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

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio. *ICML 2015*

Presented By: Sai Krishna Bollam

Outline

- Introduction
- Model Overview
- Model Details
  - Encoder
  - Decoder
  - Attention
- Experiments
- Results
- Conclusion

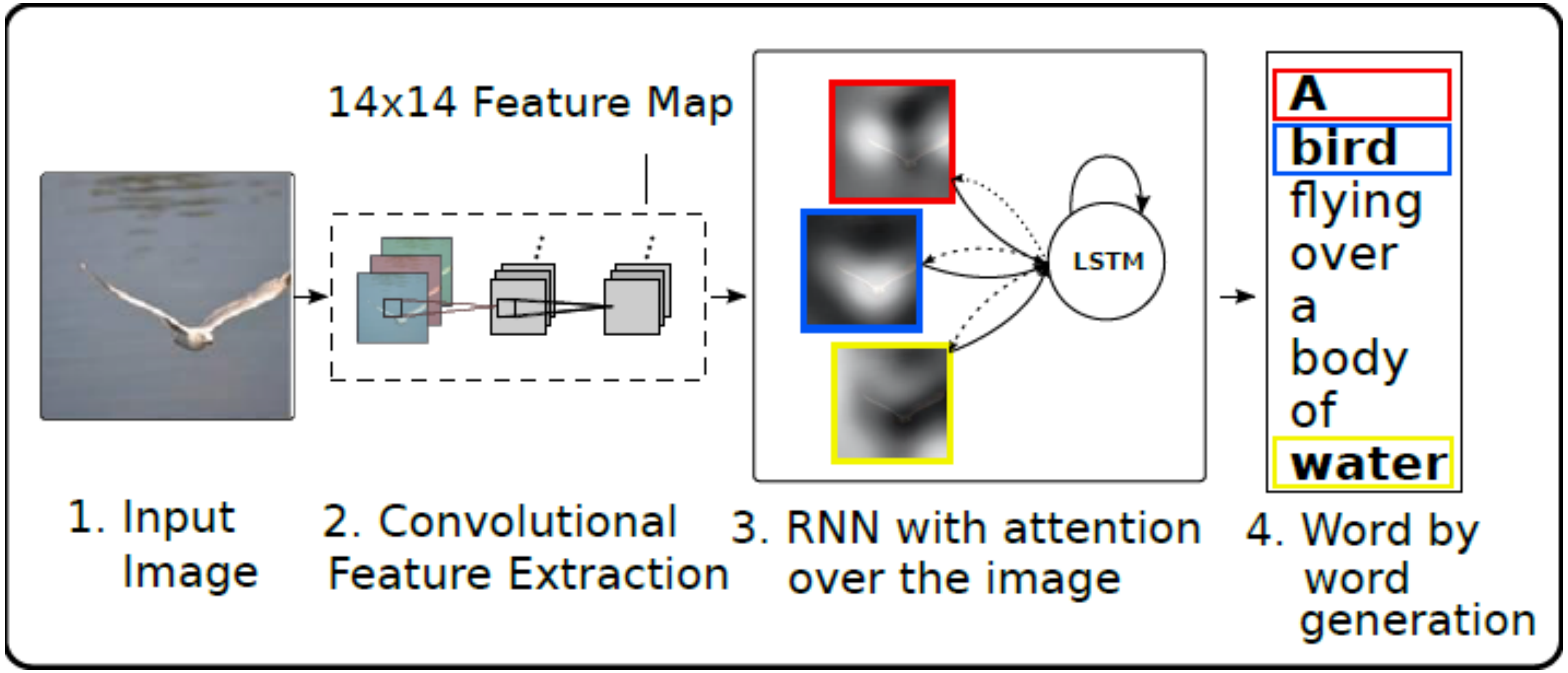
Introduction

- Multimodal Machine Learning
  - Relate information from multiple modalities: speech, image, language etc.
- Scene understanding
  - Automatic caption generation
- **Task:** Given an image, generate a sentence describing it
  - Object Detection and Machine Translation
  - *Image* to *Language* translation

A bird flying over a body of water.

A woman throwing a frisbee in a park.

Model Overview



Encoder-Decoder framework

Analogous to translation but..

