Image Representations and New Domains in Neural Image Captioning

Jack Hessel
Computer Science Dept
Cornell University
jhessel@cs.cornell.edu

Nicolas Savva
Computer Science Dept
Cornell University
nss45@cornell.edu

Michael J. Wilber
Cornell Tech
Cornell University
mwilber@mjwilber.org

Abstract

We examine the possibility that recent promising results in automatic caption generation are due primarily to language models. By varying image representation quality produced by a convolutional neural network, we find that a state-of-the-art neural captioning algorithm is able to produce quality captions even when provided with surprisingly poor image representations. We replicate this result in a new, fine-grained, transfer learned captioning domain, consisting of 66K recipe image/title pairs. We also provide some experiments regarding the appropriateness of datasets for automatic captioning, and find that having multiple captions per image is beneficial, but not an absolute requirement.

1 Introduction

Describing the content of an image is an easy task for humans, but, until recently, had been difficult or impossible for computers. Recent work in computer vision has addressed this task of automatically generating the caption of an input image with promising results (Farhadi et al., 2010; Kulkarni et al., 2013; Ordonez et al., 2011; Karpathy and Li, 2014; Mao et al., 2014; Vinyals et al., 2014; Kiros et al., 2014; Donahue et al., 2014; Fang et al., 2014). Several state-of-the-art approaches couple a pre-trained deep convolutional neural network (CNN) for image representation with a recurrent neural network (RNN) to generate captions that describe image content.

We consider the possibility that the generation of these captions, however, is not heavily reliant upon the image representation input. For instance, if one was to train a RNN directly on image captions, one could learn a fair amount about the general language of image captions. Sutskever et al. (2011) demonstrate that RNNs are capable of producing diverse and surprisingly readable sentences, given a short starting sequence of seed words. Furthermore, non-neural memoization techniques like those proposed by Wood et al. (2009) and Gasthaus et al. (2010) are capable of producing very convincing language models for particular domains.

While it is clear that existing algorithms do discriminate based on image inputs, it is still unclear if the apparently highly specific generated captions are primarily a result of language modeling rather than image modeling. If it could be determined that either image modeling or language modeling is acting as the bottleneck in this multimodal setting, research efforts could be directed appropriately.

To examine the relative multimodal modeling capacities of existing neural captioning algorithms, we execute a series of experiments where we vary image representation quality produced from a fixed CNN, and examine how the output captions are affected.

For two existing datasets and a new domain we analyze here, our results suggest that caption quality does not scale well with increased classification accuracy of a fixed CNN. In fact, as the testing/validation accuracy of a CNN with fixed architecture increases, all seven caption evaluation metrics we consider appear to saturate at surprisingly low classification accuracies. While this does not prove that better image modeling algorithms could not produce better captions, it appears that many apparently fine-grained aspects of generated natural language are the result of surprisingly coarse grained visual distinctions.

For a fixed vision model, our results indicate that there is likely little room for caption improvement via gathering more training images alone. We further postulate that progress could be made

most quickly through the development of language modeling techniques that take better advantage of existing image representations. In particular, coupling our results with independent but consistent observations made by Karpathy and Li (2014) and Vinyals et al. (2014) regarding model modifications that lead to overfitting, it's very likely that overfitting language models to image features is still a big problem for many caption generation algorithms. Our analysis highlights what we believe to be an important question for these types of algorithms going forward: if better image representations contain useful, fine-grained information, is it possible to take advantage of that information without overfitting?

To supplement our analysis of image representations, we consider a new caption generating task: generating recipe titles based on images of food. The motivation for this new task results from the intuition that image representations might matter more in visually fine-grained domains, where algorithms must be able to discriminate between minute changes in the input images. We collect a dataset consisting of images of food coupled with recipe titles (e.g. "thai chicken curry") from Yummly.com for this purpose. When compared to captioning the coarse-grained ImageNet domain, the specificity of our food dataset calls for more subtle visual discrimination.

Instead of learning a food image representing CNN from scratch to derive representations, we apply transfer learning on a dataset of 101K food images. Using this approach, we significantly surpass current state-of-the-art performance for a classification task on this dataset, despite using a somewhat outdated deep architecture. We further demonstrate that this transfer learning process does indeed improve food captioning, though we observe a similar "flattening" of all linguistic evaluation metrics, after a point.

