Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Kelvin Xu[⋆] KELVIN.XU@UMONTREAL.CA
Jimmy Lei Ba[†] JIMMY@PSI.UTORONTO.CA
Ryan Kiros[†] RKIROS@CS.TORONTO.EDU
Kyunghyun Cho[⋆] KYUNGHYUN.CHO@UMONTREAL.CA
Aaron Courville[⋆] AARON.COURVILLE@UMONTREAL.CA
Ruslan Salakhutdinov[†∗] RSALAKHU@CS.TORONTO.EDU
Richard S. Zemel[†∗] ZEMEL@CS.TORONTO.EDU
Yoshua Bengio[⋆∗∗] YOSHUA.BENGIO@UMONTREAL.CA

⋆ Université de Montréal, † University of Toronto, ∗ CIFAR

Abstract

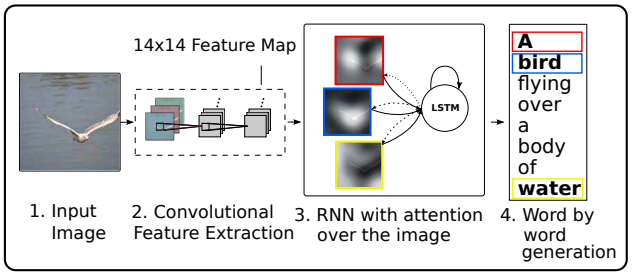
Inspired by recent work in machine translation and object detection, we introduce an attention based model that automatically learns to describe the content of images. We describe how we can train this model in a deterministic manner using standard backpropagation techniques and stochastically by maximizing a variational lower bound. We also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We validate the use of attention with state-of-the-art performance on three benchmark datasets: Flickr9k, Flickr30k and MS COCO.

1. Introduction

Automatically generating captions for an image is a task close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language. For this reason, caption generation has long been seen as a difficult problem. It amounts to mimicking the remarkable human ability to compress huge amounts of salient visual information into descriptive language and is thus an important challenge for machine learning and AI research.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Figure 1. Our model learns a words/image alignment. The visualized attentional maps (3) are explained in Sections 3.1 & 5.4



Yet despite the difficult nature of this task, there has been a recent surge of research interest in attacking the image caption generation problem. Aided by advances in training deep neural networks (Krizhevsky et al., 2012) and the availability of large classification datasets (Russakovsky et al., 2014), recent work has significantly improved the quality of caption generation using a combination of convolutional neural networks (convnets) to obtain vectorial representation of images and recurrent neural networks to decode those representations into natural language sentences (see Sec. 2). One of the most curious facets of the human visual system is the presence of attention (Rensink, 2000; Corbetta & Shulman, 2002). Rather than compress an entire image into a static representation, attention allows for salient features to dynamically come to the forefront as needed. This is especially important when there is a lot of clutter in an image. Using representations (such as those from the very top layer of a convnet) that distill information in image down to the most salient objects is one effective solution that has been widely adopted in previous work. Unfortunately, this has one potential drawback of losing information which could be useful for richer, more descrip-

tive captions. Using lower-level representation can help preserve this information. However working with these features necessitates a powerful mechanism to steer the model to information important to the task at hand, and we show how learning to attend at different locations in order to generate a caption can achieve that. We present two variants: a "hard" stochastic attention mechanism and a "soft" deterministic attention mechanism. We also show how one advantage of including attention is the insight gained by approximately visualizing what the model "sees". Encouraged by recent advances in caption generation and inspired by recent successes in employing attention in machine translation (Bahdanau et al., 2014) and object recognition (Ba et al., 2014; Mnih et al., 2014), we investigate models that can attend to salient part of an image while generating its caption.

The contributions of this paper are the following:

- We introduce two attention-based image caption generators under a common framework (Sec. 3.1): 1) a "soft" deterministic attention mechanism trainable by standard back-propagation methods and 2) a "hard" stochastic attention mechanism trainable by maximizing an approximate variational lower bound or equivalently by REINFORCE (Williams, 1992).
- We show how we can gain insight and interpret the results of this framework by visualizing "where" and "what" the attention focused on (see Sec. 5.4.)
- Finally, we quantitatively validate the usefulness of attention in caption generation with state-of-the-art performance (Sec. 5.3) on three benchmark datasets: Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014) and the MS COCO dataset (Lin et al., 2014).

2. Related Work

In this section we provide relevant background on previous work on image caption generation and attention. Recently, several methods have been proposed for generating image descriptions. Many of these methods are based on recurrent neural networks and inspired by the successful use of sequence-to-sequence training with neural networks for machine translation (Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; Kalchbrenner & Blunsom, 2013). The encoder-decoder framework (Cho et al., 2014) of machine translation is well suited, because it is analogous to "translating" an image to a sentence.

The first approach to using neural networks for caption generation was proposed by Kiros et al. (2014a) who used a multimodal log-bilinear model that was biased by features from the image. This work was later followed by Kiros et al. (2014b) whose method was designed to explicitly allow for a natural way of doing both ranking and generation. Mao et al. (2014) used a similar approach to generation but replaced a feedforward neural language model with a recurrent one. Both Vinyals et al. (2014) and Donahue et al. (2014) used recurrent neural networks (RNN) based on long short-term memory (LSTM) units (Hochreiter & Schmidhuber, 1997) for their models. Unlike Kiros et al. (2014a) and Mao et al. (2014) whose models see the image at each time step of the output word sequence, Vinyals et al. (2014) only showed the image to the RNN at the beginning. Along with images, Donahue et al. (2014) and Yao et al. (2015) also applied LSTMs to videos, allowing their model to generate video descriptions.

