Generating Visual Explanations

Hendricks, Lisa Anne, et al. "Generating visual explanations." *European Conference on Computer Vision*. Springer International Publishing, 2016.

Content

- Objective
- LRCN: Visual description model
- Relevance Loss
- Discriminative Loss
- Combined Loss
- Evaluation Results

Objective

- Jointly predicts a class label, and explains why the predicted label is appropriate for the image.

- Introspection vs Justification explanation systems

“This is a Western Grebe because filter 2 has a high activation…”

vs

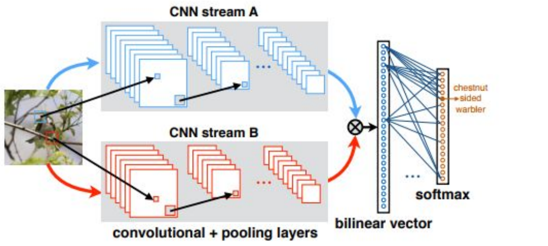
“This is a Western Grebe because it has red eyes…”

This is a **Bronzed Cowbird** *because ...*

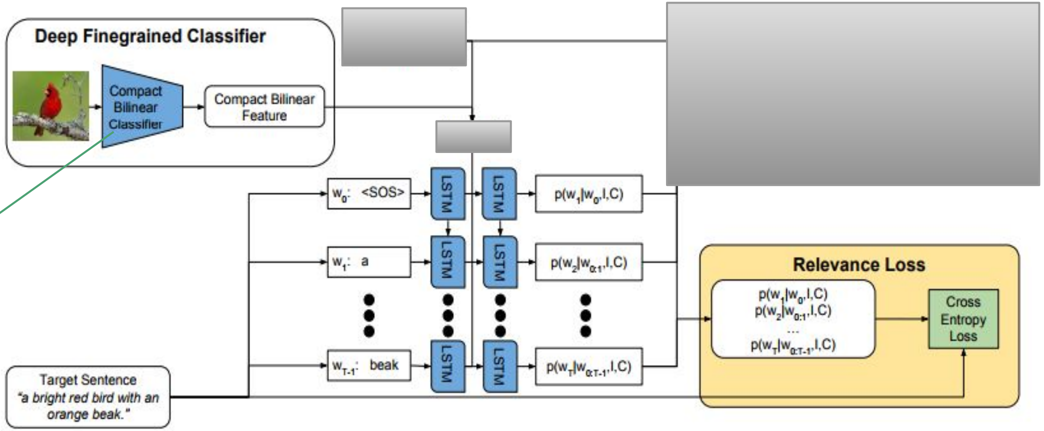
Definition:	this bird is black with blue on its wings and has a long pointy beak.
Description:	this bird is nearly all black with a short pointy bill.
Explanation-Label:	this bird is nearly all black with bright orange eyes.
Explanation-Dis.:	this is a black bird with a red eye and a white beak.
Explanation:	this is a black bird with a red eye and a pointy black beak.

Visual Description based on LRCN*

(This only generates descriptions not explanations)

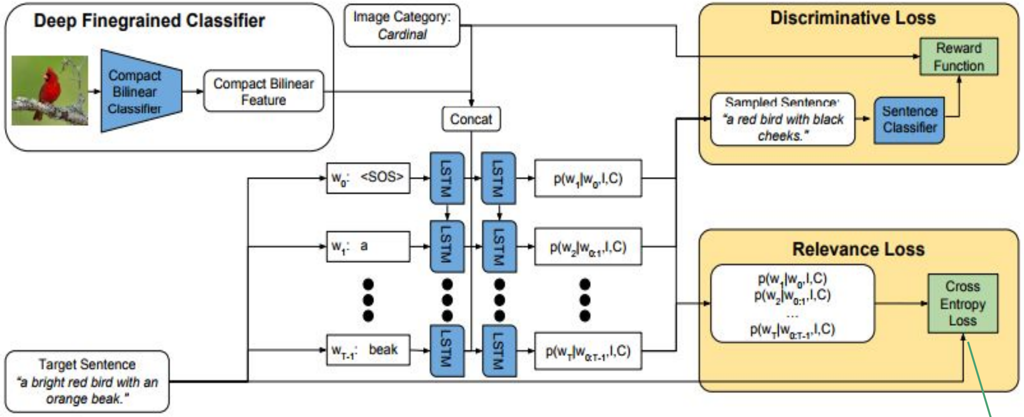


8192 dimensional penultimate layer



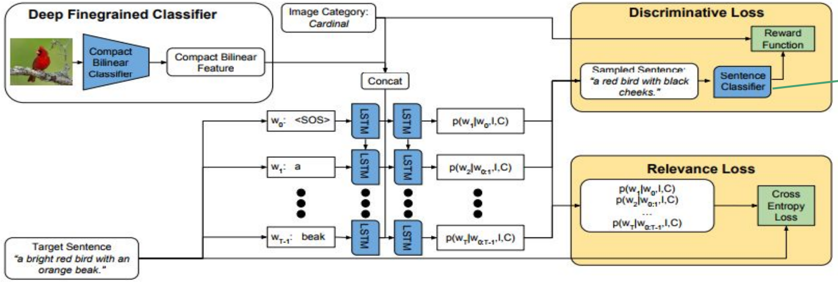
*LRCN: Long-term Recurrent Convolutional Networks

