Title: *Generating Visual Explanations*

Authors: Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach,

Jeff Donahue, Bernt Schiele, Trevor Darrell

Publication date: 28/03/2016.

Talk by Michael Aichmüller, 05/07/2018.

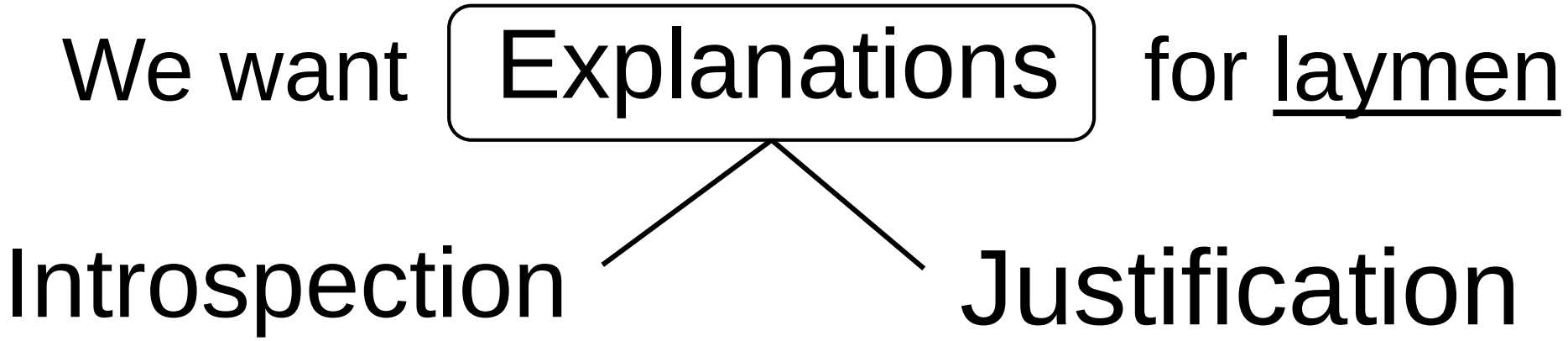
Table of contents

Motivation

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

We want Explanations for laymen

Motivation


We want
Explanations
for laymen
Introspection
Justification

Motivation

We want **Explanations** for laymen

Introspection

explain how a model determines its final output technically

"Car is a Ford, because neuron A and B activated"

Justification

Motivation

We want **Explanations** for laymen

Introspection

explain how a model determines its final output technically

"Car is a Ford, because neuron A and B activated"

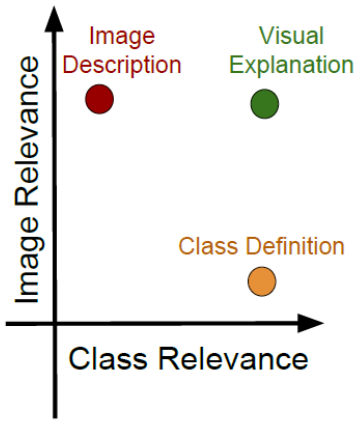
Justification

produce sentences detailing how system output and visual evidence correlate

"Car is a Ford, because the chassis has characteristic X

Motivation

The difference between descriptive, defining, and explaining sentences.



Laysan Albatross

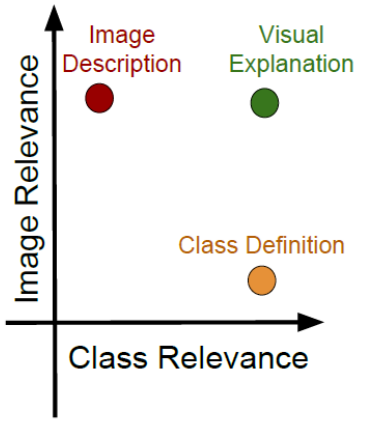
Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Motivation

The difference between descriptive, defining, and explaining sentences.



Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

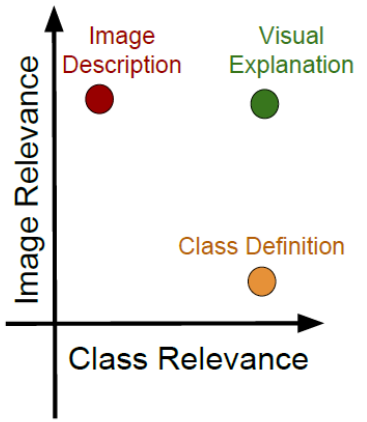
Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Motivation

The difference between descriptive, defining, and explaining sentences.



Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

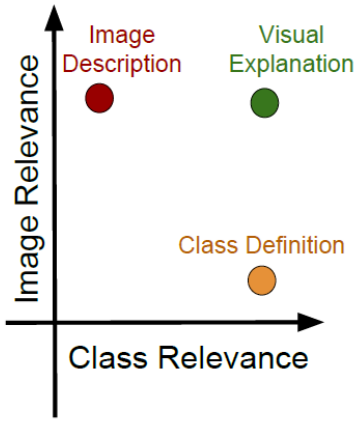
Description: This is a large bird with a white neck and a black back in the water.

Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.

Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

Motivation

The difference between descriptive, defining, and explaining sentences.



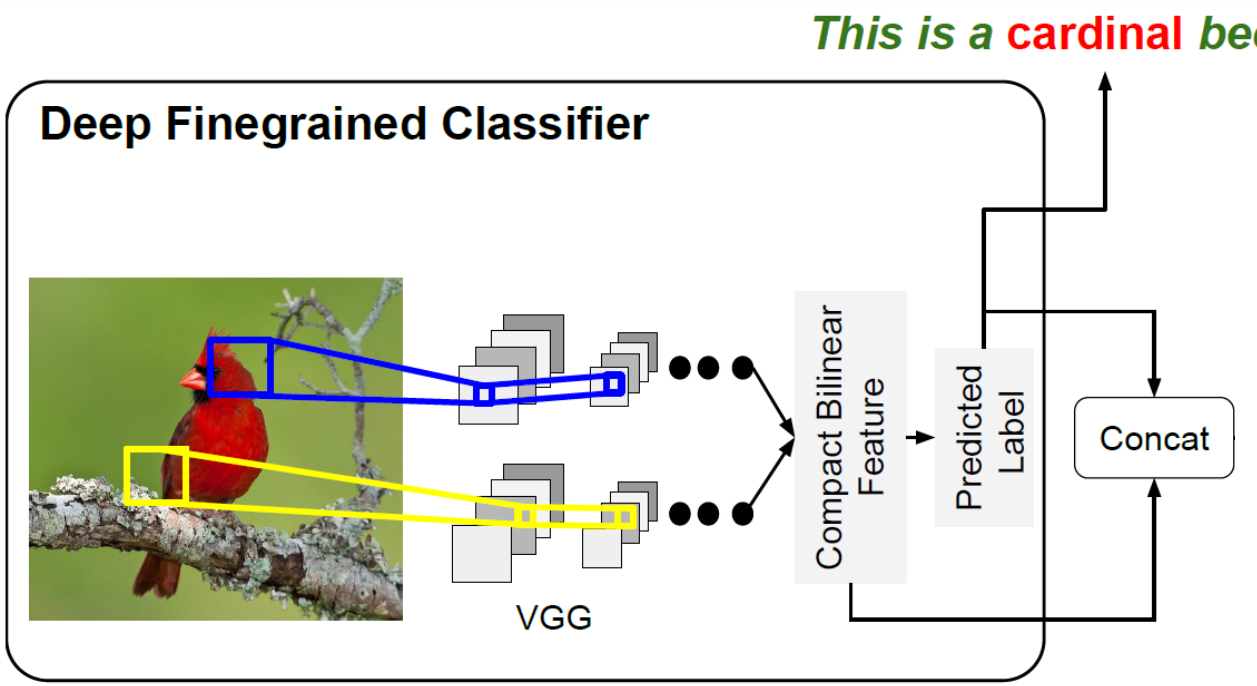
Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

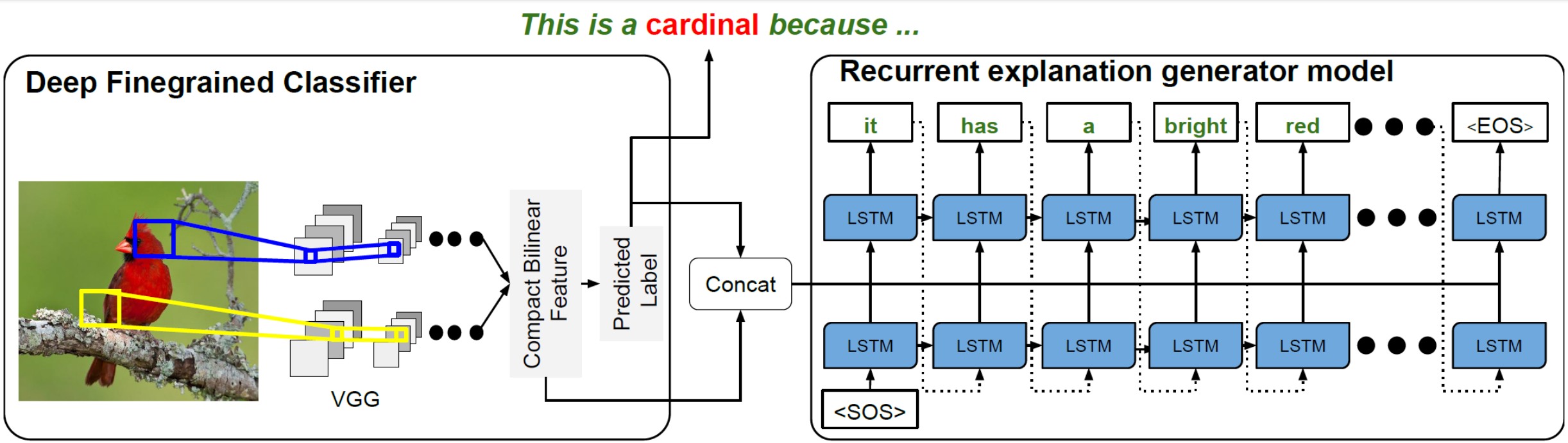
Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

The Model Formulation



- Model extracts visual features using a fine-grained classifier before language generation.