2 Related Work

2.1 Automatic Captioning

The model we choose to analyze in detail is the "Neural Image Captioning" (NIC) model detailed by Vinyals et al. (2014), though we believe the experiments we address here are relevant to researchers working on distinct but related models. In a similar fashion to Donahue et al. (2014) and Karpathy and Li (2014), NIC feeds a preclassification representation of images produced by an architecture like GoogLeNet (Szegedy et al., 2014) or AlexNet (Krizhevsky et al., 2012) to a LSTM recurrent neural network (Hochreiter and Schmidhuber, 1997) for language generation. The RNN weights are usually trained on datasets consisting of pairs of images and several corresponding human-generated annotations, such as Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), or Microsoft COCO (Lin et al., 2014). The CNN is often pre-trained on a very large set of images such as ImageNet (Deng et al., 2009) and held fixed while the RNN is trained. For many existing captioning datasets, ImageNet is a convenient starting point, presumably because images in most modern captioning datasets are of similar objects.

More complicated caption generation models have also demonstrated success on several datasets. To the knowledge of the authors, Fang et al. (2014) hold the current best result (in terms of BLEU-4) on the MSCOCO official captioning test set, though Vinyals et al. (2014) reportedly outperform Fang et al. on 2/5 evaluation metrics detailed on the MSCOCO captioning leaderboard.[1] Their pipeline involves training a language model directly on captions and a discretized image representation consisting of a likely set of objects in that image. Switching from a fine-tuned AlexNet (Krizhevsky et al., 2012) to a fine-tuned VGG-net (Simonyan and Zisserman, 2014) improved BLEU-4 by 2.4 points, and METEOR by 1.4 points. Because their image representations were discrete, it's possible that their language models were less prone to overfitting. It's not immediately obvious that a similar improvement would occur for language models that operate on extracted vector representations of images like NIC, however.

In contrast to the previous approaches that provide their RNNs with a representation of an image only at the first timestep, Mao et al. (2014) propose an extension of a single-layer RNN, dubbed the "multimodal RNN," that feeds a representation of an image to the RNN at *every* word generation step. Finally, Kiros et al. (2014) propose a model that first uses a CNN and an RNN to embed an image and its corresponding caption in the same semantic space, and then feeds vectors from this space into a "language generating structure content neural language model", an extension of a
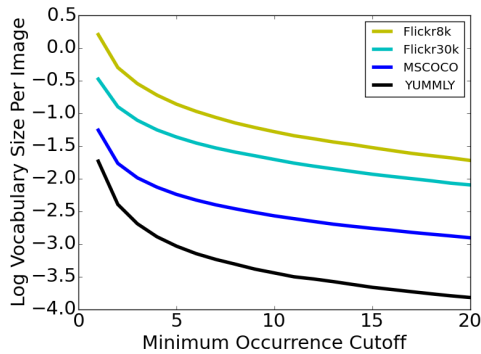

Figure 1: Word cutoff versus log-scale vocab size per image. This metric captures both dataset size and vocabulary size and shows that Yummly has the smallest vocabulary by a margin.

multiplicative RNN that "disentangles the structure of a sentence to its content."

Among models that directly input extracted features to a generating RNN, it is clear that image representations can be mishandled. Specifically, several authors note that passing image representations to the RNN at *every* timestep empirically leads to worse performance. While Karpathy and Li (2014) do not offer speculation as to why this is the case, Vinyals et al. (2014) briefly mention that this operation leads to over-fitting. These independent observations demonstrate that it is easy to overfit to image features.

2.2 Caption Evaluation Metrics

To evaluate captions, we use BLEU-{1,2,3,4} (Papineni et al., 2002) METEOR (Denkowski and Lavie, 2014) and CIDEr/CIDEr-D (Vedantam et al., 2014). BLEU-n is a precision measure over n-grams, whereas METEOR is a more sophisticated metric that involves the computation of an alignment between candidate and reference captions; both were originally conceived in the context of machine translation. CIDEr/CIDEr-D was created to evaluate captions of images and focuses on consensus, particularly in cases where there are multiple reference captions.

2.3 Recipe Title Prediction Tasks

To extend the scope of our investigation, we compile a dataset consisting of images of food coupled with recipe titles from `Yummly.com`. In this dataset, the title of a recipe is usually several words long and can be thought of as a "summary" of the image, rather than a direct description, as

not all image content is described in the caption. The image associated with "garlic butter shrimp," for instance, contains shrimp, a bowl, a lemon, and a human hand, and the captioning algorithms must learn to pick out which items are important to describe. Furthermore, there is less grammatical structure present in this dataset.

