottom rows of Table 1) These analyses indicate that the manual features still offer important complementary information for prediction, at least in the IAPR-TC12 dataset.

5. DISCUSSION & CONCLUSIONS

We have clearly demonstrated that features derived from a deep convolutional neural network match or exceed image annotation performance using larger manual feature sets. We have also provided evidence of complementary information in both the deep and manual features, suggesting they could be used in conjunction to enhance predictive performance, depending on the dataset under study. We note that we used the pre-trained networks as feature transforms without backpropagating tag prediction errors through the network. As part of current and future work, we are developing deep learning frameworks that fully integrate multimedia feature extraction with annotation. Taken together, this analysis supports more widespread adoption and further investigation of deeply learned feature representations in multimedia labeling tasks.

Acknowledgments

The authors would like to thank the INRIA-LEAR group for making their image features publicly avaiable. The authors would especially like to thank Matthieu Guillaumin and Yashaswi Verma for their helpful insights.

6. REFERENCES

[1] D. Blei and M. Jordan. Modeling Annotated Data. In *Proceedings of the 26th International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[2] M. Chen, A. Zheng, and K. Weinberger. Fast Image Tagging. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, 2013.

[3] J. Donahue. BLVC CaffeNET Reference Model. https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet, 2014. [Online; accessed 19-February-2015].

[4] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli Relevance Models for Image and Video Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[5] M. Grubinger, P. Clough, H. Muller, and T. Deselaers. The IAPR Benchmark: A New Evaluation Resource for Visual Information Systems. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2006.

[6] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. TagProp: Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation. In *IEEE 12th International Conference on Computer Vision (ICCV)*, 2009.

[7] M. Kalayeh, H. Idrees, and M. Shah. NMF-KNN: Image Annotation using Weighted Multi-View Non-Negative Matrix Factorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[8] A. Karpathy and F. Li. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[9] R. Kiros and C. Szepesvari. Deep Representations and Codes for Image Auto-Annotation. *Neural Information Processing Systems (NIPS)*, 2012.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012.

[11] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, and A. Ng. Building High-level Features Using Large Scale Unsupervised Learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[12] D. Lowe. Object Recognition from Local Scale-Invariant Features. In *IEEE 7th International Conference on Computer Vision (ICCV)*, 1999.

[13] A. Makadia, V. Pavlovic, and S. Kumar. A New Baseline for Image Annotation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

[14] V. N. Murthy, S. Maji, and R. Manmatha. Automatic Image Annotation using Deep Learning Representations. In *Proceedings of the International Conference on Multimedia Retrieval (ICMR)*, 2015.

[15] A. Oliva and A. Torralba. Modeling the Shape of the Scene: a Holistic Representation of the Spatial Envelope. *IJCV*, 2001.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge, 2014.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going Deeper with Convolutions. *CoRR*, abs/1409.4842, 2014.

[19] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The New Data and New Challenges in Multimedia Research. *arXiv preprint arXiv:1503.01817*, 2015.

[20] Y. Verma and C. Jawahar. Image Annotation Using Metric Learning in Semantic Neighborhoods. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[21] L. Von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2004.

[22] H. Xu, J. Wang, X.-S. Hua, and S. Li. Tag Refinement by Regularized LDA. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, pages 573–576, New York, NY, USA, 2009. ACM.