Focused Evaluation for Image Description with Binary Forced-Choice Tasks

Micah Hodosh and **Julia Hockenmaier**
Department of Computer Science, University of Illinois
Urbana, IL 61801, USA
{mhodosh2, juliahmr}@illinois.edu

Abstract

Current evaluation metrics for image description may be too coarse. We therefore propose a series of binary forced-choice tasks that each focus on a different aspect of the captions. We evaluate a number of different off-the-shelf image description systems. Our results indicate strengths and shortcomings of both generation and ranking based approaches.

1 Introduction

Image description, i.e. the task of automatically associating photographs with sentences that describe what is depicted in them, has been framed in two different ways: as a natural language generation problem (where each system produces novel captions, see e.g. Kulkarni et al. (2011)), and as a ranking task (where each system is required to rank the same pool of unseen test captions for each test image, see e.g. Hodosh et al. (2013)).

But although the numbers reported in the literature make it seem as though this task is quickly approaching being solved (on the recent MSCOCO challenge,[1] the best models outperformed humans according to some metrics), evaluation remains problematic for both approaches (Hodosh, 2015).

Caption generation requires either automated metrics (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2014; Vedantam et al., 2015), most of which have been shown to correlate poorly with human judgments (Hodosh et al., 2013; Elliott and Keller, 2014; Hodosh, 2015) and fail to capture the variety in human captions, while human evaluation is subjective (especially when reduced to simple questions such as “Which is a better caption?”), expensive, and difficult to replicate. Ranking-based evaluation suffers from the problem that the pool of candidate captions may, on the one hand, be too small to contain many meaningful and interesting distractors, and may, on the other hand, contain other sentences that are equally valid descriptions of the image.

To illustrate just how much is still to be done in this field, this paper examines a series of binary forced-choice tasks that are each designed to evaluate a particular aspect of image description. Items in each task consist of one image, paired with one correct and one incorrect caption; the system has to choose the correct caption over the distractor. These tasks are inspired both by ranking-based evaluations of image description as well as by more recent work on visual question answering (e.g. Antol et al. (2015)), but differ from these in that the negatives are far more restricted and focused than in the generic ranking task. Since most of our tasks are simple enough that they could be solved by a very simple decision rule, our aim is not to examine whether models could be trained specifically for these tasks. Instead, we wish to use these tasks to shed light on which aspects of image captions these models actually “understand”, and how models trained for generation differ from models trained for ranking. The models we compare consist of a number of simple baselines, as well as some publicly available models that each had close to state-of-the-art performance on standard tasks when they were published. More details and discussion can be found in Hodosh (2015).

2 A framework for focused evaluation

In this paper, we evaluate image description systems with a series of binary (two-alternative) forced choice tasks. The items in each task consist of one image from the test or development part of the Flickr30K dataset (Young et al., 2014), paired

[1] http://mscoco.org/dataset/#captions-challenge2015

Proceedings of the 5th Workshop on Vision and Language, pages 19–28,
Berlin, Germany, August 12 2016. ©2016 Association for Computational Linguistics

“Switch People” Task

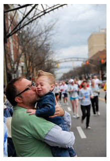
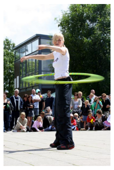
Image	Gold Caption	Distractor
	**a man** holding and kissing **a crying little boy** on the cheek	**a crying little boy** holding and kissing **a man** on the cheek
	**a woman** is hula hooping in front of **an audience**	**an audience** is hula hooping in front of **a woman**

Figure 1: The **“switch people”** task

“Replace Scene” Task

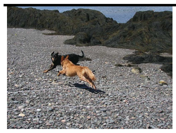
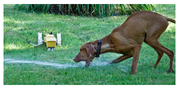
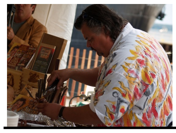
Image	Gold Caption	Distractor
	two dogs playing on **a beach**	two dogs playing on **frozen tundra**
	a brown dog is bending down trying to drink from **a jet of water**	a brown dog is bending down trying to drink from **your local brewery**
	a man in **a restaurant** having lunch	a man in **an office boardroom** having lunch

Figure 2: The **“replace scene”** task

with one correct and one incorrect caption, and the system has to choose (i.e. assign a higher score to) the correct caption over the distractor.