Most of these works represent images as a single feature vector from the top layer of a pre-trained convolutional network. Karpathy & Li (2014) instead proposed to learn a joint embedding space for ranking and generation whose model learns to score sentence and image similarity as a function of R-CNN object detections with outputs of a bidirectional RNN. Fang et al. (2014) proposed a three-step pipeline for generation by incorporating object detections. Their models first learn detectors for several visual concepts based on a multi-instance learning framework. A language model trained on captions was then applied to the detector outputs, followed by rescoring from a joint image-text embedding space. Unlike these models, our proposed attention framework does not explicitly use object detectors but instead learns latent alignments from scratch. This allows our model to go beyond "objectness" and learn to attend to abstract concepts.

Prior to the use of neural networks for generating captions, two main approaches were dominant. The first involved generating caption templates which were filled in based on the results of object detections and attribute discovery (Kulkarni et al. (2013), Li et al. (2011), Yang et al. (2011), Mitchell et al. (2012), Elliott & Keller (2013)). The second approach was based on first retrieving similar captioned images from a large database then modifying these retrieved captions to fit the query (Kuznetsova et al., 2012; 2014). These approaches typically involved an intermediate "generalization" step to remove the specifics of a caption that are only relevant to the retrieved image, such as the name of a city. Both of these approaches have since fallen out of favour to the now dominant neural network methods.

There has been a long line of previous work incorporating the idea of attention into neural networks. Some that share the same spirit as our work include Larochelle & Hinton (2010); Denil et al. (2012); Tang et al. (2014) and more recently Gregor et al. (2015). In particular however, our work directly extends the work of Bahdanau et al. (2014); Mnih et al. (2014); Ba et al. (2014); Graves (2013).

3. Image Caption Generation with Attention Mechanism

3.1. Model Details

In this section, we describe the two variants of our attention-based model by first describing their common framework. The key difference is the definition of the ϕ function which we describe in detail in Sec. 4. See Fig. 1 for the graphical illustration of the proposed model.

We denote vectors with bolded font and matrices with capital letters. In our description below, we suppress bias terms for readability.

3.1.1. Encoder: Convolutional Features

Our model takes a single raw image and generates a caption $\mathbf{y}$ encoded as a sequence of 1-of-K encoded words.

$$y = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\}, \ \mathbf{y}_i \in \mathbb{R}^K$$

where K is the size of the vocabulary and C is the length of the caption.

We use a convolutional neural network in order to extract a set of feature vectors which we refer to as annotation vectors. The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image.

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

In order to obtain a correspondence between the feature vectors and portions of the 2-D image, we extract features from a lower convolutional layer unlike previous work which instead used a fully connected layer. This allows the decoder to selectively focus on certain parts of an image by weighting a subset of all the feature vectors.

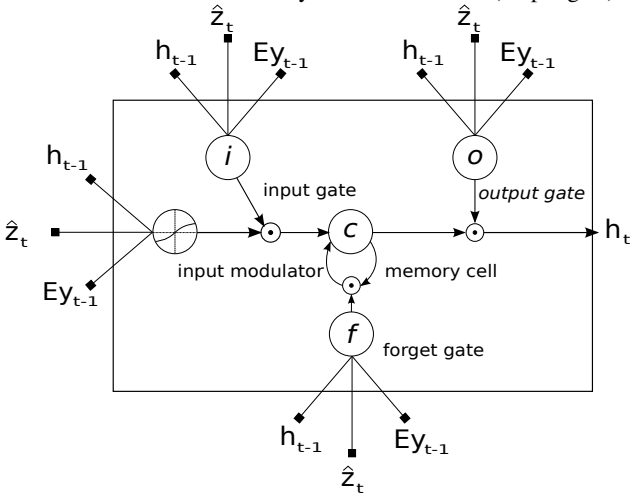
3.1.2. Decoder: Long Short-Term Memory Network

We use a long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997) that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words. Our implementation of LSTMs, shown in Fig. 2, closely follows the one used in Zaremba et al. (2014):

$$\begin{aligned}
\mathbf{i}_t &= \sigma(W_i E\mathbf{y}_{t-1} + U_i \mathbf{h}_{t-1} + Z_i \hat{\mathbf{z}}_t + \mathbf{b}_i),\\
\mathbf{f}_t &= \sigma(W_f E\mathbf{y}_{t-1} + U_f \mathbf{h}_{t-1} + Z_f \hat{\mathbf{z}}_t + \mathbf{b}_f),\\
\mathbf{c}_t &= \mathbf{f}_t \mathbf{c}_{t-1} + \mathbf{i}_t \tanh(W_c E\mathbf{y}_{t-1} + U_c \mathbf{h}_{t-1} + Z_c \hat{\mathbf{z}}_t + \mathbf{b}_c),\\
\mathbf{o}_t &= \sigma(W_o E\mathbf{y}_{t-1} + U_o \mathbf{h}_{t-1} + Z_o \hat{\mathbf{z}}_t + \mathbf{b}_o),\\
\mathbf{h}_t &= \mathbf{o}_t \tanh(\mathbf{c}_t).
\end{aligned}$$

Here, $\mathbf{i}_t$, $\mathbf{f}_t$, $\mathbf{c}_t$, $\mathbf{o}_t$, $\mathbf{h}_t$ are the input, forget, memory, output and hidden state of the LSTM respectively. $W_\bullet$, $U_\bullet$, $Z_\bullet$ and $\mathbf{b}_\bullet$ are learned weight matricies and biases. $\mathbf{E} \in \mathbb{R}^{m \times K}$ is an embedding matrix. Let m and n denote the embedding and LSTM dimensionality respectively and σ be the logistic sigmoid activation.

