Exploring Show, Attend, Tell Attention Mechanisms for Image Captioning

Nick Brenner, Alex Kozik, Brandon Li, Jake Silver, Srivatsa Kundurthy

github.com/srkvatsa/exploring-show-attend-tell

1 Introduction

Image captioning straddles the intersection of Natural Language Processing (NLP) and Computer Vision (CV). Generating a caption given an image, especially in zero-shot cases, is a difficult task due to the need for semantic alignment between two very different spaces and the inherent subjectivity in what is of importance in an image.

(a) A woman pushes a fruit cart down a narrow street.

(b) An urban walkway lined with motorcycles and bicycles.

Figure 1: Two valid interpretations of the same image.

Show and Tell [12] was the first model architecture adapt the encoder-decoder framework for image captioning, using a Convolutional Neural Network (CNN) as the encoder to produce image embeddings, and then a LSTM to sequentially decode a caption generation. Building on this, **Show Attend Tell** [13], developed by Yoshua Bengio's lab in 2015, introduced an attention mechanism that allows the decoder to dynamically focus on different image regions during caption generation. By computing a context vector z_t at each timestep and feeding it into the LSTM, the model helps overcome the information bottleneck inherent to traditional RNNs. In this work, we reimplement **Show Attend Tell** to evaluate its effectiveness and better understand the role of visual attention in image captioning.

2 Chosen Result

We reproduce the METEOR scores for both soft and hard attention models on Flickr8k from the original paper [13] (Table 1). By capturing both attention variants, we evaluate the core contributions of the paper and contrast their approaches. We focus on METEOR because it has been shown to provide a more semantically faithful evaluation than BLEU, accounting for stemming, synonymy, and paraphrasing [1, 5]. Given our compute constraints, Flickr8k [3] was the optimal dataset for our experiments. Our goal is to assess the robustness and reproducibility of these techniques under modern training & architectural regimes.

Model	B-1	B-2	B-3	B-4	METEOR
Soft-Attention	67	44.8	29.9	19.5	18.93
Hard-Attention	67	45.7	31.4	21.3	20.30

Table 1: Performance comparison of attention mechanisms on Flickr8k in the original paper for BLEU1-4 & METEOR. [13]

3 Methodology

3.1 Encoder-Decoder Model

The paper uses an encoder-decoder framework with a convolutional neural network (CNN) to encode the input image and a long short-term memory (LSTM) network to decode the resulting features into a caption. The authors use VGGNet-19 [10] as the encoder, which produces a $14\times14\times512$ feature map from the input image. We use the more modern RESNET-50 [6], pretrained on ImageNet [9] with the final classification layers removed and only the convolutional layers up to block C5 kept, to obtain the latent $14 \times 14 \times 512$ space. This yields $\{a_1, \ldots, a_{196}\} \subseteq \mathbb{R}^{512}$ vectors which we call the annotations of the image. Each of these represent localized visual information and are used to construct the contextual vector $\hat{z}_t$ at each timestep.

At each decoding step, the LSTM gates are computed from a learned transformation of the previous word embedding, the previous hidden state, and most notably, the context vector from the attention mechanism. We use learned initial states using the multi-layer-perceptrons (MLPs), $\mathbf{c}_0 = f_{init,c}\left(\frac{1}{L}\sum_i^L \mathbf{a}_i\right)$, $\mathbf{h}_0 = f_{init,h}\left(\frac{1}{L}\sum_i^L \mathbf{a}_i\right)$.

A deep output layer [6] is used to compute the output word probability given the LSTM state, the context vector, and the previous word at each timestep t.

3.2 Attention

Attention enables the decoder to dynamically focus on different parts of the image when generating each token in the caption. At time step t, energy scores are learned via an MLP, $e_{ti} = f_{att}(\mathbf{a}_i, \mathbf{h}_{t-1})$. We then use softmax to obtain a distribution $\{\alpha_{ti}\}_{i=1}^{196}$ over our compressed $14 \times 14 \times 512$ encoding space, where α_{ti} is the associated probability with annotation a_i at timestep t. In particular, we have the distribution $p(s_t \mid a)$ where s_t denotes the attention location. We then have two approaches for computing the context vector $\hat{z}_t$.

3.2.1 Soft Attention

Soft attention is a deterministic attention mechanism where the context vector $\hat{\mathbf{z}}_t$ is computed as a weighted sum over all annotation vectors $\mathbf{a}_i$. It is the expectation over the distribution $p(s_t \mid a)$, so the context is $z_t = \sum_{i=1}^L \alpha_{ti}\mathbf{a}_i$. This allows all localized visual features to influence the context with learned weights. Note that all operations are differentiable; therefore, soft attention can be trained via backpropagation.