The correct caption is either an original caption or a part of an original caption for the image. Distractors are shorter phrases that occur in the original caption, complete captions for different images that share some aspect of the correct caption, or are artificially constructed sentences based on the original caption. While all distractors are constructed around the people or scene mentions in the original caption, each task is designed to focus on a particular aspect of image description. We focus on scene and people mentions because both occur frequently in Flickr30K. Unlike MSCOCO, all images in Flickr30K focus on events and activities involving people or animals. Scene terms (*“beach”, “city”, “office”, “street”, “park”*) tend to describe very visual, unlocalized components that can often be identified by the overall layout or other global properties of the image. At the same time, they restrict what kind of entities and events are likely to occur in the image. For instance, people do not “run” , “jump”, or “swim” in an “office”. Hence, models trained and tested on standard caption datasets do not necessarily need to model what “jumping in an office” might look like. We therefore suspect that much of the generic ranking task can be solved by identifying the visual appearance of scene terms.

Some tasks require the system to choose between two captions that provide similar descriptions of the main actor or the scene. In others, the distractor is not a full sentence, but consists only of the main actor or scene description. We also evaluate a converse task in which the distractor describes the scene correctly (but everything else in the sentence is wrong), while the correct answer consists only of the NP that describes the scene. Finally, we consider a task in which the distractor swaps two people mentions, reversing their corresponding semantic roles while keeping the same vocabulary.

3 Our tasks

Our first task (**switch people**, Fig. 1) identifies the extent to which models are able to distinguish sentences that share the same vocabulary but convey different semantic information. In this task, the correct sentences contain one person mention as the main actor and another person mention that occupies a different semantic role (e.g. “*A man holding a child*”). The distractors (“*A child holding a man*”) are artificially constructed sentences in which those two people mentions are swapped. This allows us to evaluate whether models can capture semantically important differences in word order, even when the bag-of-words representation of two captions is identical (and bag-of-words-based evaluation metrics such as BLEU1, ROUGE1 or CIDER would not be able to capture the difference either).

In the **replace person** and **replace scene** task (Fig. 2), distractors are artificially constructed sentences in which the main actor (the first person mention) or the scene chunk (which typically occurs at the end of the sentence) were replaced by different people or scene mentions. These tasks

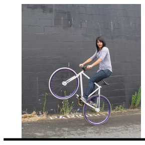
“Share Scene” Task		
Image	**GoldCaption**	**Distractor**
	a man in a suit and tie in a fancy building is speaking at **the podium**	a lady is giving a speech at **the podium**
	there is a woman riding a bike down **the road** and she popped a wheelie	two men in jeans and jackets are walking down **a small road**

Figure 3: The **“share scene”** task

aim to elicit how much systems are able to identify correct person or scene descriptions. Models should be able to understand when a person is being described incorrectly, even when the rest of the sentence remains correct. Similarly, since the scene is important to the overall understanding of the a caption, we wanted to make sure models grasp that changing the scene terms of a caption can drastically change its meaning.

The **share person** and **share scene** distractors (Fig. 3) are complete sentences from the training portion of Flickr30K whose actor or scene mentions share the same headword as the correct description for the test image. These tasks aim to elicit the extent to which systems focus only on the person or scene descriptions, while ignoring the rest of the sentence.

We also evaluate whether models are able to identify when a complete sentence is a better description of an image than a single NP. The **just person** and **just scene** distractors (Figs. 4 and 5) are NPs that consist only of the person or scene mentions of the correct description, and aim to identify whether systems prefer more detailed (correct) descriptions over shorter (but equally correct) ones. Finally, since systems can perform well on these tasks by simply preferring longer captions, we also developed a converse **just scene (+)** task, which pairs the (short, but correct) scene description with a (longer, but incorrect) sentence that shares the same scene.

