Show, Attend and Tell

Neural Image Caption Generation with Visual Attention

Sunny Sun, Michael Wei, Tony Chen, Linda Hu, Jiye Baek

Cornell University

Background Motivation

- **Problem:** Prior captioning approaches used static global image features, often missing fine-grained spatial details crucial for accurate captioning
- **Objective:** Automatically generate natural language captions that describe the content of an input image
- ***Show, Attend and Tell*** introduced soft and hard visual attention mechanisms to dynamically focus on relevant regions during caption generation
- **Motivation:** Inspired by the rise of attention mechanisms in deep learning for structured prediction tasks like image captioning
- **Our Goal:** Replicate attention-based captioning model from *Show, Attend and Tell* using PyTorch, and evaluating performance on the Flickr8k dataset
- **Target Result:** Reproduce dynamic attention behavior and achieve BLEU/METEOR scores comparable to those reported in the original paper

Methodology

- **Dataset:** Flickr8k (8,000 images), 5 human annotated captions per image
- **Model Architecture (Figure 1):**
  - **Encoder:** Pretrained CNN extracts spatial feature vectors from input image
  - **Attention Mechanism:** At each decoding step t, compute attention energies $e_{i,t} = f_{att}(a_i, h_{t-1})$, use softmax to produce attention weights to generate context vector $\hat{z}_t = \phi(\{a_i\}, \{\alpha_i\})$
  - **Decoder:** Takes the context vector $\hat{z}_t$, the previous hidden state, and the previously generated word to predict the next word in the caption sequence
- **Modifications:**
  - Only implemented soft attention (no stochastic hard attention)

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

  - Used a pretrained CNN (no training from scratch) to save computational resources

System Overview

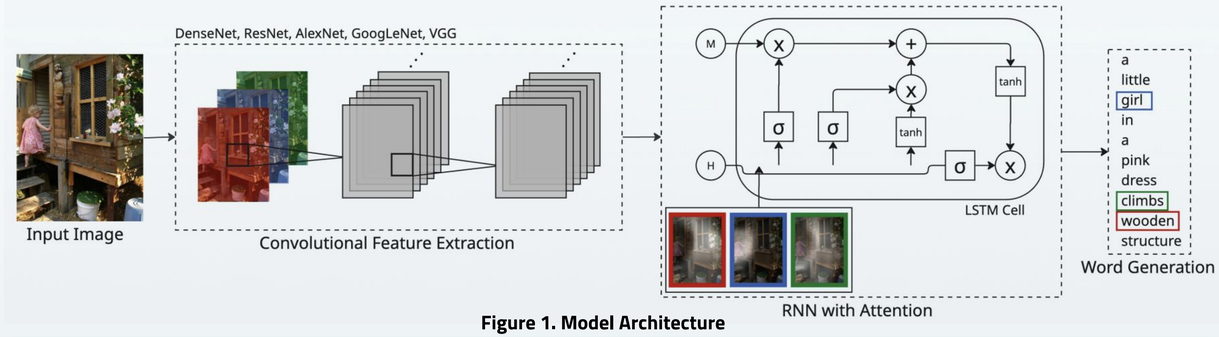

Figure 1. Model Architecture

Results

Figure 2. Visualization of Attention Mechanisms during Caption Generation

- Using newer pretrained-models (DenseNet, VGG, ResNet, GoogLeNet) boosted performance compared to AlexNet
- ResNet achieved the best overall performance, achieved results comparable to those reported for soft-attention model in original paper
- Validation accuracy plateaued after epoch 5 for pretrained models, suggesting early convergence

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Soft-Attention (Original Paper)	67.0	44.8	29.9	19.5	18.93
AlexNet	60.44	34.90	19.71	10.81	12.89
DenseNet	64.81	40.82	26.26	15.84	16.95
GoogLeNet	64.53	40.13	25.12	14.77	16.45
VGG	64.79	40.37	25.88	15.48	17.08
ResNet	**65.67**	**41.40**	**27.13**	**16.48**	**17.94**

Table 1: Comparison of BLEU and METEOR Scores across different CNN Encoders

Conclusion

- Replicated improvement over AlexNet by using new models
- Achieved marginally lower BLEU/METEOR scores (Table 1) compared to those in original paper, with ResNet outperforming all other pretrained models
- Visualized areas of image that model attends to per word similar to those in original paper (Figure 1)

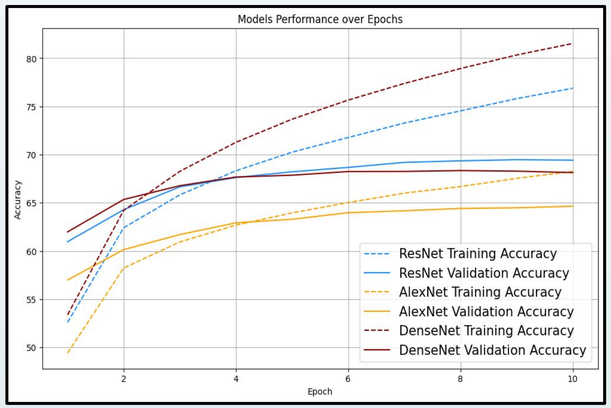

Figure 3. Accuracy of Models over Epochs

Future Work

- Use larger datasets like Flickr30k or MS COCO to improve generalization and caption quality
- Implement hard attention mechanism using REINFORCE to explore sharper spatial focus
- Fine-tune CNN encoder during training to allow the model to learn image features more tailored to caption generation

References

- Kelvin Xu et al. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2015. Paper Website. https://kelvinxu.github.io/projects/capgen.html
- Kelvin Xu. *Arctic Captions*. GitHub repository. https://github.com/kelvinxu/arctic-captions.
- Aaron C.C. Wong. *Show, Attend and Tell: PyTorch Implementation*. GitHub repository. https://github.com/AaronCCWong/Show-Attend-and-Tell.