We view this task as distinct from existing captioning tasks for three reasons. First, the captions within Yummly are both short and restricted; a caption in the Yummly setting has an average length of 4.5 words, which is very low compared to Flickr or MSCOCO settings (both have an average of 10 words per caption) and the vocabulary is very small (see Figure 1). Second, to address this data fully, models must learn very fine-grained visual distinctions. Compared to the broad ImageNet domain, the Yummly images generally consist of some food item on a plate, coupled with several words from a small vocabulary. Finally, this dataset contains a single caption for each image, thus the learning task is more difficult. Previous work (Hodosh et al., 2013) has emphasized the importance of having multiple captions per image in a caption ranking setting, though its unclear if similar observations extend to a generation setting.

While we are only aware of the work of Malmaud et al. (2015) that address food in a multimodal fashion, Bossard et al. (2014) compile the Food 101 dataset which generalizes and increases the scale of previous food image datasets (i.e. Chen et al. (2009), Yang et al. (2010)). Their dataset includes 101k images of 101 types of foods and the task they address is classification.

2.4 Choosing a CNN/RNN Architecture

While substantial improvements have been made in terms of classification accuracy on ImageNet using increasingly deep architectures, we rely on the canonical neural network described in Krizhevsky et al. (2012) to generate our representations in most of our experiments. The use of AlexNet in particular allows for more direct comparison with previous work (i.e. Bossard et al. (2014)) and faster training time when compared to other deep models. This is beneficial particularly because our experiments are not specifically designed to produce state-of-the-art results.

We perform 20 random parameter searches to determine decent parameter settings using the
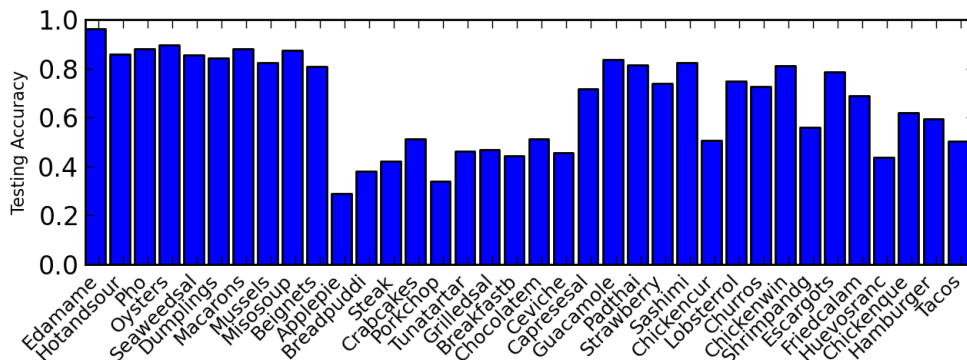
Figure 2: Transfer learned Food-101 CNN accuracy across various classes in the dataset, presented for easy comparison with Figure 6 in Bossard et al. (2014). In general, this model finds the same classes difficult to classify as the models described in previous work, suggesting that some types of fine-grained distinctions are difficult for many models.

Neuraltalk [2] library for all captioning experiments, selecting parameter settings resulting in the lowest validation set perplexity, unless specified otherwise. Settings we take as fixed include a minimum vocabulary threshold of 5, weight optimization using RMSprop (Tieleman and Hinton, 2012), and a hidden representation size of 256. We restrict our consideration to NIC because we believe it to be representative of the state-of-the-art in neural captioning. When we are evaluating models, we generate captions using a beam search of width 20. For the recipe title prediction evaluation, we include an end-of-caption token to avoid issues relating to predicted zero length captions; this has the result of artificially inflating evaluation metrics such that numerical cross-dataset comparisons are not valid.

2.5 Adapting the Food CNN through Transfer Learning

To represent food images properly, we find it appropriate to learn a model specific to the task of food recognition. Food-101 (Bossard et al., 2014) consists of only 101K images, which is a relatively low number of images to train a CNN from scratch. As such, we use a set of ImageNet-trained weights as initializations for our training of a CNN on the Food-101 classification task. This process is commonly referred to as transfer learning (Caruana, 1995; Bengio, 2012).

