Exploring *Show, Attend, Tell* Attention Mechanisms for Image Captioning

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

Cornell Bowers C·IS College of Computing and Information Science

Introduction

[Figure 3]

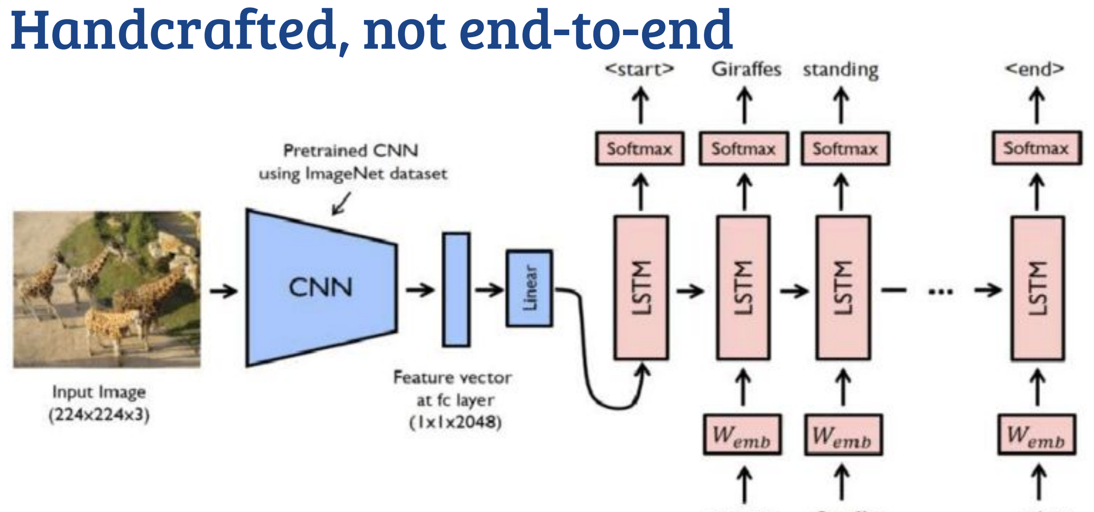
Hundreds of determined runners push forward at the start of a vibrant city marathon, each athlete focused on the road ahead under a bright, sunny sky.

- Image captioning is **complex** – requires understanding objects/relationships in images
- Reproducing the *Show, Attend, Tell* captioning model, training on Flickr-8k & optimizing for METEOR

Background & Motivation

Prior Approaches

- CNN → Single Vector → RNN (Vinyals et al., 2014)
  - ❌ Loses spatial **and** contextual detail
- CRF + Object Detectors (Fang et al., 2014)
  - ❌ Handcrafted, not end-to-end

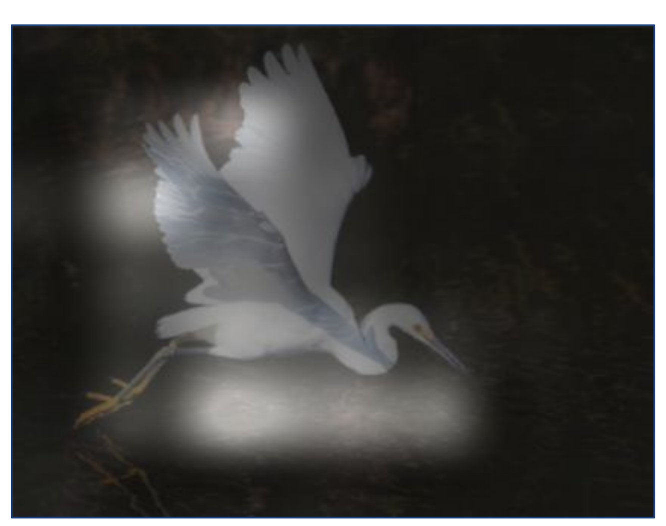


[Figure 4]