Figure 2. A LSTM cell, lines with bolded squares imply projections with a learnt weight vector. Each cell learns how to weigh its input components (input gate), while learning how to modulate that contribution to the memory (input modulator). It also learns weights which erase the memory cell (forget gate), and weights which control how this memory should be emitted (output gate).



In simple terms, the context vector $\hat{\mathbf{z}}_t$ is a dynamic representation of the relevant part of the image input at time t. We define a mechanism ϕ that computes $\hat{\mathbf{z}}_t$ from the annotation vectors $\mathbf{a}_i, i = 1, \ldots, L$ corresponding to the features extracted at different image locations. For each location i, the mechanism generates a positive weight α_i which can be interpreted either as the probability that location i is the right place to focus for producing the next word (stochastic attention mechanism), or as the relative importance to give to location i in blending the $\mathbf{a}_i$'s together (deterministic attention mechanism). The weight α_i of each annotation vector a_i is computed by an *attention model* f_{att} for which we use a multilayer perceptron conditioned on the previous hidden state $\mathbf{h}_{t-1}$. To emphasize, we note that the hidden state varies as the output RNN advances in its output sequence: "where" the network looks next depends on the sequence of words that has already been generated.

$$\begin{aligned}
e_{ti} &= f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})\\
\alpha_{ti} &= \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.
\end{aligned}$$

Once the weights (which sum to one) are computed, the context vector $\hat{z}_t$ is computed by

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right), \tag{1}$$

where ϕ is a function that returns a single vector given the

set of annotation vectors and their corresponding weights. The details of the ϕ function are discussed in Sec. 4.

The initial memory state and hidden state of the LSTM are predicted by an average of the annotation vectors fed through two separate MLPs (init,c and init,h):

$$\mathbf{c}_0 = f_{\text{init,c}}\left(\frac{1}{L}\sum_i^L \mathbf{a}_i\right), \quad \mathbf{h}_0 = f_{\text{init,h}}\left(\frac{1}{L}\sum_i^L \mathbf{a}_i\right)$$

In this work, we use a deep output layer (Pascanu et al., 2014) to compute the output word probability. Its input are cues from the image (the context vector), the previously generated word, and the decoder state (h_t).

$$p(\mathbf{y}_t|\mathbf{a}, \mathbf{y}_1^{t-1}) \propto \exp(\mathbf{L}_o(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{L}_h\mathbf{h}_t + \mathbf{L}_z\hat{\mathbf{z}}_t)), \quad (2)$$

where $\mathbf{L}_o \in \mathbb{R}^{K\times m}$, $\mathbf{L}_h \in \mathbb{R}^{m\times n}$, $\mathbf{L}_z \in \mathbb{R}^{m\times D}$, and $\mathbf{E}$ are learned parameters initialized randomly.

4. Learning Stochastic "Hard" vs Deterministic "Soft" Attention

In this section we discuss two alternative mechanisms for the attention model f_{att}: stochastic attention and deterministic attention.

4.1. Stochastic "Hard" Attention

We represent the location variable s_t as where the model decides to focus attention when generating the t-th word. $s_{t,i}$ is an indicator one-hot variable which is set to 1 if the i-th location (out of L) is the one used to extract visual features. By treating the attention locations as intermediate latent variables, we can assign a multinoulli distribution parametrized by $\{\alpha_i\}$, and view $\hat{\mathbf{z}}_t$ as a random variable:

$$p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) = \alpha_{t,i} \quad (3)$$

$$\hat{\mathbf{z}}_t = \sum_i s_{t,i}\mathbf{a}_i. \quad (4)$$

We define a new objective function L_s that is a variational lower bound on the marginal log-likelihood $\log p(\mathbf{y} \mid \mathbf{a})$ of observing the sequence of words $\mathbf{y}$ given image features $\mathbf{a}$. Similar to work in generative deep generative modeling (Kingma & Welling, 2014; Rezende et al., 2014), the learning algorithm for the parameters W of the models can be derived by directly optimizing

$$\begin{aligned} L_s &= \sum_s p(s \mid \mathbf{a}) \log p(\mathbf{y} \mid s, \mathbf{a}) \\ &\leq \log \sum_s p(s \mid \mathbf{a}) p(\mathbf{y} \mid s, \mathbf{a}) \\ &= \log p(\mathbf{y} \mid \mathbf{a}), \end{aligned} \quad (5)$$

following its gradient

$$\frac{\partial L_s}{\partial W} = \sum_s p(s \mid \mathbf{a}) \Bigg[\frac{\partial \log p(\mathbf{y} \mid s, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid s, \mathbf{a}) \frac{\partial \log p(s \mid \mathbf{a})}{\partial W}\Bigg]. \quad (6)$$

