Supplementary Material

Andrej Karpathy Li Fei-Fei
Department of Computer Science, Stanford University
{karpathy,feifeili}@cs.stanford.edu

1. Magnitude modulation

An appealing feature of our alignment model is that it learns to modulate the importance of words and regions by scaling the magnitude of their corresponding embedding vectors. To see this, recall that we compute the image-sentence similarity between image k and sentence l as follows:

$$S_{kl} = \sum_{t \in g_l} max_{i \in g_k} v_i^T s_t. \quad (1)$$

Disciminative words. As a result of this formulation, we observe that representations of visually discriminative words such as *"kayaking, pumpkins"* tend to have higher magnitude in the embedding space, which translates to a higher influence on the final image-sentence scores due to the inner product. Conversely, the model learns to map stop words such as *"now, simply, actually, but"* near the origin, which reduces their influence. Table 1 show the top 40 words with highest and lowest magnitudes $\|s_t\|$.

Disciminative regions. Similarly, image regions that contain discriminative entities are assigned vectors of higher magnitudes by our model. This can be be interpreted as a measure of visual saliency, since these regions would produced large scores if their textual description was present in a corresponding sentence. We show the regions with high magnitudes in Figure 1. Notice the common occurrence of often described regions such as balls, bikes, helmets.

Figure 1. Flickr30K test set regions with high vector magnitude.

Magnitude	Word	Magnitude	Word
0.42	now	2.61	kayaking
0.42	simply	2.59	trampoline
0.43	actually	2.59	pumpkins
0.44	but	2.58	windsurfing
0.44	neither	2.56	wakeboard
0.45	then	2.54	acrobatics
0.45	still	2.54	sousaphone
0.46	obviously	2.54	skydivers
0.47	that	2.52	wakeboarders
0.47	which	2.52	skateboard
0.47	felt	2.51	snowboarder
0.47	not	2.51	wakeboarder
0.47	might	2.50	skydiving
0.47	because	2.50	guitar
0.48	appeared	2.50	snowboard
0.48	therefore	2.48	kitchen
0.48	been	2.48	paraglider
0.48	if	2.48	ollie
0.48	also	2.47	firetruck
0.48	only	2.47	gymnastics
0.48	so	2.46	waterfalls
0.49	would	2.46	motorboat
0.49	yet	2.46	fryer
0.50	be	2.46	skateboarding
0.50	had	2.46	dulcimer
0.50	revealed	2.46	waterfall
0.50	never	2.46	backflips
0.50	very	2.46	unicyclist
0.50	without	2.45	kayak
0.51	they	2.43	costumes
0.51	either	2.43	wakeboarding
0.51	could	2.43	trike
0.51	feel	2.42	dancers
0.51	otherwise	2.42	cupcakes
0.51	when	2.42	tuba
0.51	already	2.42	skijoring
0.51	being	2.41	firewood
0.51	else	2.41	elevators
0.52	just	2.40	cranes
0.52	ones	2.40	bassoon

Table 1. This table shows the top magnitudes of vectors ($\|s_t\|$) for words in Flickr30K. Since the magnitude of individual words in our model is also a function of their surrounding context in the sentence, we report the average magnitude.

978-1-4673-6964-0/15/$31.00 ©2015 IEEE

2. Alignment model

Learned appearance of text snippets. We can query our alignment model with a piece of text and retrieve individual image regions that have the highest score with that snippet. We show examples of such queries in Figure 2 and Figure 3. Notice that the model is sensitive to compound words and modifiers. For example, *"red bus"* and *"yellow bus"* give very different results. Similarly, *"bird flying in the sky"* and *"bird on a tree branch"* give different results. Additionally, it can be seen that the quality of the results deteriorates for less frequently occurring concepts, such as *"roof"* or *"straw hat"*. However, we emphasize that the model learned these visual appearances of text snippets from raw data of full images and sentences, without any explicit correspondences.

Additional alignment visualizations. See additional examples of inferred alignments between image regions and words in Figure 4. Note that one limitation of our model is that it does not explicitly handle or support counting. For instance, the last example we show contains the phrase *"three people"*. These words should align to the three people in the image, but our model puts the bounding box around two of the people. In doing so, the model may be taking advantage of the BRNN structure to modify the "people" vector to preferentially align to regions that contain multiple people. However, this is still unsatisfying because such spurious detections only exist as a result of an error in the RCNN inference process, which presumably failed to localize the individual people.

