Supplementary Material for "Learning Visual N-Grams from Web Data"

Ang Li*
University of Maryland
College Park, MD 20742, USA
angli@umiacs.umd.edu

Allan Jabri Armand Joulin Laurens van der Maaten
Facebook AI Research
770 Broadway, New York, NY 10025, USA
{ajabri,ajoulin,lvdmaaten}@fb.com

1. Introduction

The supplementary material for the submission "Learning Visual N-Grams for Web Data" is presented below. In Section 2, we provide all license information for all images from the YFCC100M dataset that were used in the main paper. In Section 3, we present quantitative results for image and caption retrieval on the COCO caption test set of $1,000$ images (COCO-1K). In Section 4, we present additional qualitative results of phrase prediction.

2. License Information for YFCC100M Photos

We reproduce all YFCC100M photos that appear in the main paper with relevant authorship and license information in Figure 1, 2, 3 and 4.

3. Relating Images and Captions: Additional Results

As an addition to the image and caption retrieval results on COCO-5K and Flickr-30K presented in the paper, we also provide retrieval results on the COCO-1K dataset, a test set of $1,000$ images provided by Karpathy and Fei-Fei [1]. In Table 1, we show the caption retrieval (left) and image retrieval (right) performance of four baseline models and our visual n-gram models on COCO-1K. We do not report results we obtained with the last version of the neural image captioning model [4] here because that model was trained on COCO validation set that was used as the basis for the COCO-1K test set.

The results on the COCO-1K dataset are in line with the results presented in the paper: our n-gram model performs roughly on par with recurrent language models [1, 3], but like these language models, it performs worse than models that were developed specifically for retrieval tasks [2, 5].

We provide additional results to demonstrate the effectiveness of end-to-end training. We trained a Jelinek-Mercer model on the ImageNet features as an additional baseline and compare it with the end-to-end Jelinek-Mercer model in COCO-5K. The results are shown in Table 2 which reveals that an end-to-end trained Jelinek-Mercer model outperforms the one trained with ImageNet features in both non-finetuning and finetuning modes.

*This work was done while Ang Li was at Facebook AI Research.

Predicted n-grams
lights
Burning Man
Mardi Gras
parade in progress

Predicted n-grams
GP
Silverstone Classic
Formula 1
race for the

Predicted n-grams
navy yard
construction on the
Port of San Diego
cargo

Figure 1. Four high-scoring visual n-grams for three images in our test set according to our visual n-gram model, which was trained *solely* on *unsupervised* web data. We selected the n-grams that are displayed in the figure from the five highest scoring n-grams according to our model, in such a way as to minimize word overlap between the n-grams. From top to bottom, photos are courtesy of: (1) Stuart L. Chambers (CC BY-NC 2.0); (2) Martin Pettitt (CC BY 2.0); (3) Gav Owen (C).

4. Phrase Prediction: Additional Results

We show additional qualitative results for predicting unigrams and bigrams in Figure 5; these examples were omit-

Figure 2. Four highest-scoring images for n-gram queries "Washington State", "Washington DC", "Washington Nationals", and "Washington Capitals" from a collection of $931,588$ YFCC100M test images. Washington Nationals is a Major League Baseball team; Washington Capitals is a National Hockey League hockey team. The figure only shows images from the YFCC100M dataset whose license allows reproduction. From the top-left photo in clockwise direction, the photos are courtesy of: (1) Colleen Lane (CC BY-ND 2.0); (2) Ryaninc (CC BY 2.0); (3) William Warby (CC BY 2.0); (4) Cliff (CC BY 2.0); (5) Boomer-44 (CC BY 2.0); (6) Dannebrog (CC BY-ND 2.0); (7) S. Yume (CC BY 2.0); (8) Bridget Samuels (CC BY-NC-ND 2.0); (9) David G. Steadman (Public Domain Mark 1.0); (10) Hockey Club Torino Bulls (CC BY 2.0); (11) Brent Moore (CC BY-NC 2.0); (12) Andrew Malone (CC BY 2.0); (13) Terren in Virginia (CC BY 2.0); (14) Guru Sno Studios (CC BY-ND 2.0); (15) Derek Hatfield (CC BY 2.0); and (16) Bruno Kussler Marques (CC BY 2.0).

Figure 3. Four highest-scoring images for n-gram queries "Market Street", "street market", "city park", and "Park City" from a collection of $931,588$ YFCC100M images. Market Street is a common street name, for instance, it is one of the main thoroughfares in San Francisco. Park City (Utah) is a popular winter sport destination. The figure only shows images from the YFCC100M dataset whose license allows reproduction. From left to right, photos are courtesy of the following photographers (license details between brackets. **Row 1:** (1) Jonathan Percy (CC BY-NC-SA 2.0); (2) Rachel Clarke (CC BY-NC-ND 2.0); (3) Richard Lazzara (CC BY-NC-ND 2.0); and (4) AboutMyTrip dotCom (CC BY 2.0). **Row 2:** (1) Alex Holyoake (CC BY 2.0); (2) Marnie Vaughan (CC BY-NC 2.0); (3) Hector E. Balcazar (CC BY-NC 2.0); and (4) Marcin Chady (CC BY 2.0). **Row 3:** (1) Rien Honnef (CC BY-NC-ND 2.0); (2) IvoBe (CC BY-NC 2.0); (3) Daniel Hartwig (CC BY 2.0); and (4) Benjamin Chodroff (CC BY-NC-ND 2.0). **Row 4:** (1) Guido Bramante (CC BY 2.0); (2) Alyson Hurt (CC BY-NC 2.0); (3) Xavier Damman (CC BY-NC-ND 2.0); and (4) Cassandra Turner (CC BY-NC 2.0).