We approximate this gradient of L_s by a Monte Carlo method such that

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N}\sum_{n=1}^N \Bigg[\frac{\partial \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a}) \frac{\partial \log p(\tilde{s}^n \mid \mathbf{a})}{\partial W}\Bigg], \quad (7)$$

where $\tilde{s}^n = (s_1^n, s_2^n, \ldots)$ is a sequence of sampled attention locations. We sample the location s_t^n from a multinouilli distribution defined by Eq. (3):

$$\tilde{s_t^n} \sim \text{Multinoulli}_L(\{\alpha_i^n\}).$$

We reduce the variance of this estimator with the moving average baseline technique (Weaver & Tao, 2001). Upon seeing the k-th mini-batch, the moving average baseline is estimated as an accumulated sum of the previous log likelihoods with exponential decay:

$$b_k = 0.9 \times b_{k-1} + 0.1 \times \log p(\mathbf{y} \mid \tilde{s}_k, \mathbf{a})$$

To further reduce the estimator variance, the gradient of the entropy $H[s]$ of the multinouilli distribution is added to the RHS of Eq. (7).

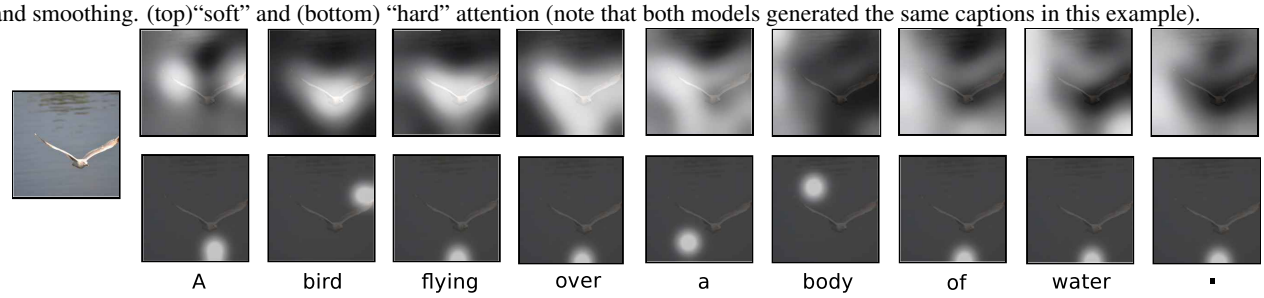
The final learning rule for the model is then

$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N}\sum_{n=1}^N \Bigg[\frac{\partial \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})}{\partial W} + \lambda_r(\log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a}) - b)\frac{\partial \log p(\tilde{s}^n \mid \mathbf{a})}{\partial W} + \lambda_e \frac{\partial H[\tilde{s}^n]}{\partial W}\Bigg]$$

where, λ_r and λ_e are two hyper-parameters set by cross-validation. As pointed out and used by Ba et al. (2014) and Mnih et al. (2014), this formulation is equivalent to the REINFORCE learning rule (Williams, 1992), where the reward for the attention choosing a sequence of actions is a real value proportional to the log likelihood of the target sentence under the sampled attention trajectory.

In order to further improve the robustness of this learning rule, with probability 0.5 for a given image, we set the sampled attention location $\tilde{s}$ to its expected value α (equivalent to the deterministic attention in Sec. 4.2).

Figure 3. Visualization of the attention for each generated word. The rough visualizations obtained by upsampling the attention weights and smoothing. (top)"soft" and (bottom) "hard" attention (note that both models generated the same captions in this example).



4.2. Deterministic "Soft" Attention

Learning stochastic attention requires sampling the attention location s_t each time, instead we can take the expectation of the context vector $\hat{\mathbf{z}}_t$ directly,

$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i \tag{8}$$

and formulate a deterministic attention model by computing a soft attention weighted annotation vector $\phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right) = \sum_i^L \alpha_i \mathbf{a}_i$ as proposed by Bahdanau et al. (2014). This corresponds to feeding in a soft α weighted context into the system. The whole model is smooth and differentiable under the deterministic attention, so learning end-to-end is trivial by using standard back-propagation.

Learning the deterministic attention can also be understood as approximately optimizing the marginal likelihood in Eq. (5) under the attention location random variable s_t from Sec. 4.1. The hidden activation of LSTM $\mathbf{h}_t$ is a linear projection of the stochastic context vector $\hat{\mathbf{z}}_t$ followed by $\tanh$ non-linearity. To the first-order Taylor approximation, the expected value $\mathbb{E}_{p(s_t|a)}[\mathbf{h}_t]$ is equivalent to computing $\mathbf{h}_t$ using a single forward computation with the expected context vector $\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t]$.