The intuition behind transfer learning in CNNs is that low-level features learned early on in the base network (which are generally observed to be

color blob and Gabor features (Yosinski et al., 2014)) are useful to networks trained on diverse classification tasks. Initializing the weights of the network to weights successful in another classification task should allow training of the new network to converge faster and to a better local optimum than if random initializations were used.

In fact, for the Food-101 dataset, we achieve a rank-1 accuracy of 66.80% when using transfer learning, when compared with the 56.40% rank-1 accuracy reported by Bossard et al. (2014) using the same AlexNet architecture; class-by-class accuracies are given in Figure 2 for comparison with previous work. Our network is learned using only 100k iterations of the Caffe library at a reduced learning rate, whereas training from scratch required Bossard et al. 450k iterations. For our tuning process, we follow the guidelines and parameter settings specified by the transfer learning example distributed with Caffe.[3]

Once the network is tuned, we compute 4096 dimensional vector representations for each image in Yummly dataset by extracting the network activations in the final fully-connected layer.

3 Yummly Dataset: Description and Baselines

After establishing that a CNN could be transfer learned to classify images of dishes at state-of-the-art performance, we were able to shift our focus to caption generation in a food domain.

The food dataset we collect contains roughly 66K recipes, each consisting of a single image-
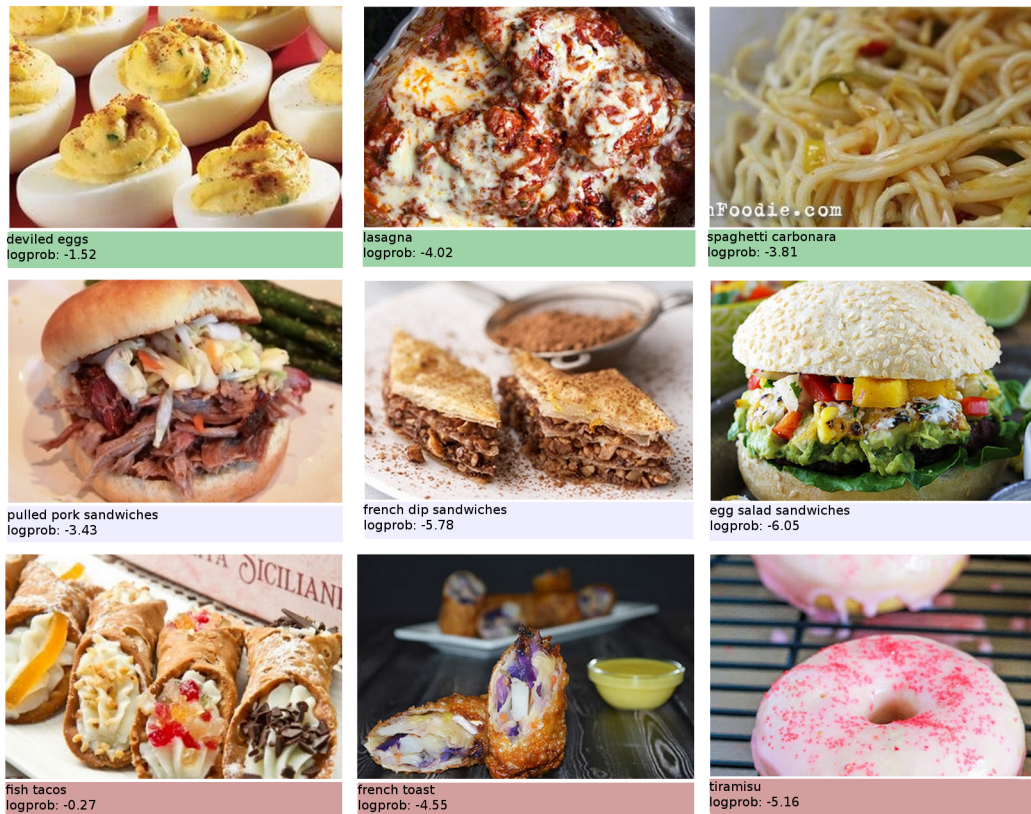

Figure 3: Examples of the captioning system output on several images. The first row of images represents images that are well captioned. The second row represents different types of images the system believes to be sandwiches. The third row represents images that the system has captioned incorrectly.

recipe pair. This data was taken from `Yummly.com`, a website that aggregates and performs analysis of millions of recipes. Out of the 66K recipes, 6K are reserved for testing, 6K are designated as a validation set, and the remaining 54K are used for model training.