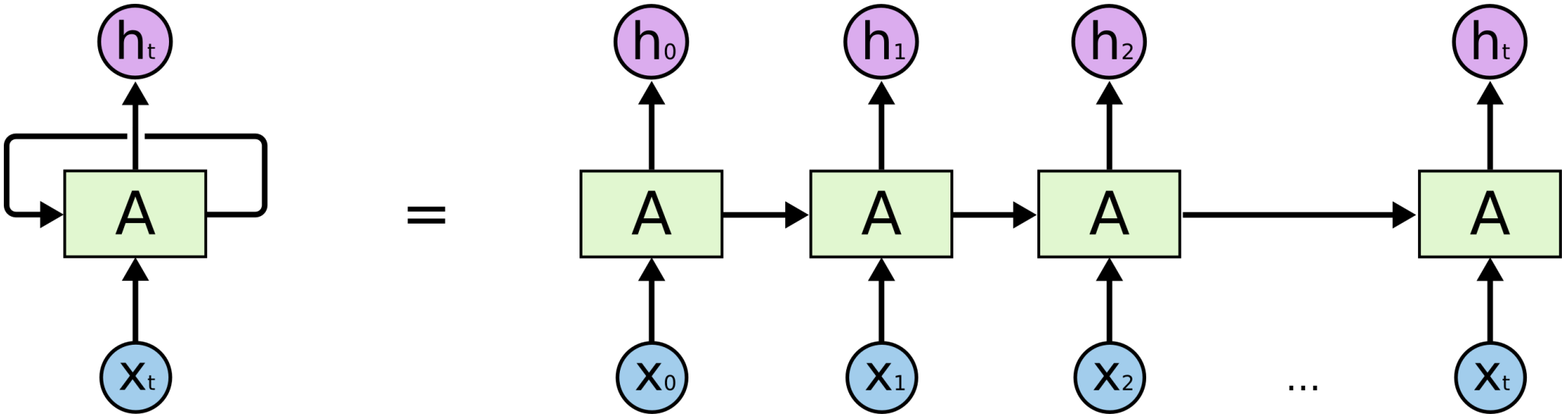
The Model Formulation



- Model extracts visual features using a fine-grained classifier before language generation.
- Additionally, unlike description models, sentence generation is conditioned on the predicted class label

The recurrent network

The unrolled network structure

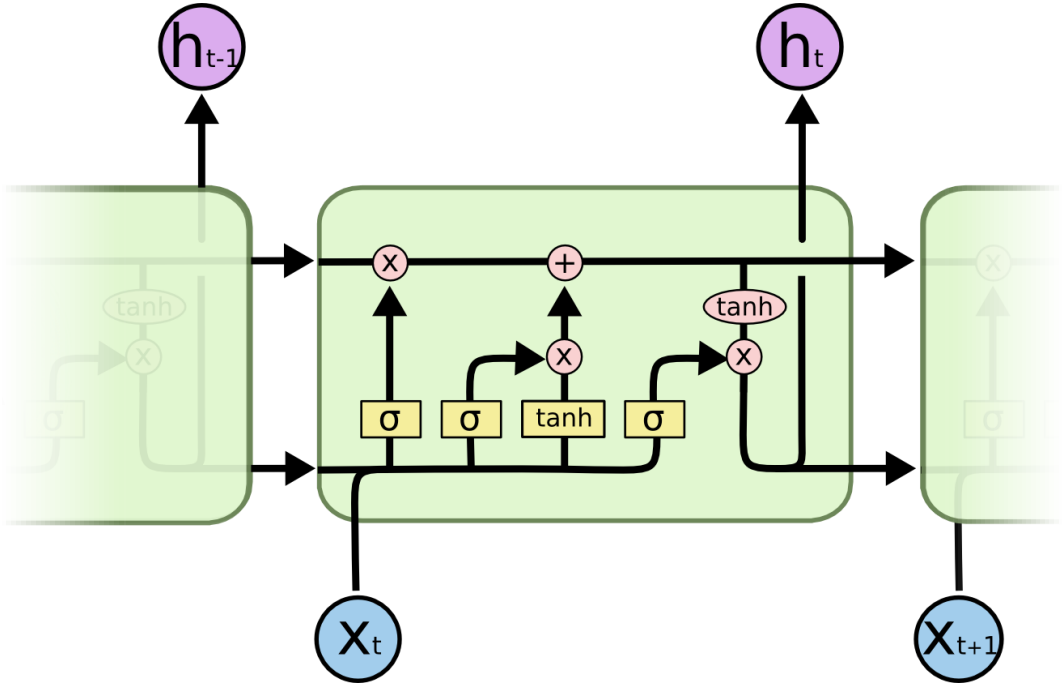


Good for sequential data, e.g. language:

$$(X_t)_{t=1,2,\ldots} = (\text{It}, \text{was}, \text{raining}, \text{so}, \text{the}, \text{road}, \text{is}, \ldots)$$