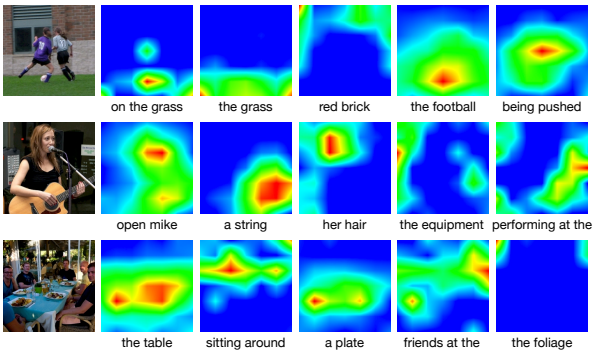

Figure 4. Discriminative regions of five n-grams for three images, computed using class activation mapping. From top to down, photos are courtesy of the following photographers (license details between brackets. **Row 1:** DebMomOf3 (CC BY-ND 2.0). **Row 2:** fling93 (CC BY-NC-SA 2.0). **Row 3:** Magnus (CC BY-SA 2.0).

COCO-1K	**Caption retrieval**			**Image retrieval**		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
Retrieval models						
Klein *et al*. [2]	38.9	68.4	80.1	25.6	60.4	76.8
Wang *et al*. [5]	50.1	79.7	89.2	39.6	75.2	86.9
Language models						
BRNN [1]	38.4	69.9	80.5	27.4	60.2	74.8
M-RNN [3]	41.0	73.0	83.5	29.0	42.2	77.0
Ours						
Naive n-gram	3.1	9.2	14.6	1.1	4.2	7.3
Jelinek-Mercer	22.5	47.6	60.7	12.8	33.5	46.5
J-M + finetuning	39.9	70.5	82.5	25.4	55.8	70.2

Table 1. Recall@k (for three cut-off levels k) of caption and image retrieval on the COCO-1K dataset for three baseline systems and our visual n-gram models (with and without finetuning). Baselines are separated in models dedicated to retrieval (top) and image-conditioned language models (bottom). Higher is better.

COCO-5K	Caption retrieval			Image retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
Imagenet + J-M	8.0	21.6	31.2	4.4	14.0	21.5
End-to-end J-M	8.7	23.1	33.3	5.0	14.5	21.9
Imagenet + J-M (finetuning)	12.7	31.0	43.0	6.5	18.9	28.1
End-to-end J-M (finetuning)	17.8	41.9	53.9	11.0	29.0	40.2

Table 2. Recall@k (for three cut-off levels k) of caption and image retrieval on the COCO-5K dataset for four variants of our visual n-gram models (with and without finetuning). Higher is better.

Unigrams	Bigrams
Sign	Neon sign
Bar	Motel in
Ave	Store in
Store	Sign for
Diner	Sacramento CA
Ferris	Ferris wheel
Blue	Lafayette Park
Wheel	Coney Island
Lafayette	Blue sky
Tower	Amusement park
Carriage	Horse drawn
Winter	Horse and
Horse	Winter in
Snow	Blizzard of
Blizzard	Snowy day
Times	Times Square
Shinjuku	Shinjuku Tokyo
Ginza	Manhattan new
Manhattan	Hong Kong
NYC	Eaton Center
Tokyo	Shinjuku Tokyo
Osaka	Tokyo Japan
Shinjuku	Vending machine
Vending	Osaka Japan
Store	Store in
Golden	Golden Gate
Marin	Suspension bridge
Suspension	Mackinac Island
Cruise	Oracle Team
Forth	Brooklyn Bridge

Figure 5. Five highest-scoring visual unigrams and bigrams for five images in our test set. From top to bottom, photos are courtesy of: (1) Mike Mozart (CC BY 2.0); (2) owlpacino (CC BY-ND 2.0); (3) brando.n (CC BY 2.0); (4) Laura (CC BY-NC 2.0); (5) inefekt69 (CC BY-NC-ND 2.0); and (6) Yahui Ming (CC BY-NC-ND 2.0).

ted from the main paper because of space limitations.

References

[1] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[2] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid Gaussian-Laplacian mixture models for image annotation. In *CVPR*, 2015.

[3] J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille. Deep captioning with multimodal recurrent neural networks. In *ICLR*, 2015.

[4] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE TPAMI*, 2016.

[5] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *CVPR*, 2016.