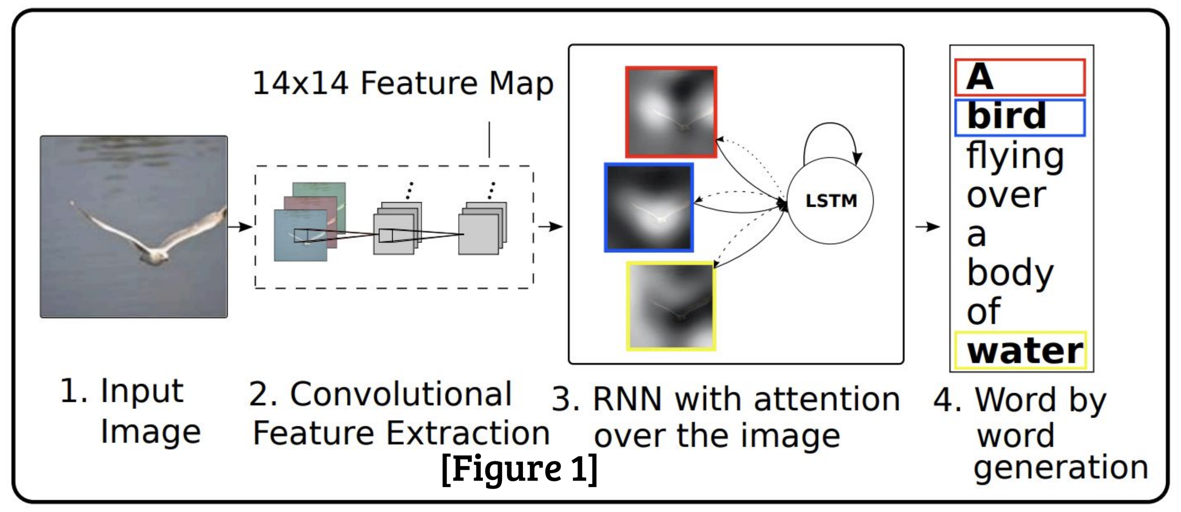
Encoder-Decoder Architecture for Captioning

Solution: Attention-Based Models

- Dynamically **focuses** on relevant image regions during captioning
- Mimics human visual attention to salient features; **interpretable**

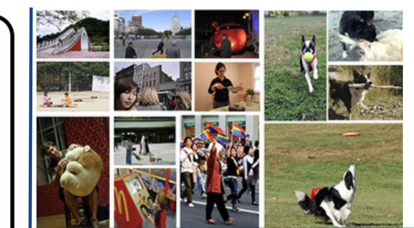


Methods



[Figure 1]

ResNet CNN, LSTM with Attention



[Figure 2]

Flickr8k Dataset

Attention

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}.$$

Soft Attention

[Figure 5]

Weighted sum over **image regions** at each timestep; fully differentiable

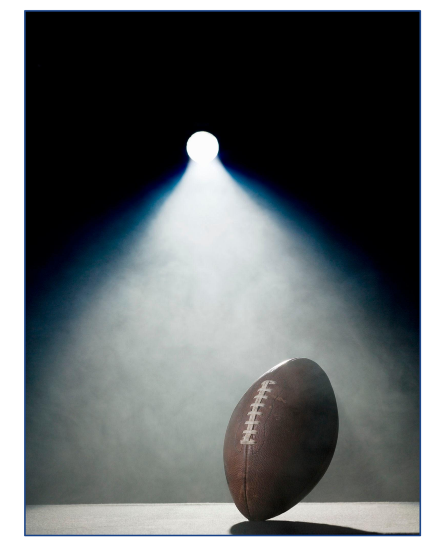
$$\mathbb{E}_{p(s_t|a)}[\hat{\mathbf{z}}_t] = \sum_{i=1}^{L} \alpha_{t,i}\mathbf{a}_i$$