The layout of an LSTM layer

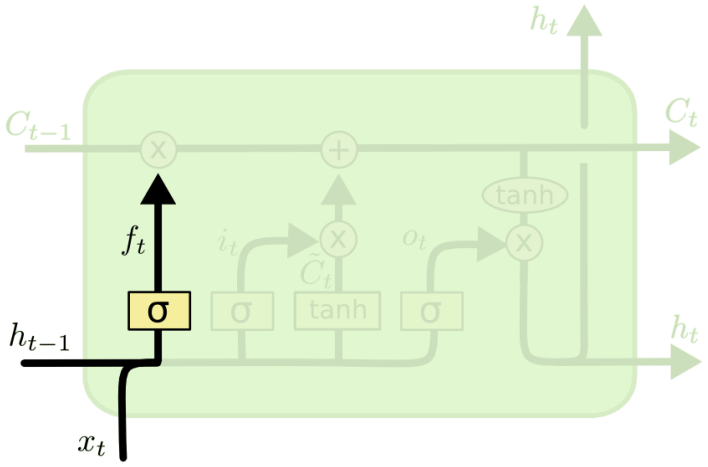
3 gates to modify data stream:



The layout of an LSTM layer

3 gates to modify data stream:

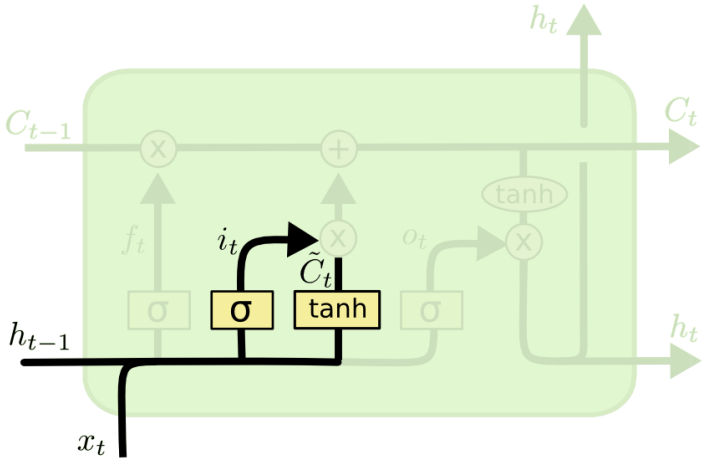
1. **Forget**: Decide which part of existing stream to keep relevant.



The layout of an LSTM layer

3 gates to modify data stream:

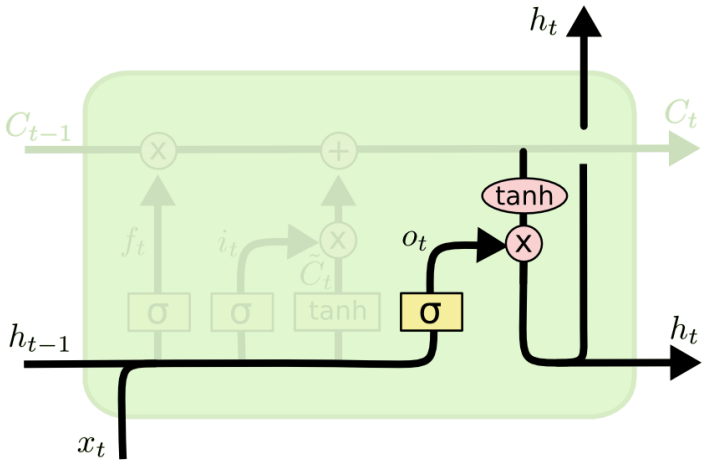
1. **Forget**: Decide which part of existing stream to keep relevant.
2. **Input**: Decide which new information to store in the stream (2-step: sigmoid to mark which values $\tilde{i}_t$, tanh to create new values $\tilde{C}_t$).



