Love Thy Neighbors: Image Annotation by Exploiting Image Metadata

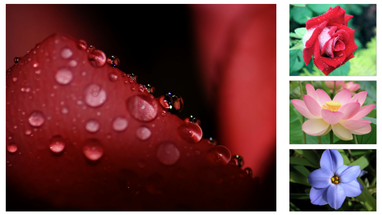
Justin Johnson*, Lamberto Ballan*, Li Fei-Fei
{jcjohns, lballan, feifeili}@cs.stanford.edu

SANTIAGO, CHILE DECEMBER 11-18, 2015
ICCV15
International Conference on Computer Vision

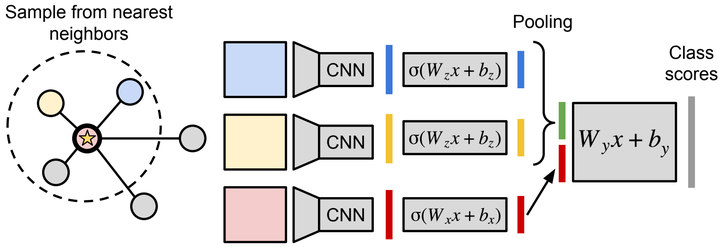
Motivation

On its own, the left image is ambiguous. In the context of a neighborhood of images with similar metadata, it is more clear that it shows a flower.

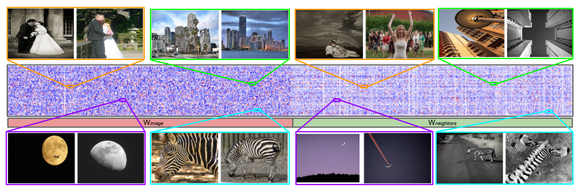
Our model uses this intuition to improve multilabel image annotation. We achieve state-of-the-art results on NUS-WIDE and can adapt to different types of metadata at test time.



Model



- Use Jaccard similarity on image metadata to compute nearest neighbors
- Sample from nearest neighbors to form an image neighborhood
- Use a CNN to extract features from the image and its neighborhood
- Use max-pooling to create a neighborhood feature vector
- Concatenate neighborhood features with image features
- Apply a linear transform to predict class scores
- Train with SGD + RMSprop



Learned weights. The model uses features from both the image and its neighbors. We show examples of images whose label scores are influenced more by the image and by its neighborhood; images with the same ground-truth labels are highlighted with the same colors. Images that are influenced by their neighbors tend to be non-canonical views.

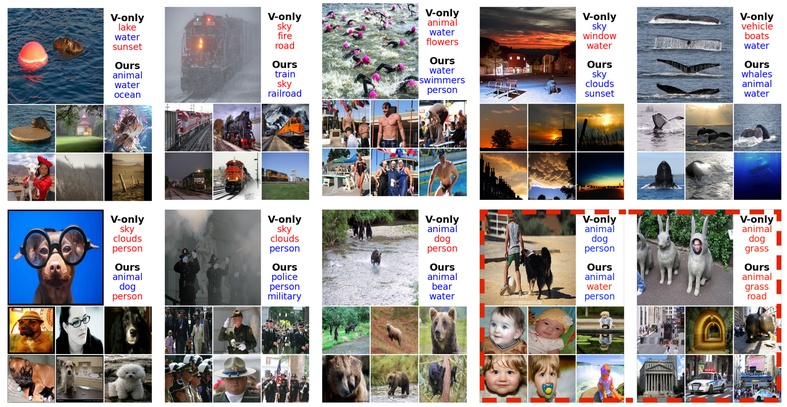
Acknowledgments

J. Johnson is supported by a Magic Grant from The Brown Institute for Media Innovation and L. Ballan is supported by a Marie Curie Fellowship from the EU (623930). We gratefully acknowledge the support of NVIDIA for their donation of GPUs and Yahoo for their donation of cluster machines used in this research.

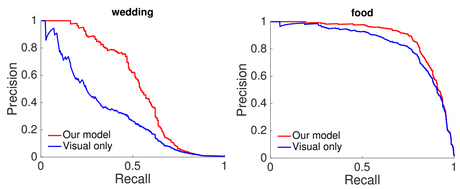
References

[15] Gong et al, ICLR 2014.
[27] Krizhevsky et al, NIPS 2012.
[36] Makadia et al, ECCV 2008.
[37] McAuley et al, ECCV 2012.

Results



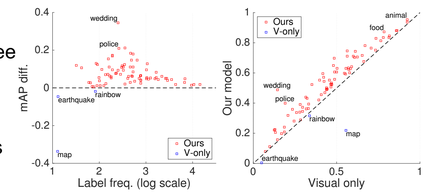
Example results. For each image we show the top 3 scoring labels using the visual-only (V-only) model and our model using tag nearest neighbors; correct labels are shown in blue and incorrect labels in red. We also show the 6 nearest neighbors to each image; its neighborhoods are drawn from these images. The red dashed lines show failure cases.