Doubly Stochastic Attention

$$L_d = -\log(P(\mathbf{y}|\mathbf{x})) + \lambda \sum_{i}^{L}(1 - \sum_{t}^{C} \alpha_{ti})^2$$

Hard Attention

Selects **one** region to focus on per timestep; non-differentiable and trained with REINFORCE



[Figure 6]

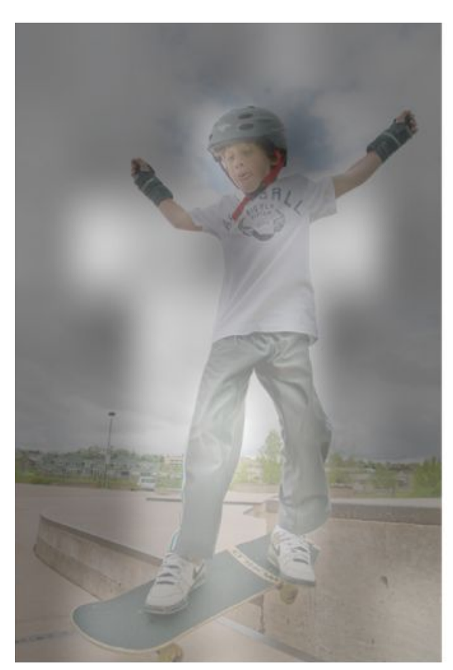
$$p(s_{t,i} = 1 \mid s_{j<t}, \mathbf{a}) = \alpha_{t,i}$$

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

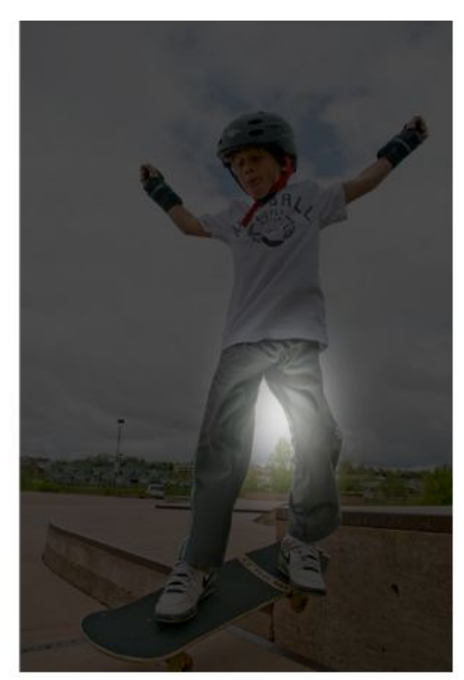
$$\frac{\partial L_s}{\partial W} \approx \frac{1}{N}\sum_{n=1}^{N}\Big[\frac{\partial \log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a})}{\partial W} + \lambda_r(\log p(\mathbf{y} \mid \tilde{s}^n, \mathbf{a}) - b)\frac{\partial \log p(\tilde{s}^n \mid \mathbf{a})}{\partial W} + \lambda_e \frac{\partial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

Moving **baseline** and **entropy** term for estimator variance reduction

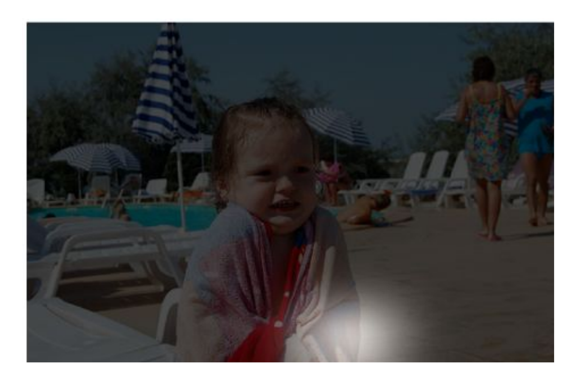
Where is the Model "Looking?"