Let us denote by $\mathbf{n}_{t,i}$ as $\mathbf{n}$ in Eq. (2) with $\hat{\mathbf{z}}_t$ set to $\mathbf{a}_i$. Then, we can write the normalized weighted geometric mean (NWGM) of the softmax of k-th word prediction as

$$\begin{aligned}\text{NWGM}[p(y_t = k \mid \mathbf{a})] &= \frac{\prod_i \exp(n_{t,k,i})^{p(s_{t,i}=1|a)}}{\sum_j \prod_i \exp(n_{t,j,i})^{p(s_{t,i}=1|a)}} \\ &= \frac{\exp(\mathbb{E}_{p(s_t|a)}[n_{t,k}])}{\sum_j \exp(\mathbb{E}_{p(s_t|a)}[n_{t,j}])}\end{aligned}$$

This implies that the NWGM of the word prediction can be well approximated by using the expected context vector $\mathbb{E}\left[\hat{\mathbf{z}}_t\right]$, instead of the sampled context vector $\mathbf{a}_i$.

Furthermore, from the result by Baldi & Sadowski (2014), the NWGM in Eq. (9) which can be computed by a single feedforward computation approximates the expectation $\mathbb{E}[p(y_t = k \mid \mathbf{a})]$ of the output over all possible attention locations induced by random variable s_t. This suggests that the proposed deterministic attention model approximately maximizes the marginal likelihood over all possible attention locations.

4.2.1. Doubly Stochastic Attention

In training the deterministic version of our model, we introduce a form a doubly stochastic regularization that encourages the model to pay equal attention to every part of the image. Whereas the attention at every point in time sums to 1 by construction (i.e $\sum_i \alpha_{ti} = 1$), the attention $\sum_i \alpha_{ti}$ is not constrained in any way. This makes it possible for the decoder to ignore some parts of the input image. In order to alleviate this, we encourage $\sum_t \alpha_{ti} \approx \tau$ where $\tau \geq \frac{L}{D}$. In our experiments, we observed that this penalty quantitatively improves overall performance and that this qualitatively leads to more descriptive captions.

Additionally, the soft attention model predicts a gating scalar β from previous hidden state $\mathbf{h}_{t-1}$ at each time step t, such that, $\phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right) = \beta \sum_i^L \alpha_i \mathbf{a}_i$, where $\beta_t = \sigma(f_\beta(\mathbf{h}_{t-1}))$. This gating variable lets the decoder decide whether to put more emphasis on language modeling or on the context at each time step. Qualitatively, we observe that the gating variable is larger than the decoder describes an object in the image.

The soft attention model is trained end-to-end by minimizing the following penalized negative log-likelihood:

$$L_d = -\log(p(\mathbf{y}|\mathbf{a})) + \lambda \sum_i^L (1 - \sum_t^C \alpha_{ti})^2, \tag{9}$$

where we simply fixed τ to 1.

4.3. Training Procedure

Both variants of our attention model were trained with stochastic gradient descent using adaptive learning rates. For the Flickr8k dataset, we found that RMSProp (Tieleman & Hinton, 2012) worked best, while for Flickr30k/MS COCO dataset we for the recently proposed Adam algorithm (Kingma & Ba, 2014) to be quite effective.

To create the annotations a_i used by our decoder, we used

Table 1. BLEU-1,2,3,4/METEOR metrics compared to other methods, † indicates a different split, (—) indicates an unknown metric, ∘ indicates the authors kindly provided missing metrics by personal communication, Σ indicates an ensemble, a indicates using AlexNet

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

the Oxford VGGnet (Simonyan & Zisserman, 2014) pre-trained on ImageNet without finetuning. In our experiments we use the 14×14×512 feature map of the fourth convolutional layer before max pooling. This means our decoder operates on the flattened 196×512 (i.e $L \times D$) encoding. In principle however, any encoding function could be used. In addition, with enough data, the encoder could also be trained from scratch (or fine-tune) with the rest of the model.

As our implementation requires time proportional to the length of the longest sentence per update, we found training on a random group of captions to be computationally wasteful. To mitigate this problem, in preprocessing we build a dictionary mapping the length of a sentence to the corresponding subset of captions. Then, during training we randomly sample a length and retrieve a mini-batch of size 64 of that length. We found that this greatly improved convergence speed with no noticeable diminishment in performance. On our largest dataset (MS COCO), our soft attention model took less than 3 days to train on an NVIDIA Titan Black GPU.

In addition to dropout (Srivastava et al., 2014), the only other regularization strategy we used was early stopping on BLEU score. We observed a breakdown in correlation between the validation set log-likelihood and BLEU in the later stages of training during our experiments. Since BLEU is the most commonly reported metric, we used BLEU on our validation set for model selection.

In our experiments with soft attention, we used Whetlab[1] (Snoek et al., 2012; 2014) in our Flickr8k experiments. Some of the intuitions we gained from hyperparameter regions it explored were especially important in our Flickr30k and COCO experiments.

We make our code for these models publicly available to encourage future research in this area[2].

5. Experiments

We describe our experimental methodology and quantitative results which validate the effectiveness of our model for caption generation.

5.1. Data

We report results on the widely-used Flickr8k and Flickr30k dataset as well as the more recenly introduced MS COCO dataset. Each image in the Flickr8k/30k dataset have 5 reference captions. In preprocessing our COCO dataset, we maintained a the same number of references between our datasets by discarding caption in excess of 5. We applied only basic tokenization to MS COCO so that it is consistent with the tokenization present in Flickr8k and Flickr30k. For all our experiments, we used a fixed vocabulary size of 10,000.