This dataset differs from the Flickr datasets and MSCOCO both in terms of vocabulary and in terms of image content. The vocabulary size per image is smaller than any of the other datasets by a wide margin (see Figure 1). While it's clear the vision task requires more subtle distinction when compared to ImageNet, because the average caption length is shorter, it's ambiguous as to whether or not the Yummly language generation task is particularly "fine-grained."

3.1 Baseline Results

Table 1 presents some baseline results using the algorithms listed. Common-3 predicts a reasonable ordering of the three most common words ("with chicken and") for all captions. Nearest n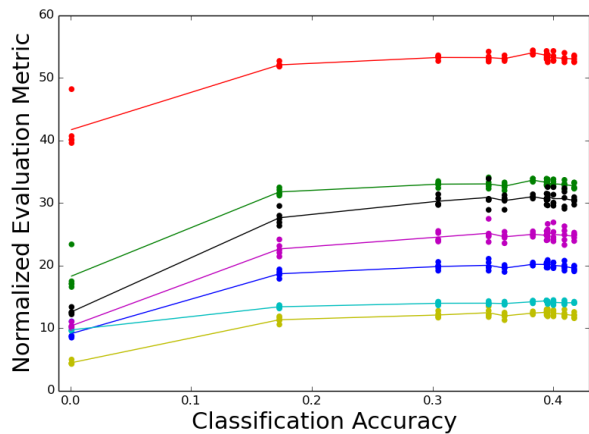eighbor predicts the caption of nearest neighbor in the transfer-learned 4096-dimensional embedding space. Common-Tri/Bi predict the most common tri/bigram in our dataset ("macaroni and cheese"/"ice cream") for all images.

Across the board, and particularly for BLEU-{2,3,4} scores, the caption generating programs outperform all baselines, which suggests the proposed task is adequately framed. However, it is worth noting that only roughly 300/6117 (roughly 5%) of generated captions are unique. This is rather low when compared with a representative result for Flickr8k, a dataset of similar size, where 200/1000 (roughly 20%) of generated captions are unique. It might be possible to re-frame the Yummly generation task as one of classification, however, it's not obvious how one might drive a fixed set of labels. In a later section we discuss whether or not only having one caption per image or other dataset features is a contributing factor to this result.
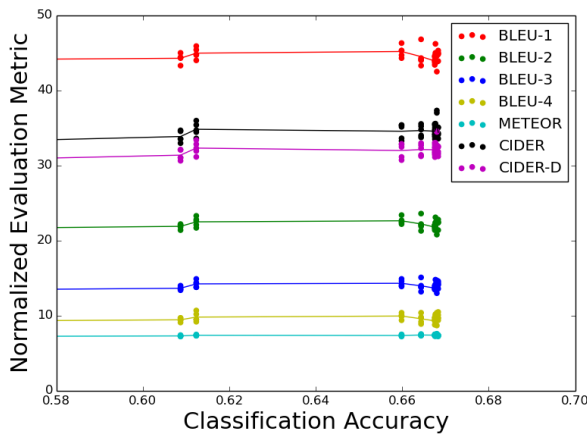
(a) Yummly: Transfer learned domain   (b) Flickr8k: Directly learned domain

Figure 4: Classification accuracy of CNN versus seven different normalized (100 is best possible) linguistic criteria for both the transfer learned (left) and directly learned (right) domains.

	B-1	B-2	B-3	B-4
Com-3	14.2	2.7	0.8	0.0
N-Neigh	20.5	2.5	0.6	0.0
Com-Tri	30.4	6.5	3.4	2.2
Com-Bi	35.4	8.9	5.2	0.0
Karpathy and Li (2014)	42.7	19.6	11.9	13.2
Vinyals et al. (2014)	46.2	23.1	14.8	10.2

Table 1: Yummly baseline BLEU-{1,2,3,4} scores for several baselines and two high performing language generation algorithms.

4 Image Representations

4.1 Experiment Descriptions

We vary image representation quality as follows: for the Flickr8k and Flickr30k datasets, we compute the representations given by snapshots of AlexNet taken mid-training on the ILSVRC2012 (Russakovsky et al., 2015) task. We use snapshots taken at intervals of 10k from 0k (random initialization) to 100k iterations. While this range of iterations is before the model has entirely converged, the rank-1 classification accuracy of the trained CNN over the ImageNet validation set increases from roughly 0% to over 40% during this time (after the model converges at 450k iterations, the rank-1 validation accuracy is 57%). From the standpoint of examining representation quality, this set of snapshots is important because this is likely where the network is learning most of its layer-by-layer abstractions, and the behavior of

the network after 100k iterations can be extrapolated based on the data we analyze here.