3.2.2 Doubly Stochastic Attention

As an extension to Soft Attention, Doubly Stochastic Attention adds a regularizer to the training objective: $L_d = -\log(P(\mathbf{y} \mid \mathbf{x})) + \lambda \sum_i^L (1 - \sum_i^C \alpha_{ti})^2$, where λ is a tuneable parameter (we use $\lambda = 0.2$). Over each of the $L = 196$ locations, we regularize so that $\sum_i^C \alpha_{ti} \approx 1$. In other words, for each location i, we regularize so that this particular location is attended to over all timesteps t, so one location is not only seen at a singular timestep. The authors note this improves performance [13].

3.2.3 Hard Attention

Hard attention is a stochastic attention mechanism in which, at each time step t, the model samples a **single** annotation vector to focus on, denoted $s_{t,i}$. So, $\hat{z}_t = \sum s_{t,i}\mathbf{a}_i$ (and only one such $s_{t,i}$ is nonzero).

3.3 Data

We trained and evaluated our model on the Flickr8k dataset [3]. This dataset consists of 8,000 images, each annotated with five human-written captions. We used the standard train/validation/test split provided by the dataset. Each test caption is evaluated against the five human reference captions using primarily the METEOR metric, though the BLEU-1 score is included for comparison [1, 5].

3.4 Training

Soft and hard attention require different training regimes due to the stochasticity in hard attention. We optimize soft attention using binary cross-entropy loss and standard backpropagation. Hard attention is not differentiable (due to stochasticity), and the paper did not specify their training algorithm used. We opted to use the REINFORCE algorithm with an exponential moving average baseline estimate and entropy term by setting the reward to be the negative per-sample cross-entropy loss. In particular, we roll out a trajectory $\tilde{s}^n$ of sampled locations from the distribution $p(s_t \mid \mathbf{a})$, keeping track of log probabilities and rewards, and compute the policy gradient estimate with Montecarlo sampling.

Because running the decoder requires time proportional to the longest length of a caption, we also added length-based sampling, where we sample a length and obtain a mini-batch of size 64 of training examples. This improved convergence speeds by allowing smaller length training examples to not be bottlenecked.

We faced characteristic challenges of REINFORCE in the hard attention regime: unstable gradients, high variance, and sensitivity to baseline tuning. Without careful calibration, the model sometimes collapsed to generic or repetitive outputs, a phenomenon echoed in the original paper's discussion. To mitigate this, our heuristics included several stabilization techniques, such as an exponential moving average baseline, entropy regularization, and label smoothing, as summarized in Table 2.

Soft Attention	**Details**
Optimizer	Adam (replace RMSProp) [4]
Encoder backbone	ResNet50 (replace VGG19) [2]
Teacher forcing	Linear decay from $1.0 \to 0.5$
Hard Attention	**Details**
Fine-tuning	20 epochs after soft attention
EMA baseline	$\alpha_{t-1} + (1-\alpha)b_t$, $\alpha = 0.95$
Entropy bonus	$\lambda_e \frac{\partial H[\tilde{s}^n]}{\partial W}$, $\lambda_e = 0.5$
CE smoothing	Label smoothing w/ $\epsilon = 0.1$ [11]
Length normalization	$\text{lp}(Y) = \left(\frac{5+\lvert Y\rvert}{6}\right)^{0.5}$
Temperature annealing	Softmax temperature $1.0 \to 0.5$ linearly
Teacher forcing	Same as original
Dropout	Reduced to 0.2
Beam search	Beam width increased to 8

Table 2: Training and inference modifications for soft and hard attention regimes.

We trained our final soft-attention model for 50 epochs, and found it was relatively stable to train. Because hard attention was optimized with REINFORCE, we anticipated weak training signals and instabilities during optimization. As such, we decided to implement hard attention by fine-tuning our soft attention model with the REINFORCE objective for an additional 20 epochs.

4 Results & Analysis

We report our METEOR & BLEU-1 scores on the Flickr8k test set for both soft and hard attention mechanisms, and compare them to paper's results, in Table 3.

Metric	**Xu et al.**	**Reproduced**
Soft Attention		
BLEU-1	**67.0**	45.9
METEOR	18.93	**18.96**
Hard Attention		
BLEU-1	**67.0**	43.2
METEOR	20.30	**20.75**

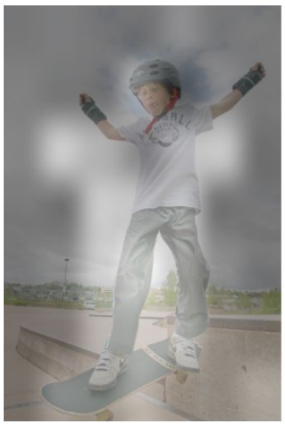
Table 3: Flickr8k BLEU-1 and METEOR scores for soft and hard attention models.