3.1 Task construction

All our tasks are constructed around people and scene mentions, based on the chunking and the dictionaries provided in Plummer et al. (2015). Person mentions are NP chunks whose head noun refer to people (*“a tall man”*) or groups of people (*“a football team”*), or “NP1-of-NP2” constructions where the head of the first NP is a collective noun and the head of the second NP refers to people (*“a group of protesters”*). Subsequent NP chunks that refer to clothing are also included (*“a girl in jeans”*, *“a team in blue”*. Scene mentions are NP chunks whose head noun refers to locations (e.g. *“beach”*, *“city”*, *“office”*, *“street”*, *“park”*).

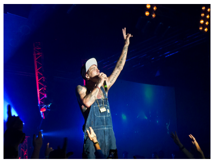
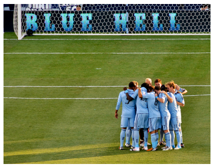
“Just Person” Task		
Image	**Gold Caption**	**Distractor**
	**a tattooed man wearing overalls** on a stage holding a microphone	**a tattooed man wearing overalls**
	**a team of soccer players** is huddled and having a serious discussion	**a team of soccer players**

Figure 4: The **“just person”** task

Switch people task We start with all captions of the 1000 development images that contain two distinct people mentions (excluding near-identical phrase pairs such as *“one man”*/*“another man”*). We filtered out examples in which the grammatical role reversal is semantically equivalent to the original (*“A man talking with a woman”*). Since we wished to maintain identical bag-of-words representations (to avoid differences between the captions that are simply due to different token frequencies) while focusing on examples that still remain grammatically correct (to minimize the effect of evaluating just how well a model generates or scores grammatically correct English text), we also excluded captions where one mention (e.g. the subject) is singular and the other (e.g. an object) is plural. When swapping two mentions, we also include the subsequent clothing chunks (e.g. *“man in red sweater”*) in addition to other premodifiers (*“a tall man”*). We automatically generate and hand prune a list of the possible permutations of the person chunks, resulting in 296 sentence pairs to use for evaluation.

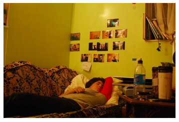
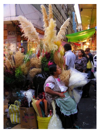
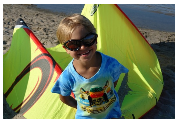
"Just Scene" Task		
Image	**Gold Caption**	**Distractor**
	a man sleeping in **a green room** on a couch	**a green room**
	a lady is sitting down tending to **her stand**	**her stand**
	a child poses wearing glasses near **water** outside	**water**

Figure 5: The **"just scene"** task

Replace person/scene tasks For the "replace person" task, we isolate person chunks, in both the training and development data. For each development sentence, we create a distractor by replacing each person chunk with a random chunk from the training data, resulting in 5816 example pairs to evaluate. For the "replace scene" task we created negative examples by replacing the scene chunk of a caption with another scene chunk from the data. Because multiple surface strings can describe the same overall scene, we use the training corpus to calculate which scene chunk's headwords can cooccur in the training corpus. We avoid all such replacements in order to ensure that the negative sentence does not actually still describe the image. In theory, this should be a baseline that all state-of-the-art image description models excel at.

Share person/scene tasks Here, the distractors consist of sentences from the Flickr30K training data which describe a similar main actor or scene as the correct caption. For each sentence in the development data, we chose a random training sentence that shares the same headword for its "actor" chunk, resulting in 4595 items to evaluate. We did the same for development sentences that mention a scene term, resulting in 2620 items.

Just person/scene tasks Finally, the "just person" and "just scene" tasks require the models to pick a complete sentence (again taken from the development set) over a shorter noun phrase that is a substring of the correct answer, consisting of either the main actor or the scene description. Although the distractors are not wrong, they typically only convey a very limited amount of information about the image, and models should prefer the more detailed descriptions provided by the complete sentences, as long as they are also correct. But since these tasks can be solved perfectly by any model that consistently prefers longer captions over shorter ones, we also investigate a converse "just scene (+)" task; here the correct answer is a noun phrase describing the scene, while the distractor is another full sentence that contains the same scene word (as in the "share scene" task). Taken together, these tasks allow us to evaluate the extent to which models rely solely on the person or scene description and ignore the rest of the sentence.