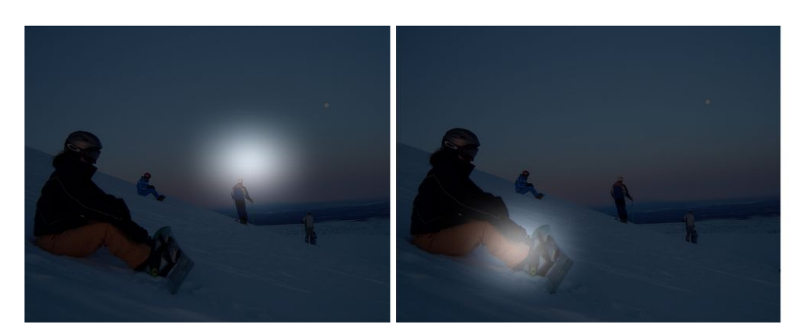
A boy does a skateboard trick.



A child in a green and white shirt and black pants skateboarding.



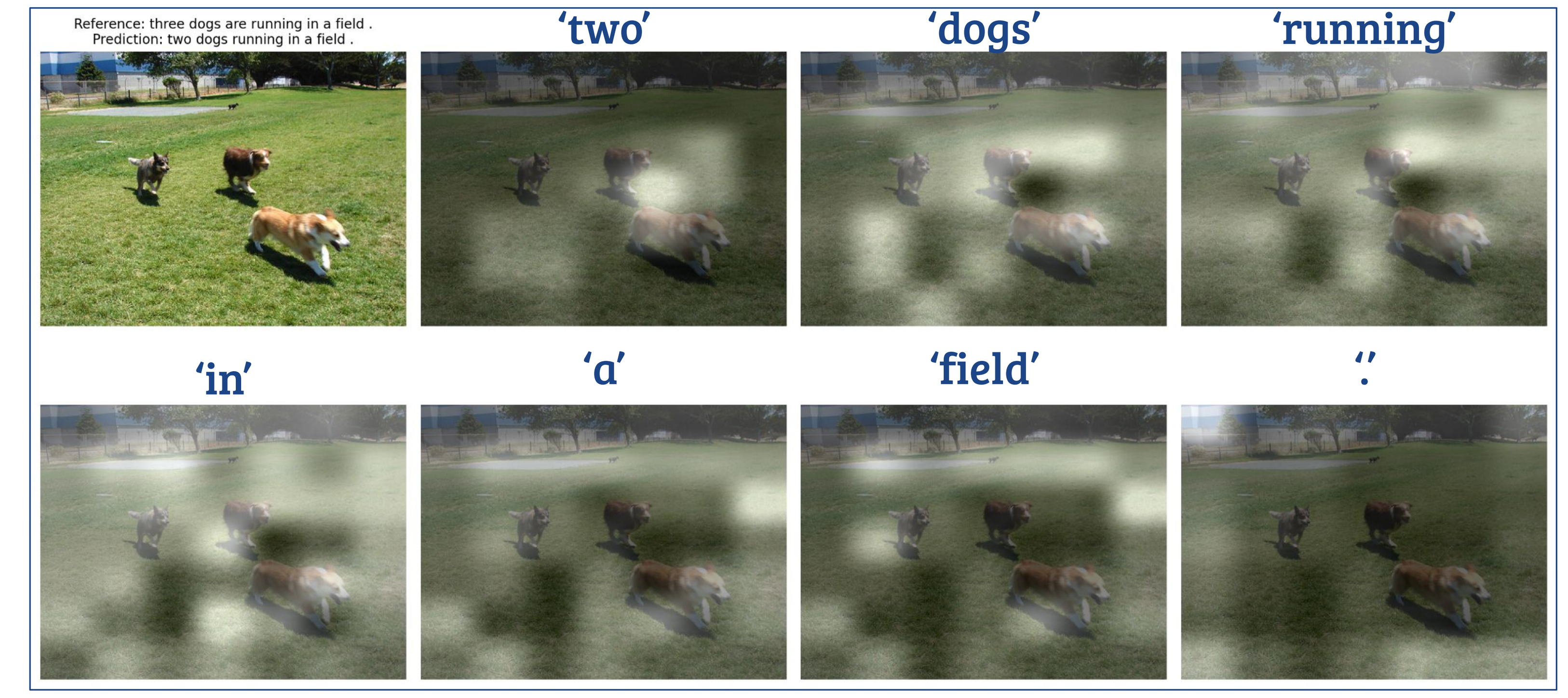
A girl in a red striped shirt. ❌



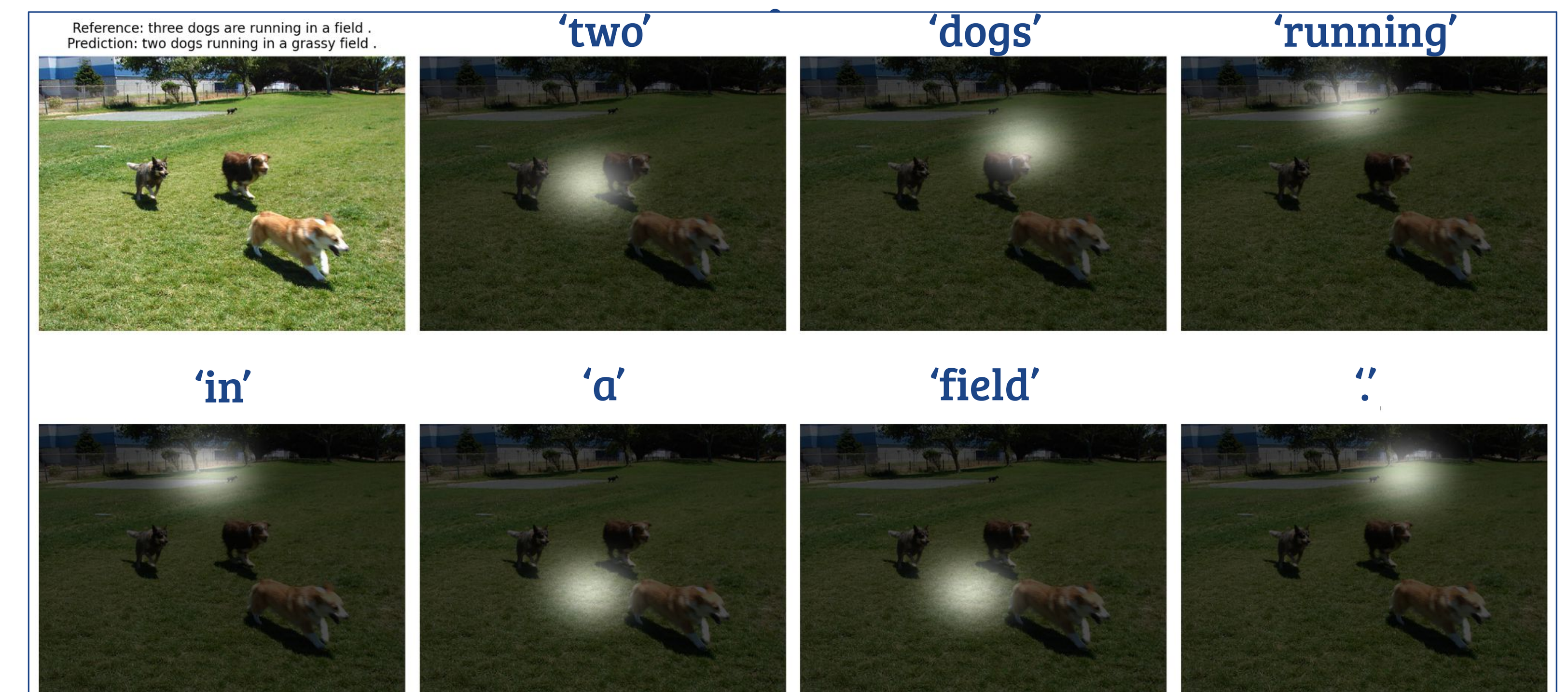
Two men skiing down a snowy hill. ❌

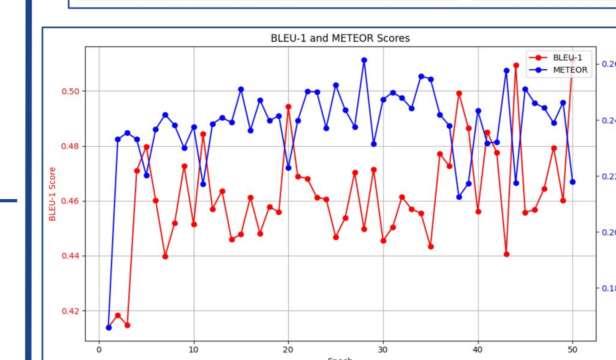
Results

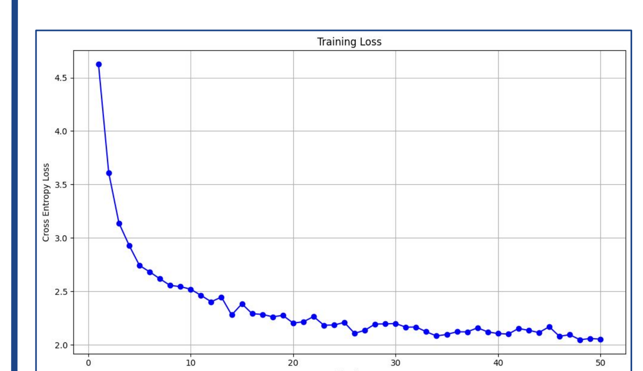
Soft



Hard





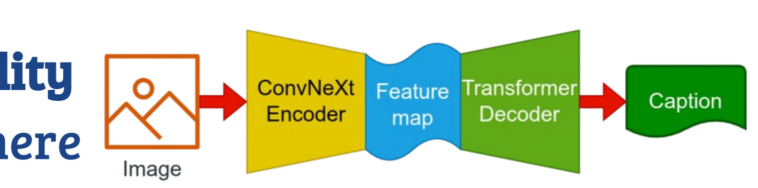


Metric	Xu et al.	Reproduced
Soft Attention		
BLEU-1	**67.0**	45.9
METEOR	18.93	**18.96**
Hard Attention		
BLEU-1	**67.0**	43.2
METEOR	20.30	**20.75**