Web demo. We have published a web demo that displays our alignments for all images in the test set [1].

Additional Flickr8K experiments. We omitted ranking experiment results from our paper due to space constraints, but these can be found in Table 2

Counting. We experimented with losses that perform probabilistic inference in the forward pass that explicitly tried to localize exactly three distinct people in the image. However, this worked poorly because while the RCNN is good at finding people, it is not very good at localizing them. For instance, a single person can easily yield multiple detections (the head, the torso, or the full body, for example). We were not able to come up with a simple approach to collapsing these into a single detection (non-maxim suppression by itself was not sufficient in our experiments). Note that this ambiguity is partly an artifact of the training data. For example, torsos of people can often be labeled alone if the body is occluded. We are therefore lead to believe that this additional modeling step is highly non-trivial and a worthy subject of future work.

Plug and play use of Natural Language Processing toolkits. Before adopting the BRNN approach, we also tried to use Natural Language Processing toolkits to process the input sentences into graphs of noun phrases and their binary relations. For instance, in the sentence *"a brown dog is chasing a young child"*, the toolkit would infer that there are two noun phrases (*"a brown dog"*, *"young child"*), joined by a binary relationship of *"chasing"*. We then developed a CRF that inferred the grounding of these noun phrases to the detection bounding boxes in the image with a unary appearance model and a spatial binary model. However, this endeavor proved fruitless. First, performing CRF-like inference during the forward pass of a Neural Network proved to be extremely slow. Second, we found that there is surprisingly little information in the relative spatial positions between bounding boxes. For instance, almost any two bounding boxes in the image could correspond to the action of *"chasing"* due to huge amount of possibly camera views of a scene. Hence, we were unable to extract enough signal from the binary relations in the coordinate system of the image and suspect that more complex 3-dimensional reasoning may be required. Lastly, we found that NLP tools (when used out of the box) introduce a large amount of mistakes in the extracted parse trees, dependency trees and parts of speech tags. We tried to fix these with complex rules and exceptions, but ultimately decided to abandon the idea. We believe that part of the problem is that these tools are usually trained on different text corpora (e.g. news articles), so image captions are outside of their domain of competence. In our experience, adopting the BRNN model instead of this approach provided immediate performance improvements and produced significant reductions in code complexity.

3. Additional examples: Image annotation

Additional examples of generated captions on the full image level can be found in Figure 5 (and our website). The model often gets the right gist of the scene, but sometimes guesses specific fine-grained words incorrectly. We expect that reasoning not only the global level of the image but also on the level of objects will significantly improve these results. We find the last example (*"woman in bikini is jumping over hurdle"*) to be especially illuminating. This sentence does not occur in the training data. Our general qualitative impression of the model is that it learns certain templates, e.g. *"<noun>in <noun>is <verb>in <noun>"*, and then fills these in based on textures in the image. In this particular case, the volleyball net has the visual appearance of a hurdle, which may have caused the model to insert it as a noun (along with the woman) into one of its learned sentence templates.

[1] http://cs.stanford.edu/people/karpathy/deepimagesent/rankingdemo/

4. Additional examples: Region annotation

Additional examples of region annotations can be found in Figure 6. Note that we annotate regions based on the content of each image region alone, which can cause erroneous predictions when not enough context is available in the bounding box (e.g. a generated description that says "container" detected on the back of a dog's head in the image on the right, in the second row). We found that one effective way of using the contextual information and improving the predictions is to concatenate the fullframe feature CNN vector to the vector of the region of interest, giving 8192-dimensional input vector the to RNN. However, we chose to omit these experiments in our paper to preserve the simplicity of the mode, and because we believe that cleaner and more principled approaches to this challenge can be developed.