Visual Explanation Model: Relevance Loss



$$L_R = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \log p(w_{t+1}|w_{0:t}, I, C)$$

Visual Explanation Model: Discriminative Loss

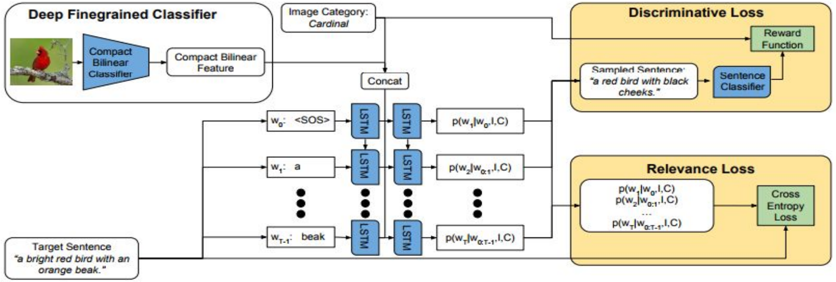


Single Layer LSTM-based classification network

- Discriminative Loss: $\mathbb{E}_{\tilde{w}\sim p(w)}\left[R_D(\tilde{w})\right],$
- Monte Carlo sampling of descriptions (w') from p(w/I,C)
- Sampling operation is non smooth i.e. $\nabla_W R_D(\tilde{w})$ is undefined.
- Using REINFORCE's equivalence property

$$\nabla_W \mathbb{E}_{\tilde{w}\sim p(w)}\left[R_D(\tilde{w})\right] = \mathbb{E}_{\tilde{w}\sim p(w)}\left[R_D(\tilde{w})\nabla_W \log p(\tilde{w})\right]$$

Visual Explanation Model: Combined Loss



- The sampled gradient term $\nabla_W \log p(\tilde{w})$ is weighted by the reward $[R_D(\tilde{w})]$
- Pushing the weights to increase likelihood of highly rewarded explanations.
- Reward is defined as

$$R_D(\tilde{w}) = p(\hat{C}|\tilde{w})$$

- Overall Loss function and gradient

$$L_R - \lambda \mathbb{E}_{\tilde{w}\sim p(w)}\left[R_D(\tilde{w})\right] \qquad \nabla_W L_R - \lambda R_D(\tilde{w}) \nabla_W \log p(\tilde{w}).$$

Visual Explanation Model: Evaluation

- Caltech UCSD Birds (CUB) dataset.
- 200 classes. 11,788 images. 5 descriptive sentences per image.
- Image relevance evaluation metrics:
  - METEOR: Matching words (and synonyms) between generated and reference sentences per image.
  - CIDEr: Additionally rewards uncommon (tf-idf weighted) n-grams in generated sentences per image.
- Class Relevance
  - Class similarity CIDEr: Ground truth is combined image descriptions within a class.
  - Class Rank Metric.
- Human Evaluation
  - Expert bird-watcher evaluation of 91 random explanations.

Visual Explanation Model: Results

Model Comparison

- Label
- Image
- Image + Label
- Image + Discriminative Loss
- Image + Label + discriminative Loss

	Image Relevance		Class Relevance		Best Explanation
	METEOR	CIDEr	Similarity	Rank (1-200)	Bird Expert Rank (1-5)
Definition	27.9	43.8	42.60	15.82	2.92
Description	27.7	42.0	35.3	24.43	3.11
Explanation-Label	28.1	44.7	40.86	17.69	2.97
Explanation-Dis.	28.8	51.9	43.61	19.80	3.22
Explanation	**29.2**	**56.7**	**52.25**	**13.12**	**2.78**

This is a **Bronzed Cowbird** *because ...*

Definition: this bird is black with blue on its wings and has a long pointy beak.

Description: this bird is nearly all black with a short pointy bill.

Explanation-Label: this bird is nearly all black with bright orange eyes.

Explanation-Dis.: this is a black bird with a red eye and a white beak.

Explanation: this is a black bird with a red eye and a pointy black beak.

End of Slides