4 The Models

We evaluate generation and ranking models that were publicly available and relatively close in performance to state of the art, as well as two simple baselines.

Generation models Our baseline model for generation (**Bigram LM**) ignores the image entirely. It returns the caption that has a higher probability according to an unsmoothed bigram language model estimated over the sentences in the training portion of the Flickr30K corpus.

As an example of an actual generation model for image description, we evaluate a publicly available implementation[2] of the generation model originally presented by Vinyals et al. (2015) (**Generation**). This model uses an LSTM (Hochreiter and Schmidhuber, 1997) conditioned on the image to generate new captions. The particular instance we evaluate was trained on the MSCOCO dataset (Lin et al., 2014), not Flickr30K (leading to a possible decrease in performance on our tasks) and uses VGGNet (Simonyan and Zisserman, 2014) image features (which should account for a significant jump in performance over the previously published results of Vinyals et al. (2015)). Works such as Vinyals et al. (2015) and Mao et al. (2014) present models that are developed for the generation task, but renormalize the probability that their models assign to sentences when they apply them to ranking tasks (even though their models include stop probabilities that should enable them to directly compare sentences of different lengths). To examine the ef-

[2] http://cs.stanford.edu/people/karpathy/neuraltalk/

fect of such normalization schemes, we also consider normalized variants of our two generation models in which we replace the original sentence probabilities by their harmonic mean. We will see that the unnormalized versions of these models tend to perform poorly when the gold caption is measurably longer than the distractor term, and well in the reverse case, while normalization attempts to counteract this trend.

Ranking models Ranking models learn embeddings of images and captions in the same space, and score the affinity of images and captions in terms of their Euclidian distance in this space. We compare the performance of these generation models with two (updated) versions of the ranking model originally presented by Kiros et al. (2014)[3] (**LSTM Ranking**), one trained on MSCOCO, and the other on Flickr30K. This model uses an LSTM to learn the embedding of the captions. While the Flickr30K trained model should be more appropriate for our test data, the MSCOCO trained model might be more directly comparable to the generation model of Vinyals et al. A comparison between the two variants can offer insight into the degree of domain shift between the two datasets.

Our ranking baseline model (**BOW Ranking**) replaces the LSTM of Kiros et al. (2014) with a simple bag-of-words text representation, allowing us to examine whether the expressiveness of LSTMs is required for this task. We use the average of the tokens' GloVe embeddings (Pennington et al., 2014) as input to a fully connected neural network layer that produces the final learned text embedding[4]. More formally, for a sentence consisting of tokens $w_1...w_n$, GloVe embeddings $\phi()$, and a non-linear activation function σ_w, we define the learned sentence embedding as $F(w_1...w_n) = \sigma_w(W_w \cdot (\frac{1}{n}) \sum_i \phi(w_i) + b_w)$. Similarly, the embedding of an image represented as a vector p is defined as $G(p) = \sigma_i(W_i \cdot p + b_i)$. We use a ranking loss similar to Kiros et al. (2014) to train the parameters of our model, $\theta = (W_w, W_i, b_w, b_i)$. We define the distance of the embeddings of image i and sentence s as $\Delta(i, s) = \cos(F(i), G(s))$. Using S to refer to the set of sentences in the training data, S_i for the training sentences associated with image i, S_{-i} for the set of sentences not associated with i, I for the set of training images, I_s for the image associated with sentence s, and I_{-s} for the set of all other training images, and employing a free parameter m for the margin of the ranking, our loss function is:

$$\begin{aligned} L(\theta) = & \sum_{i \in I, s \in S_i, s' \in S_{-i}} \max(0, m - \Delta(i,s) + \Delta(i,s')) \\ & + \sum_{s \in S, i \in I_s, i' \in I_{-s}} \max(0, m - \Delta(i,s)) + \Delta(i',s))) \end{aligned}$$

As input image features, we used the 19 layer VGGNet features (Simonyan and Zisserman, 2014), applied as by Plummer et al. (2015). We first process the GloVe embeddings by performing whitening through zero-phase component analysis (ZCA) (Coates and Ng, 2012) based on every token appearance in our training corpus. We set σ_w to be a ReLU and simply use the identity function for σ_i (i.e. no non-linearity) as that resulted in the best validation performance. We train this model on the Flickr30K training data via stochastic gradient descent, randomly sampling either 50 images (or sentences), and randomly sampling one of the other training sentences (images). We adjust the learning rate of each parameter using Adagrad (Duchi et al., 2010) with an initial learning rate of 0.01, a momentum value of 0.8, and a parameter decay value of 0.0001 for regularization.