Encoder output is not a single vector

Learn alignments from scratch

Using Attention over low level feature maps

Instead of joint object-text embedding

Bahdanau et al. (2014)

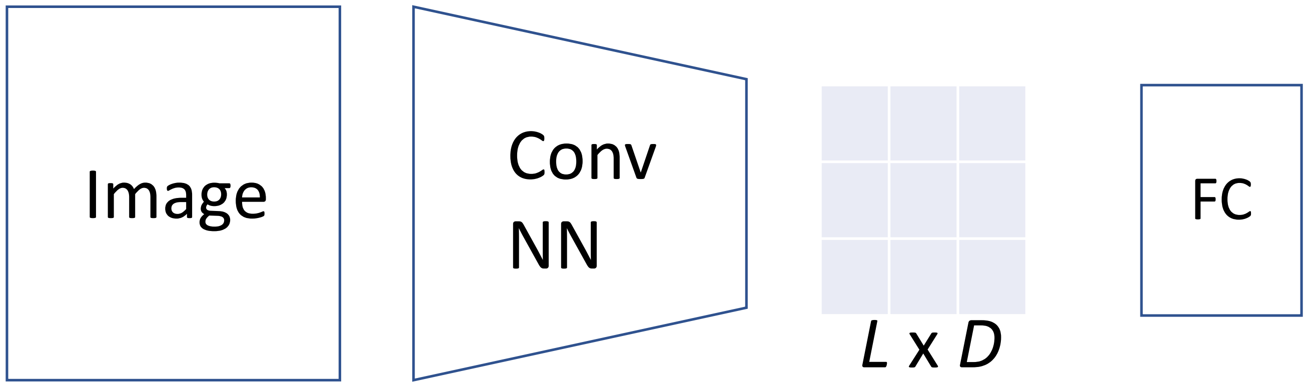
Model Details: Encoder

- Model:
  - Input: Raw image
  - Output: Sequence of *C* words from vocabulary of size *K*

$$\boldsymbol{y} = \{y_1, \dots, y_C\}, y_i \in \mathbb{R}^K$$

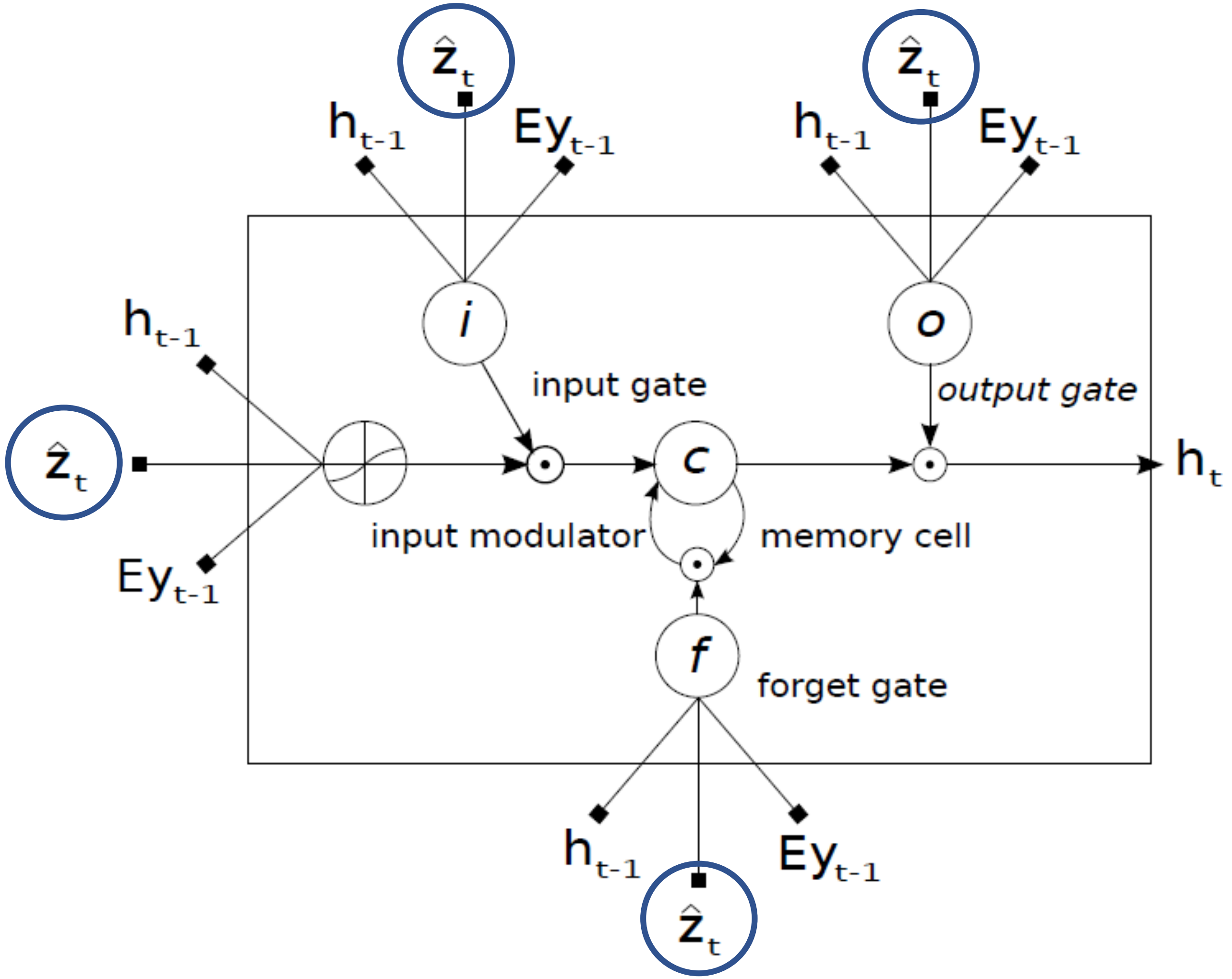
- Encoder: Convolutional Neural Network
  - Input: Raw image
  - Output: multiple feature vectors (annotation vectors) from lower conv layers