PR curves for our model and a CNN baseline. Our model shows large improvements for labels with high intra-class variability (e.g. wedding) and for labels where the visual model performs well (e.g. food).

Left: AP for each label of our model vs the baseline; we improve for all but three labels (map, earthquake, rainbow).

Right: difference in class AP between our model and the visual-only model vs label frequency.



Method	mAP_L	mAP_I	Rec_L	$Prec_L$	Rec_I	$Prec_I$
Tag-only Model + linear SVM [37]	46.67	-	-	-	-	-
Graphical Model (all metadata) [37]	49.00	-	-	-	-	-
CNN + softmax [15]	-	-	31.22	31.68	59.52	47.82
CNN + ranking [15]	-	-	26.83	31.93	58.00	46.59
CNN + WARP [15]	-	-	35.60	31.65	60.49	48.59
Upper bound	100.00±0.00	100.00±0.00	68.52±0.35	60.68±1.32	92.09±0.10	66.83±0.12
Tag-only + logistic	43.88±0.32	77.06±0.14	47.52±2.59	46.83±0.89	71.34±0.16	51.18±0.16
CNN [27] + kNN-voting [36]	44.03±0.26	73.72±0.10	30.83±0.37	44.41±1.05	68.06±0.15	49.49±0.11
CNN [27] + logistic (visual-only)	45.78±0.18	77.15±0.11	43.12±0.39	40.90±0.39	71.60±0.19	51.56±0.11
Image neighborhoods + CNN-voting	50.40±0.23	77.86±0.15	34.52±0.47	**56.05**±1.47	72.12±0.21	51.91±0.20
Our model: tag neighbors	52.78±0.34	**80.34**±0.07	43.61±0.47	46.98±1.01	74.72±0.16	**53.69**±0.13
Our model: tag neighbors + tag vector	**61.88**±0.36	80.27±0.08	**57.30**±0.44	54.74±0.63	**75.10**±0.20	53.46±0.09

Results on NUS-WIDE. Precision and recall are measured using n = 3 labels per image. Metrics are reported both per-label (mAP_L) and per-image (mAP_I). We run on 5 splits of the data and report mean and standard deviation.

Experiments

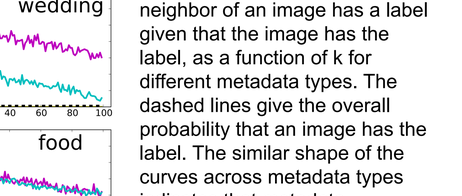
Hyperparameter Analysis

Method	mAP_L	mAP_I
CNN [27] + logistic (visual-only)	45.78±0.18	77.15±0.11
Our model: visual neighbors	47.45±0.19	78.56±0.14
Our model: group neighbors	48.87±0.22	79.11±0.13
Our model: set neighbors	48.02±0.33	78.40±0.25
Our model: tag neighbors	**52.78**±0.34	**80.34**±0.07

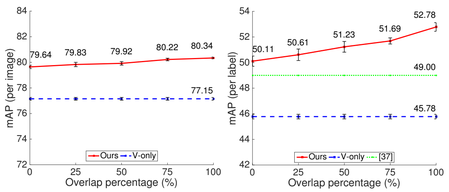
Performance of our model as we vary the type of metadata used to compute image neighborhoods vs the visual-only CNN baseline.

Left: Effect of neighborhood size (m) and maximum neighbor rank (M) on model performance.

Right: Effect of tag vocabulary size on model performance.



Vocabulary Generalization

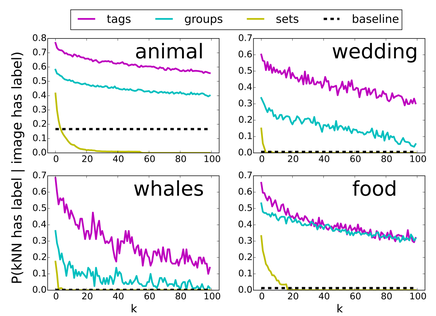


Performance as we vary overlap between tag vocabularies used for training and testing. Our model gives strong results even in the case of disjoint vocabularies.

Metadata Generalization

Train: \ Test:	Tags	Sets	Groups
Tags	52.78±0.34	47.12±0.35	48.14±0.33
Sets	52.21±0.29	48.02±0.33	48.49±0.16
Groups	50.32±0.28	47.82±0.24	48.87±0.22

Metadata generalization results: we use different types of metadata during training and testing, and report mAP_L.



Probability that the kth nearest neighbor of an image has a label given that the image has the label, as a function of k for different metadata types. The dashed lines give the overall probability that an image has the label. The similar shape of the curves across metadata types indicates that metadata generalization is possible.