5 Results

Results for all tasks can be found in Table 1.

The "switch people" task The generation models are much better than the ranking models at capturing the difference in word order that distinguishes the correct answer from the distractor in this task. At 52% accuracy, the ranking models perform only marginally better than the ranking baseline model, which ignores word order, and therefore performs at chance. But the 69% accuracy obtained by the generation models is about the same as the performance of the bigram baseline that ignores the image. This indicates that neither of the models actually "understands" the sentences (e.g. the difference between men carrying children and children carrying men), although generation models perform significantly better than chance because they are often able to distinguish the more common phrases that occur in the correct answers ("man carries child") from those that appear in the constructed sentences that serve as distractors here ("child carries man"). It

[3] `https://github.com/ryankiros/visual-semantic-embedding`

[4] Deeper and more complex representations showed no conclusive benefit

	Switch	**Replace**	**Replace**	**Share**	**Share**	**Just**	**Just**	**Just**
	People	**Person**	**Scene**	**Person**	**Scene**	**Person**	**Scene**	**Scene(+)**
# of pairs	296	5816	2513	4595	2620	5811	2624	2620
Bigram LM	**69.8**	83.0	77.5	49.6	47.9	1.1	0.0	**99.6**
Normalized Bigram LM	**69.8**	69.9	76.5	50.2	50.9	31.3	28.2	71.0
Generation (COCO)	69.3	**85.2**	85.2	56.5	54.7	3.8	7.4	94.2
Normalized Generation (COCO)	68.9	74.0	85.5	61.6	59.2	79.5	**97.3**	5.5
BOW Ranking (Flickr30K)	50.0	84.9	**89.3**	**93.6**	**89.9**	81.2	84.6	71.3
LSTM Ranking (COCO)	52.0	79.4	86.6	89.9	88.0	79.8	86.5	58.2
LSTM Ranking (Flickr30K)	52.0	81.1	87.0	92.5	89.3	**82.6**	78.8	75.5

Table 1: Accuracies of the different models on our tasks

seems that localization of entities (Plummer et al., 2015; Xu et al., 2015) may be required to address this issue and go beyond baseline performance.

The "replace person/scene" tasks On the "replace person" task, the (unnormalized) bigram baseline has a relatively high accuracy of 83%, perhaps because the distractors are again artifically constructed sentences. The ranking baseline model and the (unnormalized) generation model outperform this baseline somewhat at around 85%, while the ranking models perform below the bigram baseline. The ranking model trained on Flickr30K has a slight advantage over the same model trained on MSCOCO, an (unsurprising) difference that also manifests itself in the remaining tasks, but both models perform below the ranking baseline. Normalization hurts both generation models significantly. It is instructive to compare performance on this task with the "replace scene" task. We see again that normalization hurts for generation, while the baseline ranking model outperforms the more sophisticated version. But here, all models that consider the image outperform the bigram model by a very clear eight to almost twelve percent. This indicates that all image description models that we consider here rely heavily on scene or global image features. It would be interesting to see whether models that use explicit object detectors could overcome this bias.

The "share person/scene" tasks The distractors in these tasks are captions for other images that share the same actor or scene head noun. Since the bigram language models ignore the image, they cannot distinguish the two cases (it is unclear why the unnormalized bigram model's accuracy on the "share scene" task is not closer to fifty percent). And while normalization helps the generation model a little, its accuracies of 61.6% and 59.2% are far below those of the ranking models, indicating that the latter are much better at distinguishing between the correct caption and an equally fluent, but incorrect one. This is perhaps not surprising, since this task is closest to the ranking loss that these models are trained to optimize. By focusing on an adversarial ranking loss between training captions, the ranking model may be able to more correctly pick up important subtle differences between in-domain images, while the generation model is not directly optimized for this task (and instead has to also capture other properties of the captions, e.g. fluency). With an accuracy of 93.6% and 89.9%, the bag-of-word ranking baseline model again outperforms the more complex LSTM. But examining its errors is informative. In general, it appears that it makes errors when examples require more subtle understanding or are atypical images for the words in the caption, as shown in Figure 6.