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



Model Details: Decoder

- LSTM Network



$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{Ey}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

$\widehat{z}_t$: Context vector

Model Details: Decoder – Context Vector

Context Vector ($\hat{z}_t$): A dynamic representation of relevant part of image at time t

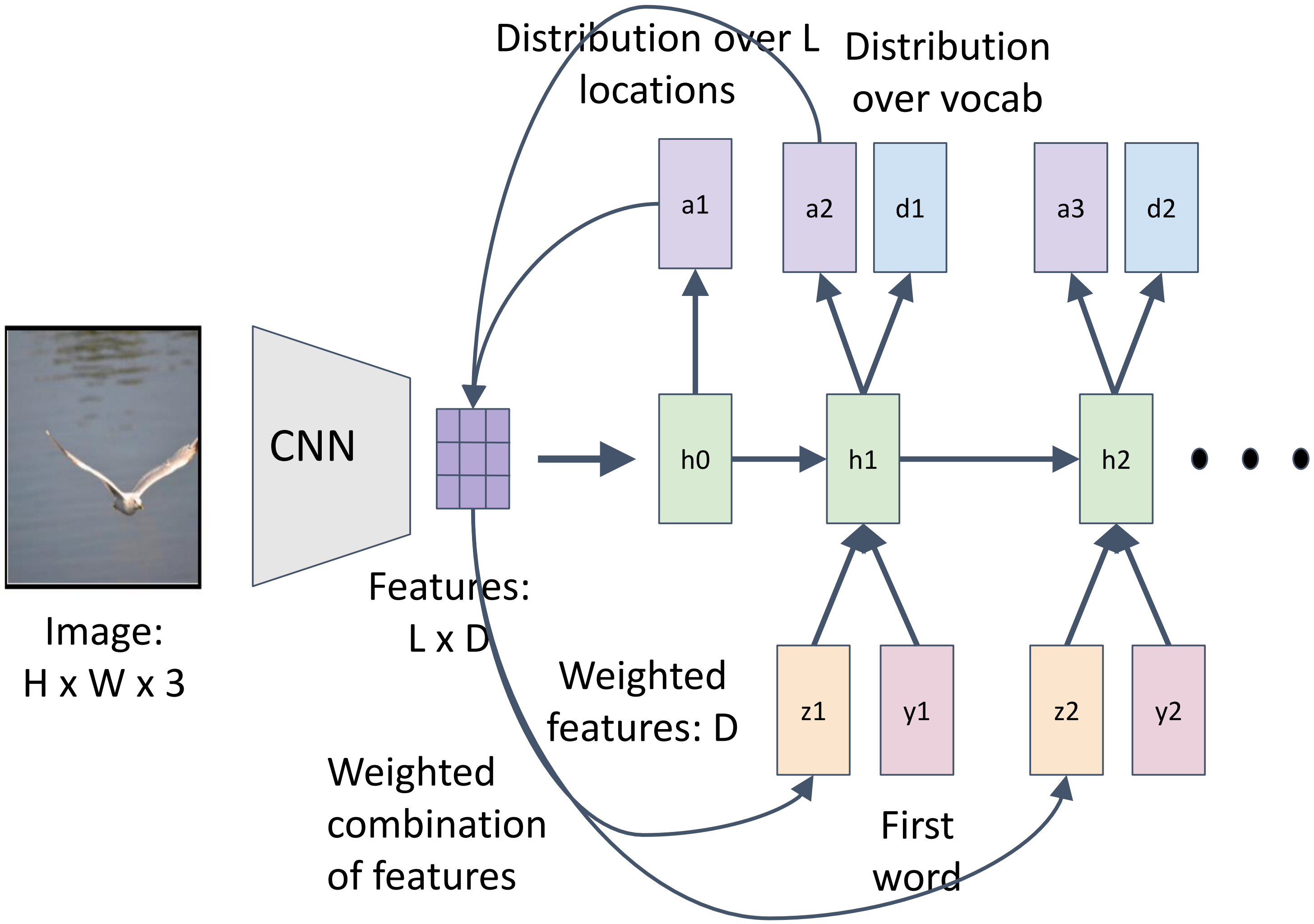
$$\hat{z}_t = \phi(\underbrace{\{a_i\}}_{\text{Annotation vectors}}, \underbrace{\{\alpha_i\}}_{\text{Attention. Calculated using } f_{att}})$$