In a similar fashion, for Yummly we compute representations generated by snapshots of the transfer learned network at intervals of 10k from 0k to 90k, though our starting point is a fully-converged CNN that produces 57% rank-1 accuracy on ImageNet's validation set.

We train 5 NIC models from a random initialization per CNN for Flickr8k and Yummly, and 2-4 NIC models per CNN for Flickr30k. Every data point described in the following section is the result of up to six days of parallel computation using a modern 4/8-core machine. It should be noted that test/validation accuracy of these CNNs is not monotonically increasing with snapshot number. While the trend is that training CNNs for more iterations results in higher accuracy, there is some noise. For instance, for the Food-101 transfer learned CNN, rank-1 test accuracy drops from 61% to 60% over the snapshots extracted at 10k and 20k iterations respectively, before abruptly jumping to 66% testing accuracy in the next 10k iterations.

4.2 Results

We evaluate predicted captions using seven caption evaluation metrics, namely, BLEU-{1,2,3,4}, METEOR, and CIDEr/CIDEr-D. Figure 4 shows our main results for both the directly learned and transfer learned domains. In both cases, all captioning metrics appear to level off early, and do not improve significantly with increased classification rate after a point. This suggests that weight

settings for a fixed CNN with higher classification rates are unlikely to produce significantly better captions in terms of these seven evaluation metrics, after a point.

To quantify this lack of improvement, for each dataset we select a CNN that performs its associated visual classification task relatively poorly, and compare it to all better-classifying CNNs. For Flickr8k, for instance, we consider a CNN that produces 30.5% rank-1 accuracy on ImageNet's validation set, and compare its caption performance against that of 8 "better" CNNs that achieve between 34.6% and 41.7% accuracy; there are a total of 56 comparisons, in this case.

Though it is difficult to compute accurate statistics with only 5 observations in each group, we conduct three separate statistical tests, each with different variance/normality assumptions/efficiencies. The tests we perform are Students' t-test, Mann-Whitney U-test, and Welch's unpaired t-test.

In the case of Flickr8k, there are very few significant differences between the 30.5%-CNN and more accurate CNNs. In fact, in 14/56 cases (including half the time among BLEU-1/2 scores) the lower classifying CNN actually produced better captions. The results significant at the 5% level for any statistical test suggested that the 38%-CNN outperformed the 30.5%-CNN in terms of BLEU-1/2, and that the 39.5%-CNN outperformed the 30.5%-CNN in terms of METEOR.

The results for Flickr30k were very similar to the results for Flickr8k. In Figure 5 we present results from this dataset presented against CNN iteration number rather than CNN classification accuracy. We modify the presentation of our data simply to demonstrate that caption quality and iteration number (not just testing/validation accuracy) are also apparently independent after a point. No evidence of improvement was observed after the 30.5%-CNN, though only 2-4 observations per CNN could be made due to computational restrictions.

In total, in the directly-learned domain (Flickr8k/30k) all metrics appear to saturate after AlexNet reaches 30% classification accuracy over the ImageNet validation set. It is possible that training to convergence could result in slightly higher quality captions. However, our results indicate that efforts on ImageNet which result in less than a roughly 10% rank-1 classification
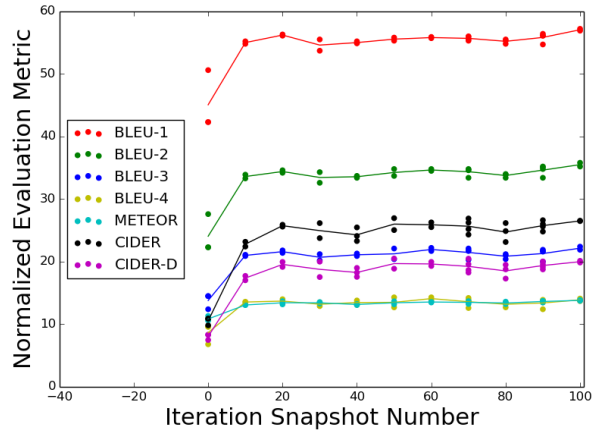

Figure 5: Caption quality versus CNN iteration (in thousands of iters) that representations were derived from. It is clear that a caption quality saturation happens very early on, and there is little to no improvement in captions as the CNNs are trained for more time.

accuracy increase for a fixed network are likely not worth undertaking if one's end goal is higher quality captions.