The layout of an LSTM layer

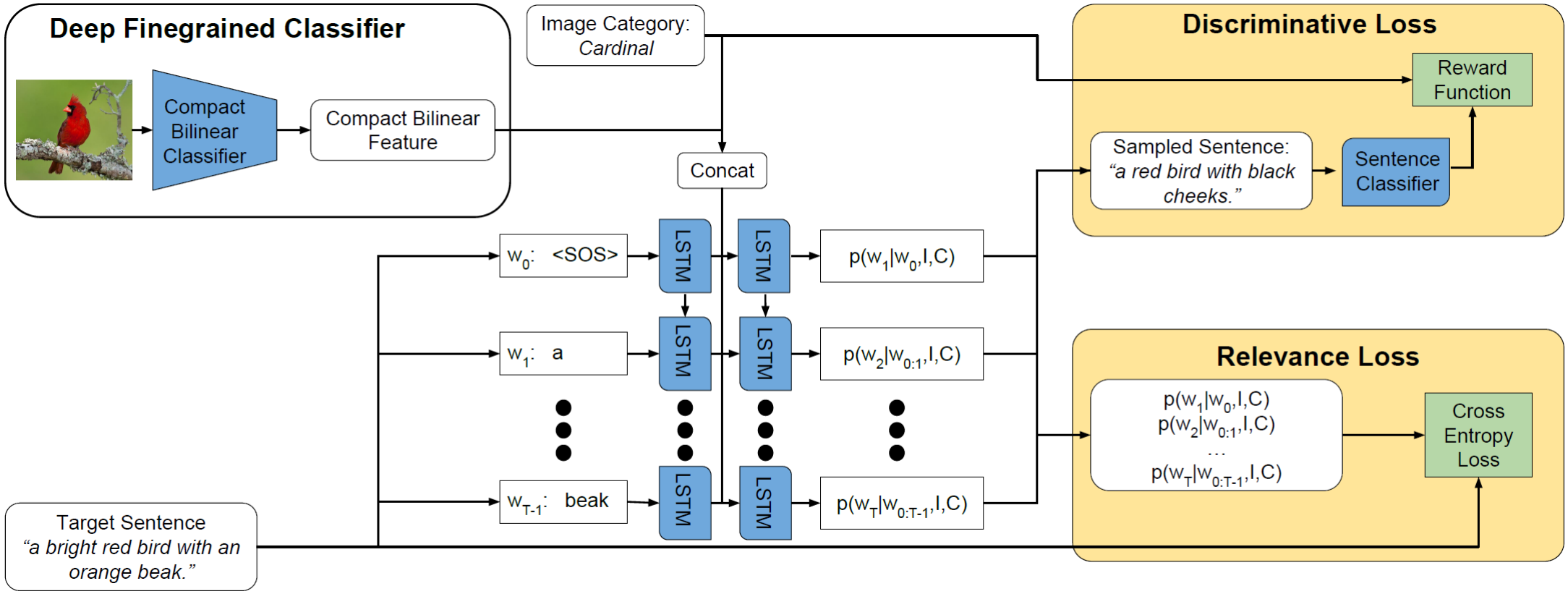
3 gates to modify data stream:

1. **Forget**: Decide which part of existing stream to keep relevant.
2. **Input**: Decide which new information to store in the stream (2-step: sigmoid to mark which values $\tilde{\imath}_t$, tanh to create new values $\tilde{C}_t$).
3. **Output**: Decide which part to output from the memory stream C_{t-1} and the hidden state h_{t-1}. Creates new hidden state h_t, the output.