f_{att} : Attention Model, an MLP conditioned on previous hidden state

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

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

Model Representation



Attention Mechanism: Stochastic Attention

Stochastic "Hard" Attention

At every time step, focus on exactly 1 location (a_i)

$s_{t,i} = 1$ iff i^{th} location is used to extract visual features

$$p(\, s_{t,i} = 1 \mid s_{j<t}, \mathbf{a} \,) = \alpha_{t,i} = \mathrm{softmax}\,(f_{att}(\mathrm{a}_i, \mathbf{h}_{t-1}))$$

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

Sample a_i based on Multinoulli distribution

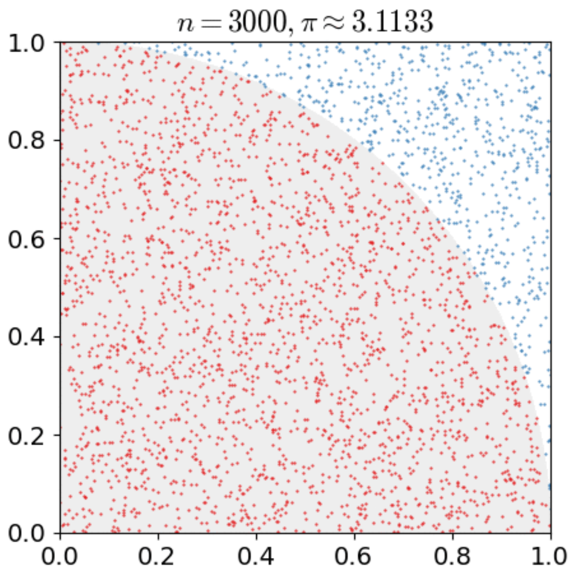
$$L_s = \sum_s p(\, s \mid \mathbf{a} \,) \log p(\, \mathbf{y} \mid s, \mathbf{a} \,) \leq \log[p(\, s \mid \mathbf{a} \,)p(\, \mathbf{y} \mid s, \mathbf{a} \,)]$$

$$= \log p(\mathbf{y} \mid \mathbf{a})$$

Attention Mechanism: Stochastic Attention

$$\frac{\partial L_s}{\partial W} = \sum_s p(s \mid \mathbf{a}) \left[\frac{\partial \log p(\mathbf{y} \mid s, \mathbf{a})}{\partial W} + \log p(\mathbf{y} \mid s, \mathbf{a}) \frac{\partial \log p(s \mid \mathbf{a})}{\partial W} \right]$$

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



Monte Carlo based sampling approximation

Wikipedia

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

REINFORCE learning rule

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

Attention Mechanisms: Deterministic Attention

Deterministic "Soft" Attention

Expectation of context vector, instead of sampling. Differentiable!