5. Training the Multimodal RNN

There are a few tricks needed to get the Multimodal RNN to train efficiently. We found that **clipping the gradients** (we only experimented with simple per-element clipping) at an appropriate value consistently gave better results and helped on the validation data. As mentioned in our paper, we experimented with SGD, SGD+Momentum, Adadelta, Adagrad, but found **RMSProp** to give best results. However, some SGD checkpoints usually also converged to nearby validation performance vicinity. Moreover, the distribution of the words in English language are highly non-uniform. Therefore, the model spends the first few iterations mostly learning the biases for the Softmax classifier such that it is predicting every word at random with the appropriate dataset frequency. We found that we could obtain faster convergence early in the training (and nicer loss curves) by explicitly **initializing the biases** of all words in the dictionary (in the Softmax classifier) to log probability of their occurrence in the training data. Therefore, with small weights and biases set appropriately the model right away predicts word at random according to their chance distribution. After submission of our original paper we performed additional experiments with comparing an RNN to an LSTM and found that **LSTMs** consistently produced better results, but took longer to train. Lastly, we initially used word2vec vectors as our word representations x_i, but found that it was sufficient to train these vectors from random initialization without changes in the final performance. Moreover, we found that the word2vec vectors have some unappealing properties when used in multimodal language-visual tasks. For example, all colors (e.g. red, blue, green) are clustered nearby in the word2vec representation because they are relatively interchangeable in most language contexts. However, their visual instantiations are very different.

References

[1] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 5

[2] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. *arXiv preprint arXiv:1406.5679*, 2014. 5

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

[4] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014. 5

[5] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *TACL*, 2014. 5

“chocolate cake”

“glass of wine”

“red bus”

“yellow bus”

“closeup of zebra”

“sprinkled donut”

“wooden chair”

“wooden office desk”

“shiny laptop”

Figure 2. Examples of highest scoring regions for queried snippets of text, on 5,000 images of our MSCOCO test set.