Train Loss/Validation Meteor Curves, METEOR/BLEU-1 Inference

Conclusion

- Attention-based models improve caption quality and **interpretability**
- Soft and hard attention guide where the model "looks" when generating
- Outperformed paper METEOR results



[Figure 7]

[1] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv preprint arXiv:1502.03044. M. Hodosh, P. Young and J. Hockenmaier (2013)

[2] "Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics", Journal of Artifical Intellegence Research, Volume 47, pages 853-899 http://www.jair.org/papers/paper3994.html

[3] https://images.ctfassets.net/7ajcefednbt4/3L9DDPPSZiCIBtTaKq3L3U/a08cbaa48d19a1111a0472c8a643934c/Marathon_runners___BABAROGA.jpg

[4] https://iq.opengenus.org/content/images/2021/09/encoder--decoder.JPG

[5] https://www.bridgerev.com/hs-fs/hubfs/Untitled%20design%20-%202021-04-29T134930.082.png?width=525&height=350&name=Untitled%20design%20-%202021-04-29T134930.082.png

[6] https://render.fineartamerica.com/images/rendered/default/poster/6/8/break/images/artworkimages/medium/2/american-football-in-spotlight-siri-stafford.jpg

[7] https://starbeamrainbowlabs.com/blog/images/20220904-image-captioning-ai-arch.png