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

$$\phi(\{\mathrm{a}_i\},\{\alpha_i\}) = \sum_i^L \alpha_i \mathrm{a}_i$$

Soft attention weighted vector

Normalized Weighted Geometric Mean

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

$$NWGM[p(y_t = k \mid \mathbf{a})] \approx \mathbb{E}[p(y_t = k \mid \mathbf{a})]$$

Attention Mechanisms: Deterministic Attention

Doubly Stochastic Attention

Introduce regularization: $\sum_t \alpha_{t,i} \approx 1$

Encourages model to pay equal attention to every part of image over time

$$\beta_t = \sigma(f_\beta(\mathbf{h}_{t-1}))$$

$$\phi(\{a_i\}, \{\alpha_i\}) = \beta \sum_i^L \alpha_i a_i$$

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

Attention Mechanisms

Image:
H x W x 3

CNN

a	b
c	d

Grid of features
(Each D-dimensional)

From RNN:

p_a	p_b
p_c	p_d

Distribution over grid locations
$p_a + p_b + p_c + p_c = 1$

Context vector z
(D-dimensional)

Soft attention:
Summarize ALL locations
$z = p_a a + p_b b + p_c c + p_d d$

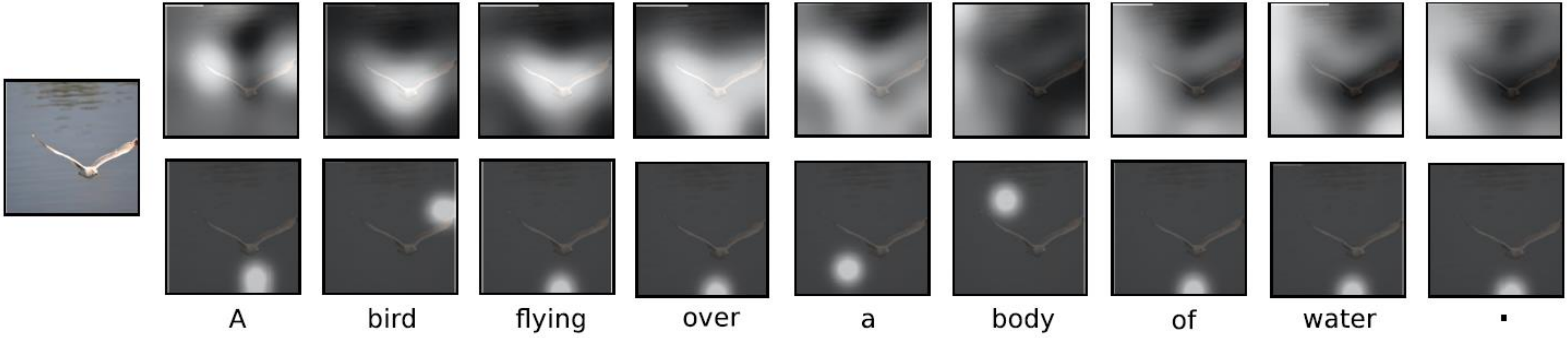
Derivative dz/dp is nice!
Train with gradient descent

Hard attention:
Sample ONE location according to p, z = that vector

With argmax, dz/dp is zero almost everywhere …
Can’t use gradient descent; need reinforcement learning

Visualizing Attention

Soft attention



Hard attention

Experiments

Encoder:

Oxford VGGnet

- pretrained on ImageNet
- Feature maps from 4th conv layer before pooling. 14x14x512 flattened to 196 x 512 (L x D)

Training:

Flickr8k: RMSProp

Flickr30k/MS COCO: Adam

Dropout

Early stopping on BLEU

Batching by sentence lengths

Datasets:

Flickr8k	: 8,000
Flickr30k	: 30,000
MS COCO	: 82,783
Vocabulary	: 10,000 words

5 reference sentences per image

Metrics:

BLEU-1, 2, 3, 4

- No brevity penalty

METEOR

Results

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

Dataset	Model	BLEU				
		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

Key Points

- Learn latent alignments from scratch
  - Better context to decoder
  - Attends to non object regions
  - Joint representation
- Visualizing attention to interpret functioning
  - Stochastic Attention
  - Deterministic Attention

Thank you

Questions?