In the transfer learned domain, it is clear that domain adaptation improves caption quality, even after a small number of iterations. All statistical tests for all evaluation metrics indicate a highly significant difference ($p < .01$) between captions generated by a CNN trained directly on ImageNet, and one that has been transfer-learned using Food-101 for just 10K iterations (producing a rank-1 testing accuracy of 61.2% on that dataset). After a point, however, we observe the same independence of caption quality and classification accuracy.

It seems that "knowing more" about the image does not help the RNN generate more accurate captions after a point because the language patterns it learns are sufficient. This result is akin to prior work (e.g. Sutskever et al. (2011)) which demonstrates that RNNs are able to generate reasonable natural language, given a relatively weak seeding signal. The "weak" signal in this case is provided by image representations, rather than by a short sequence of starting words.

4.3 The Effect of Changing CNN Architectures

Our analysis thus-far has focused on a single image model, AlexNet, for extracting image representations. In this experiment, we compare the

captions generated on Flickr8k when using an improved CNN. We train 15 NIC models based on features extracted from a fully converged AlexNet, and 15 NIC models based on features extracted form a fully converged 16-layer VGGNet (Simonyan and Zisserman, 2014). The former model produces a 57.1% rank-1 accuracy over ImageNet's validation set, while the later outperforms this mark, producing 75.6% rank-1 validation accuracy. The default train/valiation/test split of 6k/1k/1k images is used for training.

Our results are summarized in Table 2. In addition to the seven caption evaluation metrics we've used in previous experiments, this table also includes the proportion of the 1k generated captions that are unique, and the train/validation perplexities.

Counter-intuitively, we find that, despite producing 18% lower rank-1 validation accuracy across ImageNet's validation set, AlexNet generates *better* captions than VGG net by all evaluation metrics. Notably, the models using VGG features produce lower perplexity across the validation split. Because we used validation perplexity as a metric for hyperparameter selection, it's likely that the VGG net models are overfitting to the particular Flickr8k validation split we used. However, the AlexNet trained models do not suffer a similar performance degradation. Here, it appears that not overfitting to image features is more important than taking advantage of very detailed image representations.

Our results from this experiment illustrate that better image representations might actually cause models like NIC to become more prone to overfitting. It's possible, too, that the early saturation of caption quality observed in the previous sections could be primarily due to overfitting. Future work would be well suited to evaluate different methods of hyperparameter selection.

4.4 One caption per image?

We conclude with a final experiment to address one potential shortcoming of domains similar to Yummly, where one is only able to extract a single caption per image. Though Yummly differs from the other datasets we explore in several ways (caption length/vocab size) a fundamental question arises from its examination: for a fixed amount of training data, is it better to have more captions per image, or more images with single captions? In

short, we hope to experimentally examine Hodosh et al.'s (2013) suggestion that having multiple captions per image is vital.

To address this question, we use Flickr30k, which provides five captions per image. We subset this dataset in two ways. In the first, we remove 4 captions randomly from each image in the training set, but keep all images (the "more images" method). In the second, we randomly remove 80% of training images, but keep all 5 captions for the remaining (the "more captions" method). This subsetting scheme is such that the overall number of image/caption pairs is the same between both methods, but the training data is of a different form.

We extract image representations from the ImageNet CNN at 100k iterations (which produces roughly 40% rank-1 classification accuracy over the ImageNet validation set) and train NIC on 6 random datasets constructed via the "more images" subsetting method, and 7 random datasets constructed via the "more captions" subsetting method. Finally, we generate captions and compare performance. A good hyperparameter setting for Flickr30k is borrowed from the random search conducted over the whole dataset experiments described in the previous section.

Our findings, summarized in Table 3, generally align with the accepted notion that having more captions and less images is better than having more images with single captions. For all seven evaluation metrics, the mean score for the models trained on the "more captions" datasets was greater than the mean score for the models trained on the "more images" datasets, and the results were significant at the 5% level for all three statistical tests in the case of BLEU-1 and BLEU-2. Interestingly, for CIDEr/CIDEr-D, the results were somewhat significant (all 6 p-values less than .15) but the results for METEOR were the least significant (all 3 p-values greater than .94).