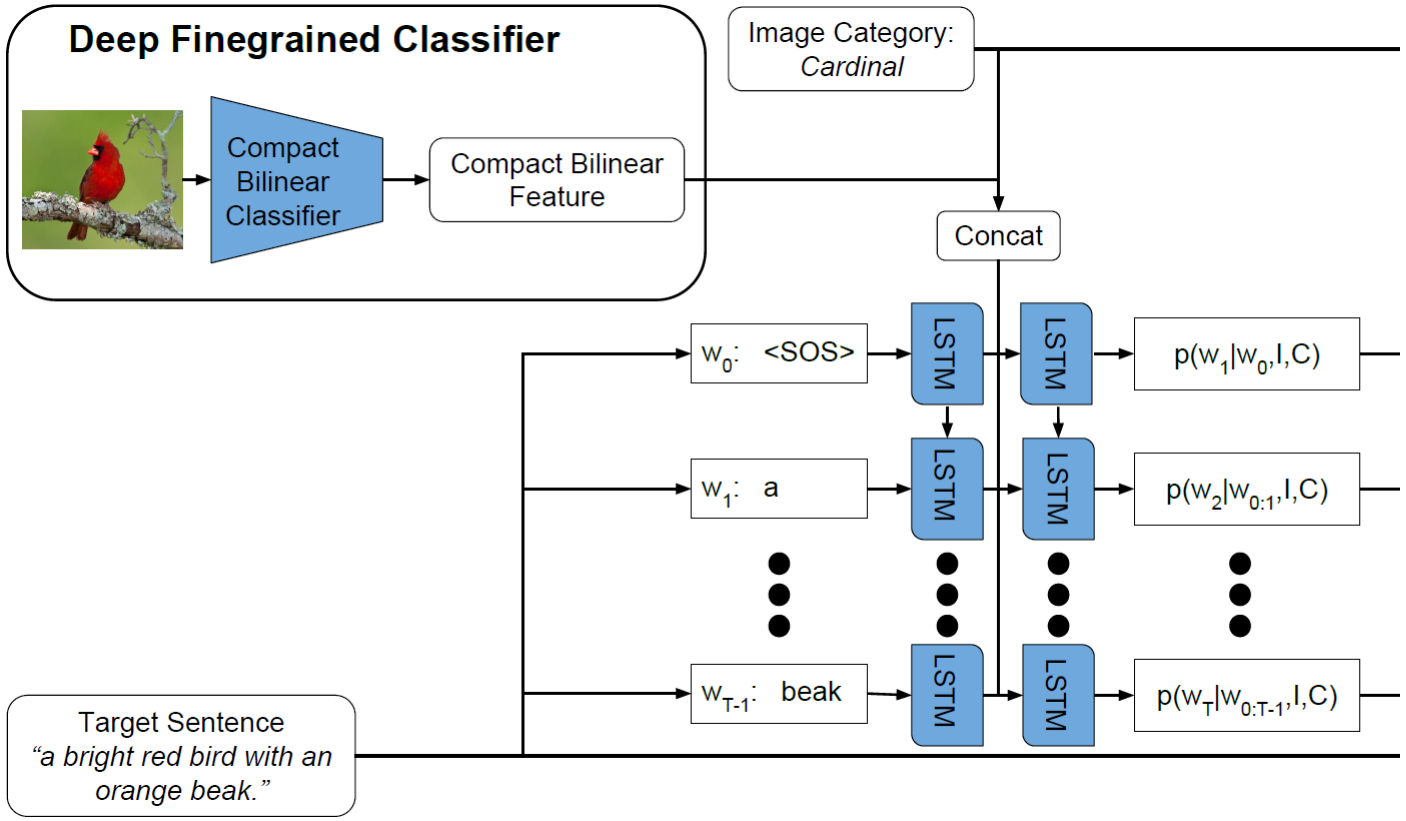
Loss Functions

How does the model train?



Loss Functions

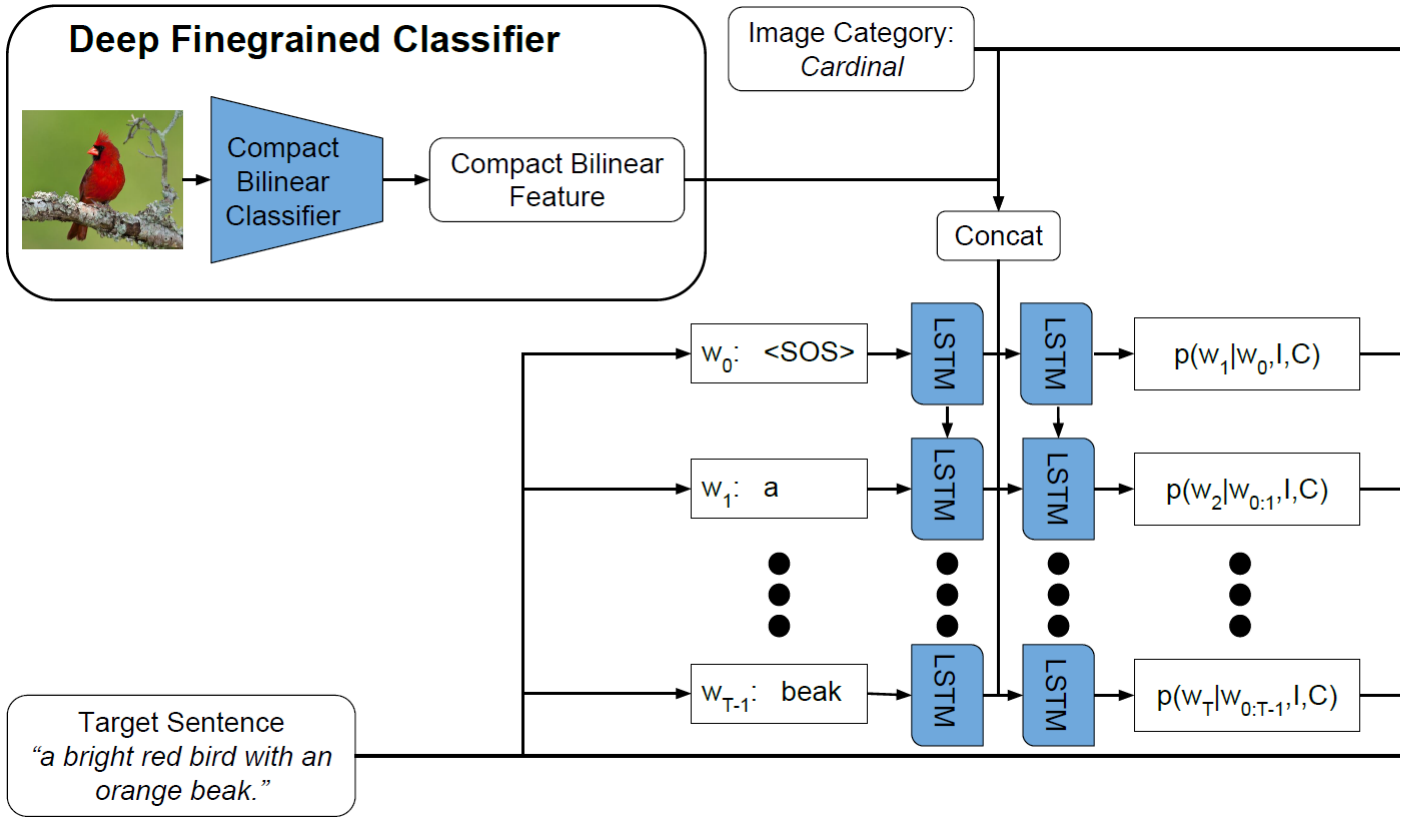
How does the model train?