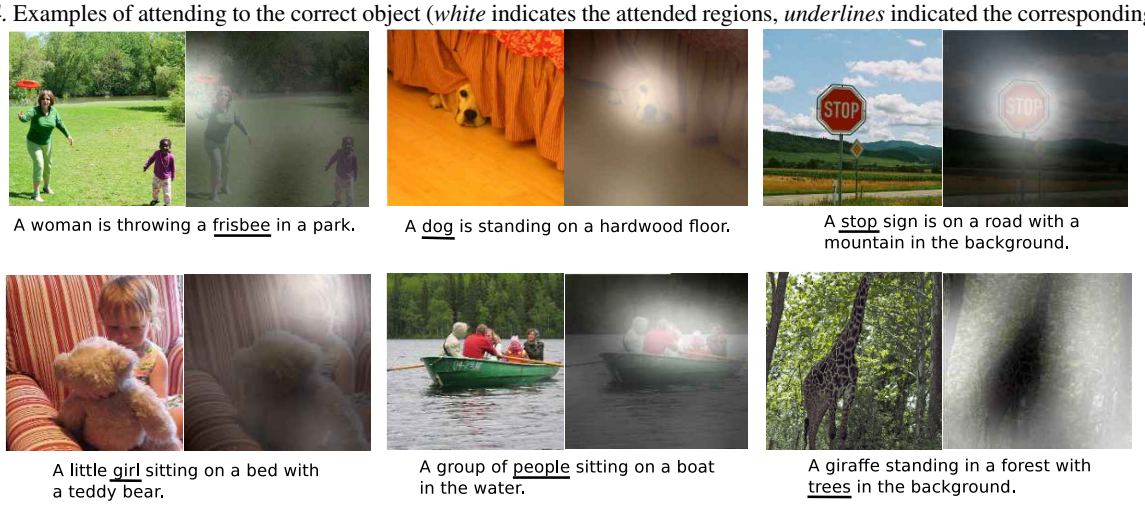
Results for our attention-based architecture are reported in Table 1. We report results with the frequently used BLEU metric[3] which is the standard in image caption generation

[1] https://www.whetlab.com/

[2] https://github.com/kelvinxu/arctic-captions

[3] We verified that our BLEU evaluation code matches the au-

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)



research. We report BLEU[4] from 1 to 4 without a brevity penalty. There has been, however, criticism of BLEU, so we report another common metric METEOR (Denkowski & Lavie, 2014) and compare whenever possible.

5.2. Evaluation Procedures

A few challenges exist for comparison, which we explain here. The first challenge is a difference in choice of convolutional feature extractor. For identical decoder architectures, using a more recent architectures such as GoogLeNet (Szegedy et al., 2014) or Oxford VGG (Simonyan & Zisserman, 2014) can give a boost in performance over using the AlexNet (Krizhevsky et al., 2012). In our evaluation, we compare directly only with results which use the comparable GoogLeNet/Oxford VGG features, but for METEOR comparison we include some results that use AlexNet.

The second challenge is a single model versus ensemble comparison. While other methods have reported performance boosts by using ensembling, in our results we report a single model performance.

Finally, there is a challenge due to differences between dataset splits. In our reported results, we use the predefined splits of Flickr8k. However, for the Flickr30k and COCO datasets is the lack of standardized splits for which results are reported. As a result, we report the results with the publicly available splits[5] used in previous work (Karpathy & Li, 2014). We note, however, that the differences in splits do not make a substantial difference in overall performance.

5.3. Quantitative Analysis

In Table 1, we provide a summary of the experiment validating the quantitative effectiveness of attention. We obtain state of the art performance on the Flickr8k, Flickr30k and MS COCO. In addition, we note that in our experiments we are able to significantly improve the state-of-the-art performance METEOR on MS COCO. We speculate that this is connected to some of the regularization techniques we used (see Sec. 4.2.1) and our lower-level representation.

5.4. Qualitative Analysis: Learning to attend

By visualizing the attention learned by the model, we are able to add an extra layer of interpretability to the output of the model (see Fig. 1). Other systems that have done this rely on object detection systems to produce candidate alignment targets (Karpathy & Li, 2014). Our approach is much more flexible, since the model can attend to "non-object" salient regions.