The "just person/scene" tasks The "just person" and "just scene" tasks differ from all other tasks we consider in that the distractors are also correct descriptions of the image, although they are consistently shorter. To actually solve these tasks, models should be able to identify that the additional information provided in the longer caption is correct. By contrast, the "just scene (+)" task requires them to identify that the additional information provided in the longer caption is not correct. But a simple preference for longer or shorter captions can also go a long way towards "solving" these tasks. In this case, we would expect to see a model's performance on the "just scene" task to be close to the complement of its performance on the converse"just scene (+)" task. This is indeed the case for the bigram and the generation models (but not for the ranking models). This preference is particularly obvious in the case of the unnormalized bigram model (which

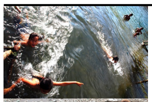
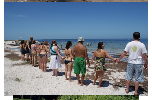
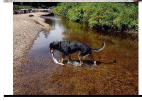
Image	Gold Caption	Distractor: Shares a Scene
	a group of children in the ocean (0.194)	a person in a kayak rides waves in the ocean (0.344)
	two women are sitting in ditches of dirt with two buckets and a purse close by (0.378)	the young toddlers is dressed in yellow and purple while sitting on the ground with three bucks filling them with dirt (0.393)
	a group of people hold hands on the beach (0.609)	a group of people are lounging at a beach (0.613)
	a dog drags a white rag through an almost dried up creek (0.330)	a dog jumps over a creek (0.433)

Figure 6: Examples from the "share scene" task that the BOW ranking model gets wrong, together with its scores for each of the captions.

does not take the image into account), and, to a slightly lesser extent, by the unnormalized generation model (which does). Both models have near perfect accuracy on the "just scene (+)" task, and near complete failure on the other two tasks. Length normalization reduces this preference for shorter captions somewhat in the case of the bigram model, and seems to simply reverse it for the generation model. None of the ranking models show such a marked preference for either long or short captions. But although each model has similar accuracies on the "just scene" and on the "just scene (+)" task, accuracies on the "just scene" task are higher than on the "just scene (+)" task. This indicates that they are not always able to identify when the additional information is incorrect (as in the "just scene (+)" task). Accuracies on the "just person" task tend to be lower, but are otherwise generally comparable to those on the "just scene" task. We see the biggest drops for the length-normalized generation model, whose accuracy goes down from 97.3% on the scene task to 79.5% (indicating that something else besides a preference for longer captions is at play), and the MSCOCO-trained ranking model which goes down from 86.5% to 79.8%.

It is unclear why the performance on the "just person" task tends to be lower than on the "just scene" task. Since scenes correspond to global image properties, we stipulate that models are better at identifying them than most people terms. Although some people descriptions (e.g. "baseball player", "audience") are highly indicative of the scene, this is not the case for very generic terms ("man", "woman"). We also note that identifying when the additional information is correct can be quite difficult. For example, in the second example in Figure 4, the phrase "huddled and having a serious discussion" has to be understood in the context of soccer. While the dataset contains other images of discussions, there are no other instances of discussions taking place on soccer fields, and the people in those cases tend to occupy a much larger portion of the image. Further analyzing and isolating these examples (and similar ones) is key for future progress. Figure 7 shows items from the "just scene" task that the BOW model gets right, paired with items for the same image where it makes a mistake. For the first item, it seems that the model associates the terms "crowd" or "crowded" with this image (while not understanding that "busy" is synonymous with "crowded" in this context). The error on the second item may be due to the word "rock" in the correct answer (Flickr30K contains a lot of images of rock climbing), while the error on the fourth item may be due to the use of words like "parents" rather than the more generic "people."