Finegrained Classifier picks out detail features and category.

Loss Functions

How does the model train?



Finegrained Classifier picks out detail features and category.

At training time:

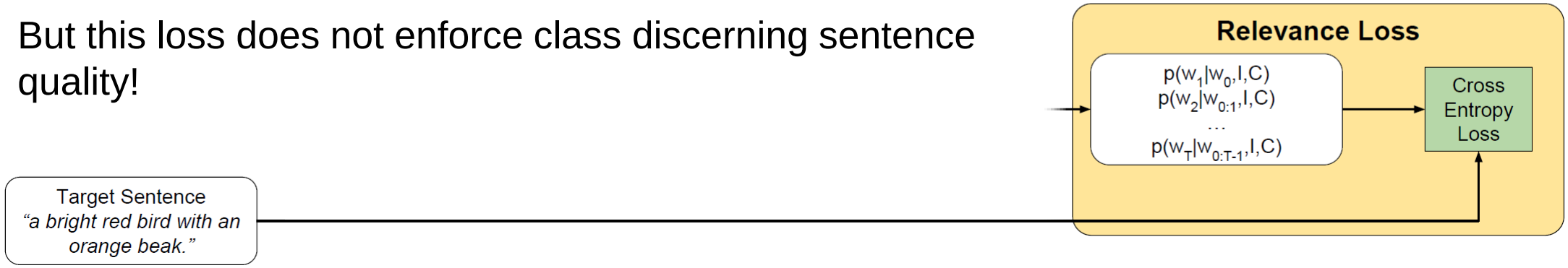
- 1st LSTM stack is given the target words as input sequentially
- 1st stack output, features & category are inputs for 2nd stack.
- Outputs of 2nd stack:
  word probabilities conditioned on previous words, image features and category

The Relevance Loss

Cross Entropy (approximation) ensures image relevance for generated sentences
(word in position t w_t, Image I, category C, batch size N):

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

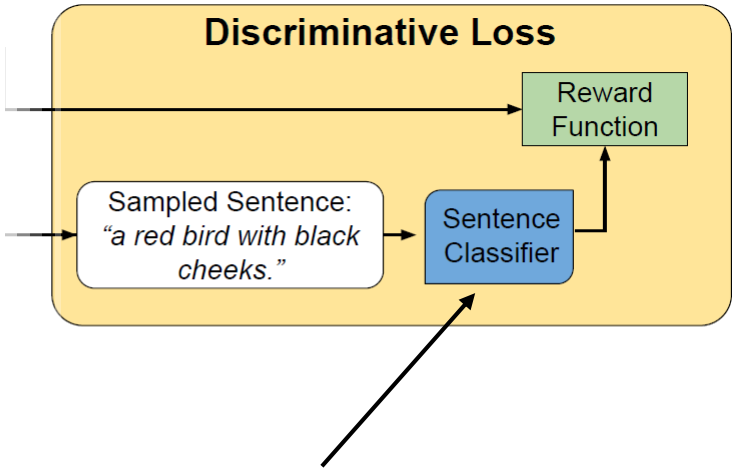
But this loss does not enforce class discerning sentence quality!



The Discriminative Loss

Introduce Discriminator Reward $R_D(\widetilde{w}) = p(C|\widetilde{w})$ to build overall loss function

$$L_R - \lambda \mathbb{E}_{\widetilde{w} \sim p(w|I,C)}(R_D(\widetilde{w}))$$



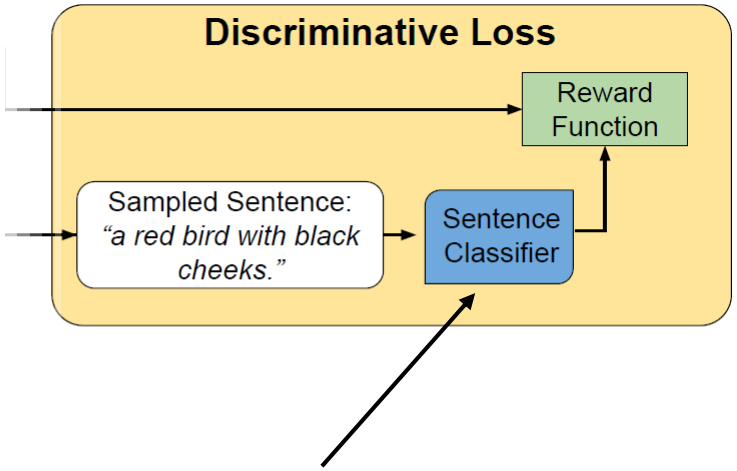
Sentence Classifier (pretrained) predicts class of original image the sentence is supposed to describe.

The Discriminative Loss

Introduce Discriminator Reward $R_D(\widetilde{w}) = p(C|\widetilde{w})$ to build overall loss function

$$L_R - \lambda \mathbb{E}_{\widetilde{w} \sim p(w|I,C)}(R_D(\widetilde{w}))$$

- Expectation is untractable -> sample description sentences $\widetilde{w}$ from LSTM stack.