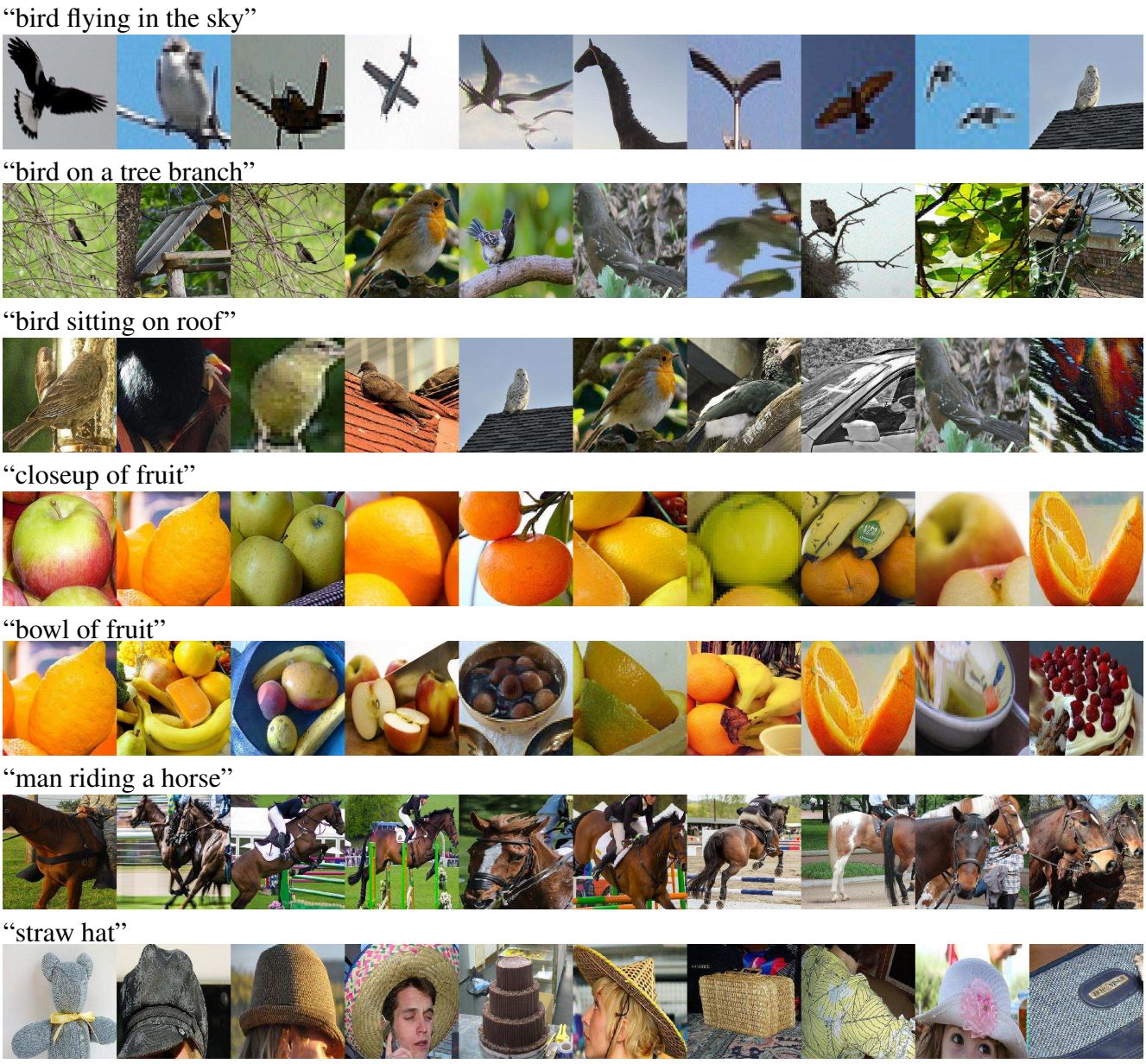

Figure 3. Examples of highest scoring regions for queried snippets of text, on 5,000 images of our MSCOCO test set.

	Image Annotation				Image Search			
Model	**R@1**	**R@5**	**R@10**	**Med *r***	**R@1**	**R@5**	**R@10**	**Med *r***
Flickr8K								
DeViSE (Frome et al. [1])	4.5	18.1	29.2	26	6.7	21.9	32.7	25
SDT-RNN (Socher et al. [5])	9.6	29.8	41.1	16	8.9	29.8	41.1	16
Kiros et al. [3]	13.5	36.2	45.7	13	10.4	31.0	43.7	14
Mao et al. [4]	14.5	37.2	48.5	11	11.5	31.0	42.4	15
DeFrag (Karpathy et al. [2])	12.6	32.9	44.0	14	9.7	29.6	42.5	15
Our implementation of DeFrag [2]	13.8	35.8	48.2	10.4	9.5	28.2	40.3	15.6
Our model: DepTree edges	14.8	37.9	50.0	9.4	11.6	31.4	43.8	13.2
Our model: BRNN	**16.5**	**40.6**	**54.2**	**7.6**	**11.8**	**32.1**	**44.7**	**12.4**

Table 2. Ranking experiment results for the Flickr8K dataset.