The validation perplexity of the "more images" method is lower when compared to the more captions method, whereas the training perplexity is higher. Despite the fact that the output captions are better overall, this is an indication that having multiple captions per image can actually make NIC more prone to overfitting.

Finally, the NIC models trained on the "more caption" subsets produced higher proportions of unique captions on the test set. This suggests

	AlexNet	VGG
Top-1 ImageNet Val Acc	57.1%	75.6%
B-1	54.187	53.913
B-2	33.967	33.527
B-3**	20.640	20.007
B-4**	12.833	12.213
METEOR	14.559	14.559
CIDEr	32.416	31.362
CIDEr-D*	26.200	25.242
Proportion Unique***	20.5%	17.0%
Training Perplexity***	10.79	11.04
Validation Perplexity***	17.84	17.66

Table 2: Effect on caption quality when using the fully converged AlexNet and VGGNet on Flickr8k. Significance for all 3 statistical tests that there was a true difference between the subsetting techniques: ***$p < .001$, **$p < .01$, *$p < .05$

	More Captions	More Images
B-1**	55.167	54.243
B-2*	33.567	32.814
B-3	20.633	20.300
B-4	13.133	13.014
METEOR	13.105	13.096
CIDEr	21.428	20.418
CIDEr-D	16.350	15.550
Proportion Unique**	14.8%	9.96%
Training Perplexity**	14.69	16.01
Validation Perplexity*	25.86	25.33

Table 3: Evaluations for the NIC models trained on subsets of Flickr30k containing more captions (5 captions per image, 1/5 the total number of images) and more images (1 caption per image, all training images). Significance for all 3 statistical tests that there was a true difference between the subsetting techniques: **$p < .01$, *$p < .05$

that the single-caption per image feature of the Yummly dataset contributed to a lack of caption innovation.

Despite only having one caption per image, however, NIC was still able to produce good results on the single-captioned subsets. This indicates that quality captioning datasets can be built with only one caption per image. The number of additional images one needs to gather to compensate for this feature, however, is likely greater than the number of captions one would need to add to existing images.

5 Conclusion

We demonstrate the relationship between CNN classification accuracy and the quality of captions generated by a state of the art neural captioning algorithm. Training increasingly accurate image classifiers does not lead to better captions, after a point. This early saturation of caption quality is an indication that the performance of neural caption generating algorithms likely cannot be increased directly by producing more accurate CNNs. Furthermore, many of the apparently highly-specific generated captions output by models like NIC are likely due to language models capturing coarse grained information and generating corresponding plausible natural language sequences.

The role of overfitting to image features is dif-ficult to quantify. On one hand, there is extra information contained in image representations that NIC, for instance, does not take advantage of, and even commonly overfits to. However, it's not clear that this extra, fine-grained information is even worth taking into account. The success of models that generate language based on discretized image representations (e.g. (Young et al., 2014)) demonstrates that algorithms are capable of state-of-the-art performance without consideration of rich, real-valued vector features. It's likely that these types of models are less prone to overfitting, as well.

6 Acknowledgments

References

Yoshua Bengio. 2012. Deep learning of representations for unsupervised and transfer learning. *Unsupervised and*

37

Transfer Learning Challenges in Machine Learning, Volume 7, page 19.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*.

Rich Caruana. 1995. Learning many related tasks at the same time with backpropagation. *Advances in neural information processing systems*, pages 657–664.

Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. 2009. Pfid: Pittsburgh fast-food image dataset. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 289–292. IEEE.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

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

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John Platt, et al. 2014. From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer.

Jan Gasthaus, Frank Wood, and Yee Whye Teh. 2010. Lossless compression based on the sequence memoizer. In *Data Compression Conference (DCC), 2010*, pages 337–345. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*.

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

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, pages 1–42, April.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Ilya Sutskever, James Martens, and Geoffrey E Hinton. 2011. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *arXiv preprint arXiv:1411.5726*.

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

Frank Wood, Cédric Archambeau, Jan Gasthaus, Lancelot James, and Yee Whye Teh. 2009. A stochastic memoizer for sequence data. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1129–1136. ACM.

Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. 2010. Food recognition using statistics of pairwise local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2249–2256. IEEE.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.