Sentence Classifier (pretrained) predicts class of original image the sentence is supposed to describe.

The Discriminative Loss

Introduce Discriminator Reward $R_D(\tilde{w}) = p(C|\tilde{w})$ to build overall loss function

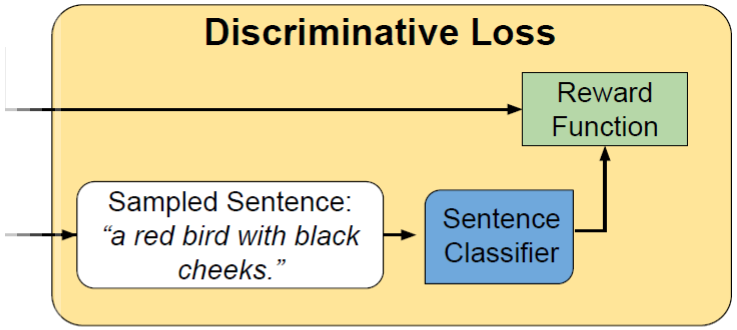
$$L_R - \lambda \mathbb{E}_{\tilde{w} \sim p(w|I,C)}(R_D(\tilde{w}))$$

Use REINFORCE loss function equivalence

$$\nabla_W \mathbb{E}_{\tilde{w} \sim p(w|I,C)}(R_D(\tilde{w})) = \mathbb{E}_{\tilde{w} \sim p(w)}(R_D(\tilde{w}) \nabla_W \log p(\tilde{w}))$$

to update the loss with

$$\nabla_W L_R - \lambda \mathbb{E}_{\tilde{w} \sim p(w|I,C)}(R_D(\tilde{w}) \nabla_W \log p(\tilde{w}))$$



Loss Functions

Examples of the model output

This is a Kentucky warbler because this is a yellow bird with a black cheek patch and a black crown.

This is a pine grosbeak because this bird has a red head and breast with a gray wing and white wing.

This is a pied billed grebe because this is a brown bird with a long neck and a large beak.

This is an artic tern because this is a white bird with a black head and orange feet.

Evaluation Metrics and Setup

Model applied to Caltech UCSD Birds 200-2011 (CUB) dataset:

- 200 classes of North American bird species and 11788 images
- Each image contains 5 detailed description sentences

Bilinear Classifier pretrained on this dataset with up to 8192 dimensional features extraction and 84% accuracy.

Evaluation Metrics and Experimental Setup

Baseline comparisons to models of the same setup, except they are…

- conditioned only on the images **= Description model**
- conditioned only on the class label of the images **= Definition model**
- trained without discriminative loss **= Explanation label**
- trained without the predicted class label **= Explanation-discriminative**

Evaluation Metrics and Experimental Setup

Linguistic metrics used to verify…

Image relevance:

- **METEOr** computed by matching words in generated and reference sentences (also synonyms).
- **CIDEr** measures the similarity of two sentences by matching n-grams that are TF-IDF weighted.

Class relevance:

- **Class similarity metric** CIDEr calc. with all ground truth sentences of its own class.
- **Class Rank** CIDEr calc. with all ground truth sentences of each class.
- **Human bird experts evaluation.**

Comparison Results

Label
Image
Image + Label
Image + Discr. Loss
Image + Label + Discr.Loss

Better is…	**Image Relevance** METEOR ↑	CIDEr ↑	**Class Relevance** Similarity ↑	Rank (1-200) ↓	**Best Explanation** 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**

Comparison Results

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**.

Correct
Somewhat correct
Wrong

} **attributes**

Comparison Results

This is a **White Necked Raven** *because ...*

Definition:	this bird is **black in color** with a **black beak** and **black eye rings**.
Description:	this bird is **black** with a **white spot** and has a **long pointy beak**.
Explanation-Label:	this bird is **black** in color with a **black beak** and **black eye rings**.
Explanation-Dis.:	this is a **black** bird with a **white nape** and a **black beak**.
Explanation:	this is a **black** bird with a **white nape** and a **large black beak**.

Correct
Somewhat correct
Wrong

} **attributes**

Comparison Results

This is a **Hooded Merganser** *because ...*

Definition:	this bird has a **black crown** a **white eye** and a **large black bill**.
Description:	this bird has a **brown crown** a **white breast** and a **large wingspan**.
Explanation-Label:	this bird has a **black and white head** with a large **long yellow bill** and **brown tarsus and feet.**
Explanation-Dis.:	this is a **brown bird** with a **white breast** and a **white head**.
Explanation:	this bird has a **black and white head** with a **large black beak**.

Correct
Somewhat correct
Wrong

} **attributes**

Thanks for your attention

References:

[1] *Generating Visual Explanations, Hendricks et al*

[2] LONG SHORT-TERM MEMORY, Sepp Hochreiter, Jürgen Schmidhuber

[3] Christoper Olah, colah.github.io