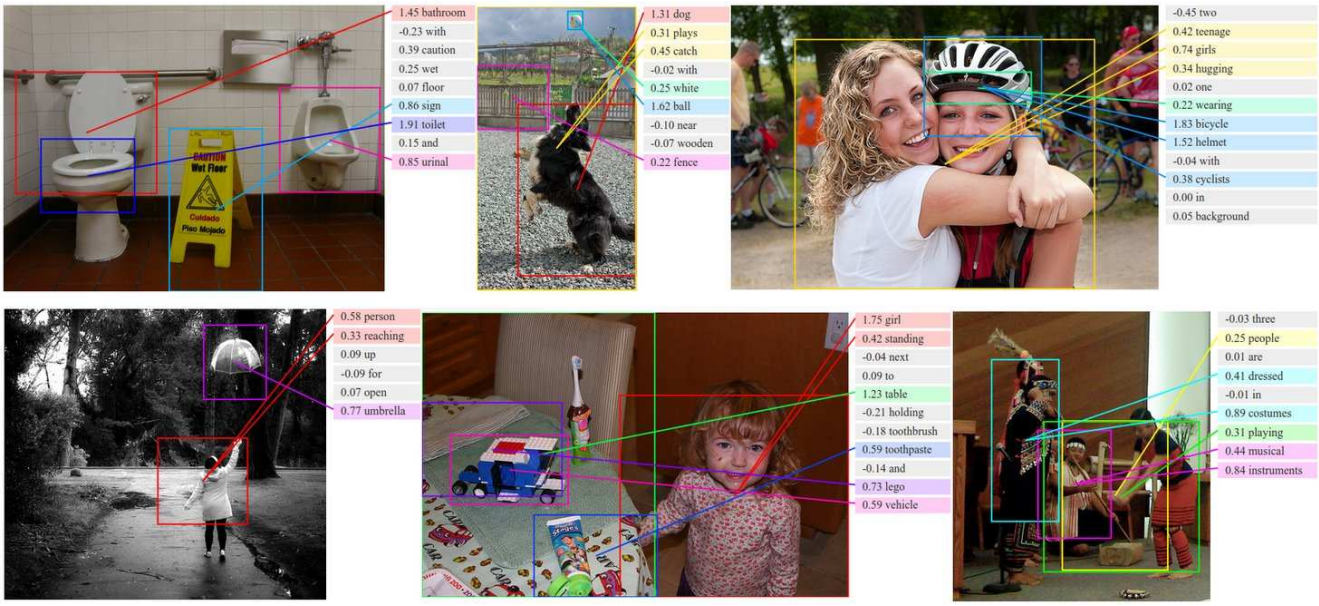

Figure 4. Additional examples of alignments. For each query test image above we retrieve the most compatible sentence from the test set and show the alignments.

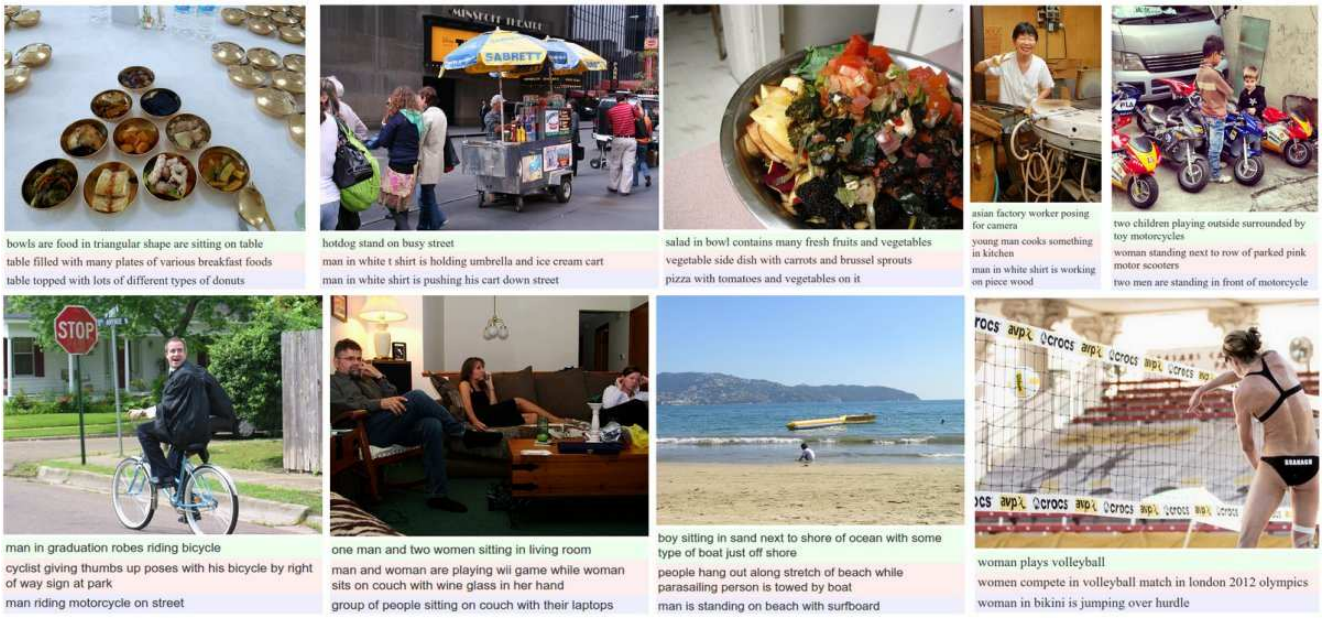

Figure 5. Additional examples of captions on the level of full images. Green: Human ground truth. Red: Top-scoring sentence from training set. Blue: Generated sentence.

Figure 6. Additional examples of region captions on the test set of Flickr30K.