Notably, despite not having access to the original paper's hyperparameters and optimization details, our models achieved **superior** METEOR scores to those reported on the Flickr8k set, with a +0.03 improvement for soft attention and a +0.45 improvement for hard attention. Overall, our results affirm the utility of visual attention in image captioning. Our models not only matched but exceeded reported METEOR performance, illustrating the **robustness and replicability** of the original approach.

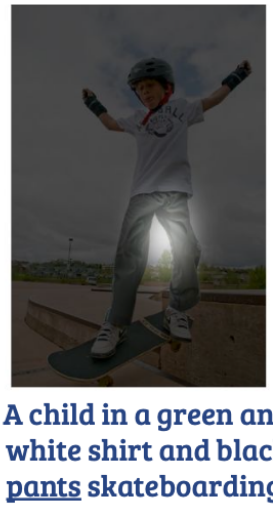
While reproducing BLEU-1 scores was not in our scope, we include them for completeness. BLEU-1 scores are lower across both reimplementations, and this is likely due to differences in tokenization. Notoriously,BLEU is sensitive to exact n-gram overlaps, making consistent tokenization a major factor in being able to fairly compare BLEU scores [5]. However,

because the caption tokenization scheme was not published in the original paper, BLEU scores suffer, which further validates our successful above-paper reproduction of the more expressive METEOR metric [1, 5]. Because the paper omitted several training details, including hyperparameter settings, number of epochs, and data pre-processing information, we conducted our own experiments and made heuristic decisions. Moreover, our training was conducted on a single GPU with runtime constraints, which limited batch size and the number of epochs.

An important implication of this paper is the **interpretability** of results, not only for correct but also incorrect examples. For example, we can understand "where" the model is looking when generating word t using both attention variants. By using our encoded annotation vectors, we can visualize attention by generating a heat map using the outputted distributions:



(a) Soft Attention when generating "boy" distributes the attention about the boy's figure and ignores the background.



(b) Hard Attention when generating "pants" sharply attends to the boy's pants.

Figure 2: Visual comparison of soft vs hard attention mechanisms.

5 Reflections

Implementing both attention mechanisms allowed us to empirally visualize the performance boost of techniques that combat bottlenecking in RNNs. While training the soft attention mechanism was relatively simple, we saw how difficult is is to train hard attention with REINFORCE due to the variability and intense sensitivity to hyperparameter configurations.

Future extensions of our work would include evaluating transformer-based decoding approaches, since these architectures have since dominated state-of-the-art results in VLM tasks and show promising zero-shot capabilities for captioning. Current research directions are improving the shortcomings of transformer models for understanding visual features [14]. Moreover, retrieval-augmented generation (RAG) approaches have emerged to supplement captioning abilities [7, 8], such as a a trigger-augmented (TA) generation approach to enhance visual alignment [15].

Our code, models, and results are accessible on GitHub. Since our approach outperformed the paper results, we hope that by making our methods open-source we are able to stimulate more transparent reproductions and evaluations of the original work.

References

[1] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[3] Micah Hodosh, Peter Young, and Julia Hockenmaier. Flickr8k dataset. University of Illinois at Urbana-Champaign, 2013.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[6] Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. How to construct deep recurrent neural networks. *Proceedings of the Second International Conference on Learning Representations*, 2014.

[7] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented transformer for image captioning. In *Proceedings of the 2022 ACM International Conference on Multimedia Retrieval*, pages 123–131. ACM, 2022.

[8] Rita Ramos, Desmond Elliott, and Bruno Martins. Retrieval-augmented image captioning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3666–3681. Association for Computational Linguistics, 2023.

[9] 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. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

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

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

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

[13] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *International Conference on Machine Learning*, pages 2048–2057, 2015.

[14] Shizhuo Dylan Zhang, Curt Tigges, Zory Zhang, Stella Biderman, Maxim Raginsky, and Talia Ringer. Transformer-based models are not yet perfect at learning to emulate structural recursion. *Transactions on Machine Learning Research*, 2024.

[15] Wei Zhang and Jing Zhang. Tpcap: Unlocking zero-shot image captioning with trigger-augmented and multi-modal purification modules. *arXiv preprint arXiv:2502.11024*, 2025.