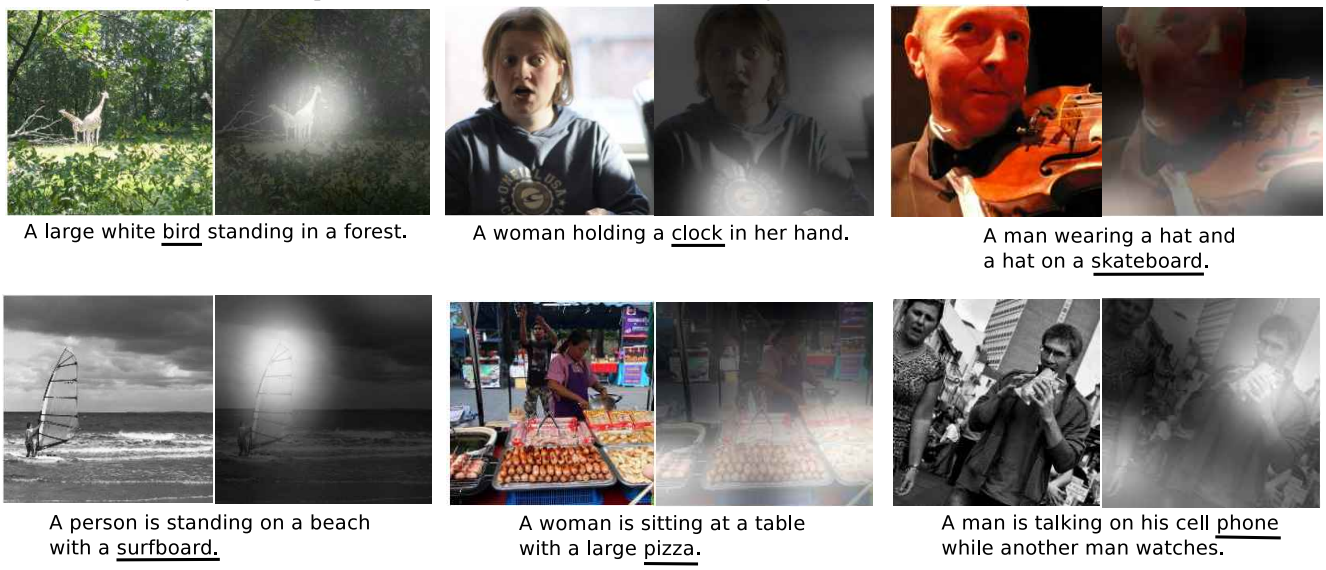
The 19-layer OxfordNet uses stacks of 3x3 filters meaning the only time the feature maps decrease in size are due to the max pooling layers. The input image is resized so that the shortest side is 256-dimensional with preserved aspect ratio. The input to the convolutional network is the center-cropped 224x224 image. Consequently, with four max pooling layers, we get an output dimension of the top convolutional layer of 14x14. Thus in order to visualize the attention weights for the soft model, we upsample the weights by a factor of $2^4 = 16$ and apply a Gaussian filter

thors of Vinyals et al. (2014), Karpathy & Li (2014) and Kiros et al. (2014b). For fairness, we only compare against results for which we have verified that our BLEU evaluation code is the same.

[4] BLEU-n is the geometric average of the n-gram precision. For instance, BLEU-1 is the unigram precision, and BLEU-2 is the geometric average of the unigram and bigram precision.

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

Figure 5. Examples of mistakes where we can use attention to gain intuition into what the model saw.



to emulate the large receptive field size.

As we can see in Figs. 3 and 4, the model learns alignments that agree very strongly with human intuition. Especially from the examples of mistakes in Fig. 5, we see that it is possible to exploit such visualizations to get an intuition as to why those mistakes were made. We provide a more extensive list of visualizations as the supplementary materials for the reader.

6. Conclusion

We propose an attention based approach that gives state of the art performance on three benchmark datasets using the BLEU and METEOR metric. We also show how the learned attention can be exploited to give more interpretability into the models generation process, and demonstrate that the learned alignments correspond very well to human intuition. We hope that the results of this paper will encourage future work in using visual attention. We also expect that the modularity of the encoder-decoder approach combined with attention to have useful applications in other domains.

Acknowledgments

The authors would like to thank the developers of Theano (Bergstra et al., 2010; Bastien et al., 2012). We acknowledge the support of the following organizations for research funding and computing support: NSERC, Samsung, NVIDIA, Calcul Québec, Compute Canada, the Canada Research Chairs and CIFAR. The authors would also like to thank Nitish Srivastava for assistance with his ConvNet package as well as preparing the Oxford convolutional network and Relu Patrascu for helping with numerous infrastructure-related problems.

References

Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. Multiple object recognition with visual attention. *arXiv:*1412.7755 [cs.LG], December 2014.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv:*1409.0473 [cs.CL], September 2014.

Baldi, Pierre and Sadowski, Peter. The dropout learning algorithm. *Artificial intelligence*, 210:78–122, 2014.

Bastien, Frederic, Lamblin, Pascal, Pascanu, Razvan, Bergstra, James, Goodfellow, Ian, Bergeron, Arnaud, Bouchard, Nicolas, Warde-Farley, David, and Bengio, Yoshua. Theano: new features and speed improvements. Submited to the Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

Bergstra, James, Breuleux, Olivier, Bastien, Frédéric, Lamblin, Pascal, Pascanu, Razvan, Desjardins, Guillaume, Turian, Joseph, Warde-Farley, David, and Bengio, Yoshua. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, 2010.

Chen, Xinlei and Zitnick, C Lawrence. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*, 2014.

Cho, Kyunghyun, van Merrienboer, Bart, Gulcehre, Caglar, Bougares, Fethi, Schwenk, Holger, and Bengio, Yoshua. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, October 2014.

Corbetta, Maurizio and Shulman, Gordon L. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002.

Denil, Misha, Bazzani, Loris, Larochelle, Hugo, and de Freitas, Nando. Learning where to attend with deep architectures for image tracking. *Neural Computation*, 2012.

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

Donahue, Jeff, Hendrikcs, Lisa Anne, Guadarrama, Segio, Rohrbach, Marcus, Venugopalan, Subhashini, Saenko, Kate, and Darrell, Trevor. Long-term recurrent convolutional networks for visual recognition and description. *arXiv:*1411.4389v2 [cs.CV], November 2014.

Elliott, Desmond and Keller, Frank. Image description using visual dependency representations. In *EMNLP*, pp. 1292–1302, 2013.