5.1 Discussion

We compared generation models for image description, which are trained to produce fluent descriptions of the image, with ranking-based models, which learn to embed images and captions in a common space in such a way that captions appear near the images they describe. Among the models we were able to evaluate, ranking-based approaches outperformed generation-based ones on most tasks, and a simple bag-of-words models per-

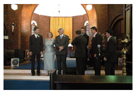
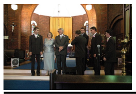
Image	Gold Caption	Distractor: A Scene Chunk
	a single man in a black tshirt standing above the crowd at a busy bar (0.329)	a busy bar (0.203)
	a man is making a rock gesture while standing on a stool in a crowded bar (0.216)	**a crowded bar (0.327)**
	**some people in formal attire stand in front of the altar in a church sanctuary (0.434)**	a church sanctuary (0.325)
	a son and his parents are taking a group picture in a church (0.274)	**a church (0.399)**

Figure 7: Items from the “Just Scene” task with the scores from the BOW ranking model in parentheses (bold = the caption preferred by the model).

formed similarly to a comparable LSTM model.

The “switch people” results indicate that ranking models may not capture subtle semantic differences created by changing the word order of the original caption (i.e. swapping subjects and objects). But although generation models seem to perform much better on this task, their accuracy is only as good as, or even slightly lower than, that of a simple bigram language model that ignores the image. This indicates that generation models may have simply learned to distinguish between plausible and implausible sentences.

The “share person/scene” and “just person/scene” results indicate that ranking models may be better at capturing subtle details of the image than generation models. But our results also indicate that both kinds of models still have a long way to before they are able to describe images accurately with a “human level of detail.”

Our comparison of the LSTM-based model of Kiros et al. (2014) against our bag-of-words baseline model indicates that the former may not be taking advantage of the added representational power of LSTMs (in fact, most of the recent improvements on this task may be largely due to the use of better vision features and dense word embeddings trained on large corpora). However, RNNs (Elman, 1990) and LSTMs offer convenient ways to define a probability distribution across the space of all possible image captions that cannot be modeled as easily with a bag-of-words style approach. The question remains if that convenience comes at a cost of no longer being able to easily train a model that understands the language to an acceptable amount of detail. It is also important to note that we were unable to evaluate a model that combines a generation model with a reranker such Fang et al. (2014) and the follow up work in Devlin et al. (2015). In theory, if the generation models are able produce a significantly enough diverse set of captions, the reranking can make up the gap in performance while still being able to generate novel captions easily.

6 Conclusion

It is clear that evaluation still remains a difficult issue for image description. The community needs to develop metrics that are more sensitive than the ranking task while being more directly correlated to human judgement than current automated metrics used for generation. In this paper, we developed a sequence of binary forced-choice tasks to evaluate and compare different models for image description. Our results indicate that generation and ranking-based approaches are both far from having “solved” this task, and that each approach has different advantages and deficiencies. But the aim of this study was less to analyze the behavior of specific models (we simply used models whose performance was close to state of the art, and whose implementations were available to us) than to highlight issues that are not apparent under current evaluation metrics, and to stimulate a discussion about what kind of evaluation methods are appropriate for this burgeoning area. Our data is available,[5] and will allow others to evaluate their models directly.

[5] http://nlp.cs.illinois.edu/HockenmaierGroup/data.html

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grants No. 1205627, 1405883 and 1053856. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2425–2433.

Adam Coates and Andrew Y. Ng. 2012. Learning feature representations with k-means. In *Neural Networks: Tricks of the Trade - Second Edition*, pages 561–580.

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*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 2: Short Papers*, pages 100–105.

John Duchi, Elad Hazan, and Yoram Singer. 2010. Adaptive subgradient methods for online learning and stochastic optimization. Technical Report UCB/EECS-2010-24, EECS Department, University of California, Berkeley, Mar.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland, June.

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2014. From captions to visual concepts and back. *CoRR*, abs/1411.4952.

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

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artifical Intellegence*, 47:853–899.

Micah Hodosh. 2015. *Natural language image description: data, models, and evaluation*. Ph.D. thesis, University of Illinois.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2011. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1608.

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 *European Conference on Computer Vision (ECCV)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, July.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In *The IEEE International Conference on Computer Vision (ICCV)*, December.

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

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3156–3164.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.

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.