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

Graves, Alex. Generating sequences with recurrent neural networks. Technical report, arXiv preprint arXiv:1308.0850, 2013.

Gregor, Karol, Danihelka, Ivo, Graves, Alex, and Wierstra, Daan. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

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

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

Kalchbrenner, Nal and Blunsom, Phil. Recurrent continuous translation models. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1700–1709. Association for Computational Linguistics, 2013.

Karpathy, Andrej and Li, Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv:*1412.2306 [cs.CV], December 2014.

Kingma, Diederik P. and Ba, Jimmy. Adam: A Method for Stochastic Optimization. *arXiv:*1412.6980 [cs.LG], December 2014.

Kingma, Durk P. and Welling, Max. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

Kiros, Ryan, Salahutdinov, Ruslan, and Zemel, Richard. Multimodal neural language models. In *International Conference on Machine Learning*, pp. 595–603, 2014a.

Kiros, Ryan, Salakhutdinov, Ruslan, and Zemel, Richard. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv:*1411.2539 [cs.LG], November 2014b.

Krizhevsky, Alex, Sutskever, Ilya, and Hinton, Geoffrey. ImageNet classification with deep convolutional neural networks. In *NIPS*. 2012.

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

Kuznetsova, Polina, Ordonez, Vicente, Berg, Alexander C, Berg, Tamara L, and Choi, Yejin. Collective generation of natural image descriptions. In *Association for Computational Linguistics: Long Papers*, pp. 359–368. Association for Computational Linguistics, 2012.

Kuznetsova, Polina, Ordonez, Vicente, Berg, Tamara L, and Choi, Yejin. Treetalk: Composition and compression of trees for image descriptions. *TACL*, 2(10):351–362, 2014.

Larochelle, Hugo and Hinton, Geoffrey E. Learning to combine foveal glimpses with a third-order boltzmann machine. In *NIPS*, pp. 1243–1251, 2010.

Li, Siming, Kulkarni, Girish, Berg, Tamara L, Berg, Alexander C, and Choi, Yejin. Composing simple image descriptions using web-scale n-grams. In *Computational Natural Language Learning*, pp. 220–228. Association for Computational Linguistics, 2011.

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

Mao, Junhua, Xu, Wei, Yang, Yi, Wang, Jiang, and Yuille, Alan. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv:*1412.6632 *[cs.CV]*, December 2014.

Mitchell, Margaret, Han, Xufeng, Dodge, Jesse, Mensch, Alyssa, Goyal, Amit, Berg, Alex, Yamaguchi, Kota, Berg, Tamara, Stratos, Karl, and Daumé III, Hal. Midge: Generating image descriptions from computer vision detections. In *European Chapter of the Association for Computational Linguistics*, pp. 747–756. Association for Computational Linguistics, 2012.

Mnih, Volodymyr, Hees, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. Recurrent models of visual attention. In *NIPS*, 2014.

Pascanu, Razvan, Gulcehre, Caglar, Cho, Kyunghyun, and Bengio, Yoshua. How to construct deep recurrent neural networks. In *ICLR*, 2014.

Rensink, Ronald A. The dynamic representation of scenes. *Visual cognition*, 7(1-3):17–42, 2000.

Rezende, Danilo J., Mohamed, Shakir, and Wierstra, Daan. Stochastic backpropagation and approximate inference in deep generative models. Technical report, arXiv:1401.4082, 2014.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., and Fei-Fei, Li. ImageNet Large Scale Visual Recognition Challenge, 2014.

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

Snoek, Jasper, Larochelle, Hugo, and Adams, Ryan P. Practical bayesian optimization of machine learning algorithms. In *NIPS*, pp. 2951–2959, 2012.

Snoek, Jasper, Swersky, Kevin, Zemel, Richard S, and Adams, Ryan P. Input warping for bayesian optimization of non-stationary functions. *arXiv preprint arXiv:1402.0929*, 2014.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15:1929–1958, 2014.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc VV. Sequence to sequence learning with neural networks. In *NIPS*, pp. 3104–3112, 2014.

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

Tang, Yichuan, Srivastava, Nitish, and Salakhutdinov, Ruslan R. Learning generative models with visual attention. In *NIPS*, pp. 1808–1816, 2014.

Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5 - RMSProp. Technical report, 2012.

Vinyals, Oriol, Toshev, Alexander, Bengio, Samy, and Erhan, Dumitru. Show and tell: A neural image caption generator. *arXiv:*1411.4555 [cs.CV], November 2014.

Weaver, Lex and Tao, Nigel. The optimal reward baseline for gradient-based reinforcement learning. In *Proc. UAI'2001*, pp. 538–545, 2001.

Williams, Ronald J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Yang, Yezhou, Teo, Ching Lik, Daumé III, Hal, and Aloimonos, Yiannis. Corpus-guided sentence generation of natural images. In *EMNLP*, pp. 444–454. Association for Computational Linguistics, 2011.

Yao, Li, Torabi, Atousa, Cho, Kyunghyun, Ballas, Nicolas, Pal, Christopher, Larochelle, Hugo, and Courville, Aaron. Describing videos by exploiting temporal structure. *arXiv preprint arXiv:1502.08029*, April 2015.

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

Zaremba, Wojciech, Sutskever, Ilya, and Vinyals, Oriol. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, September 2014.