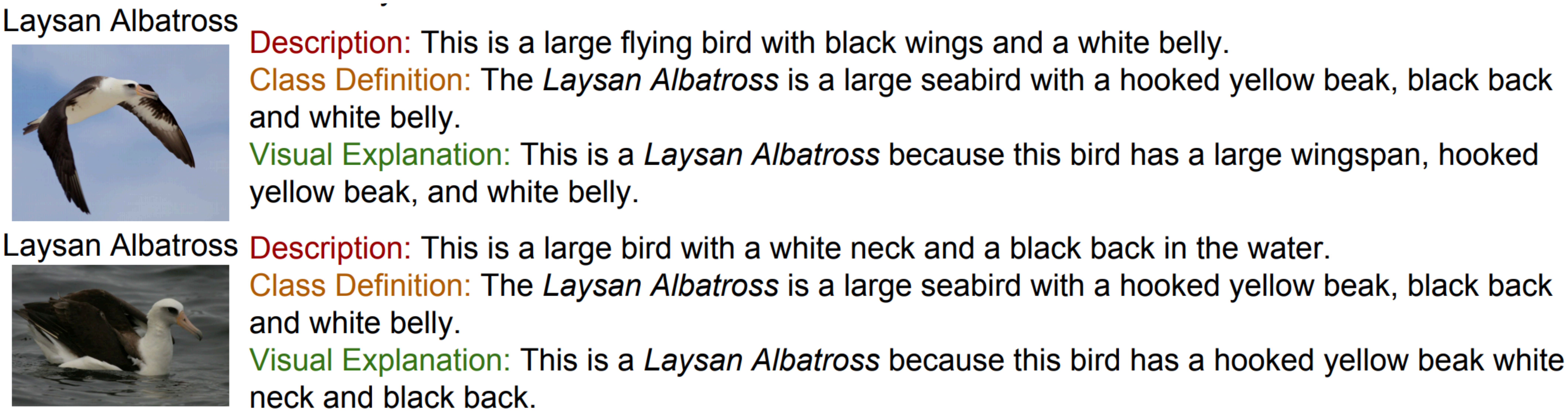
Text-based Explanations

- Can we generate a natural language explanation of a model's behavior?
- Possible advantages:
  - Easy for untrained users to understand
  - Easy for annotators to provide ground truth human explanations (which may also help our models)
- Possible disadvantages:
  - Hard to generate grammatical/semantically meaningful text
  - Can text truly explain a model's behavior?

Explanations of Bird Classification



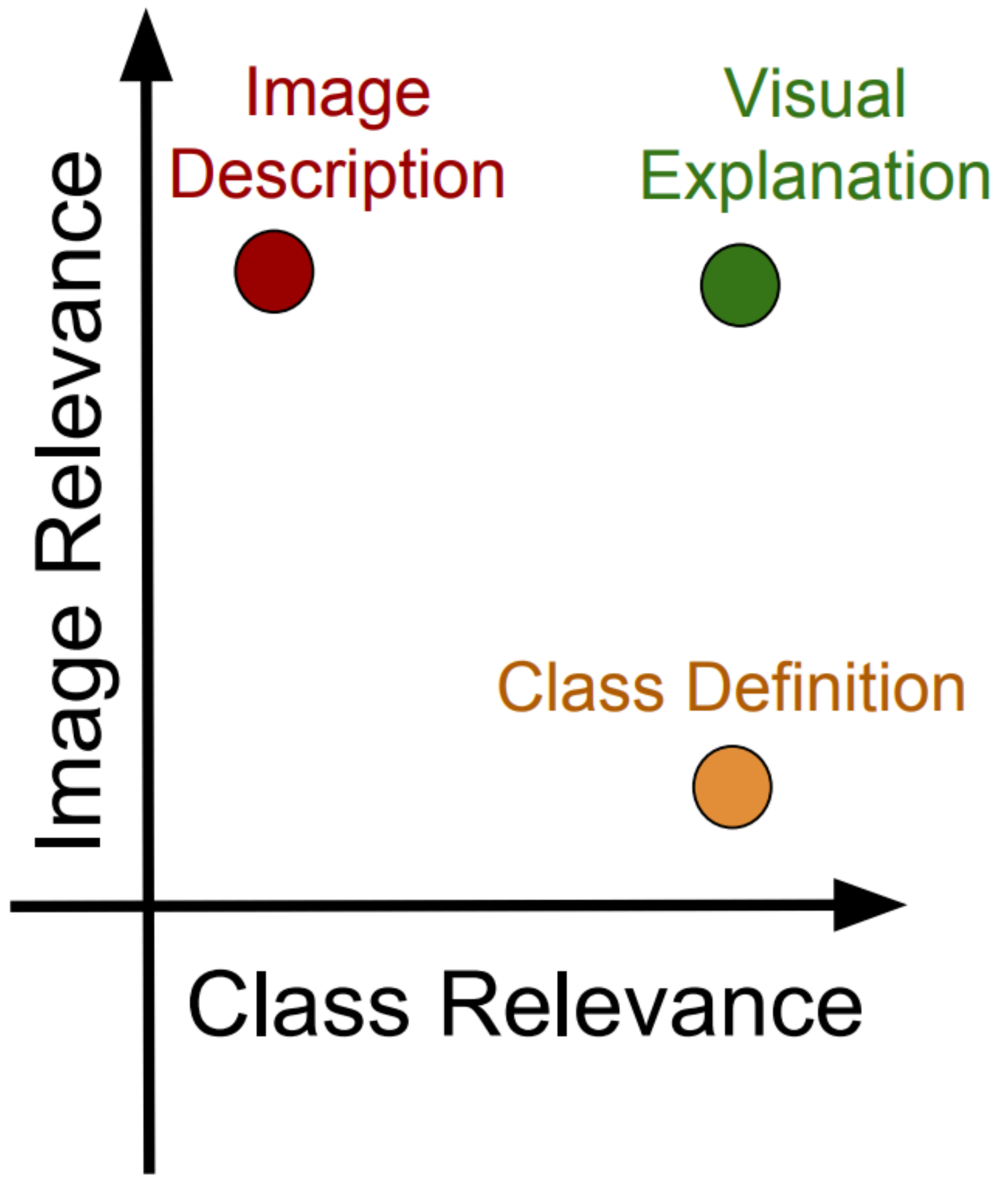
Laysan Albatross

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

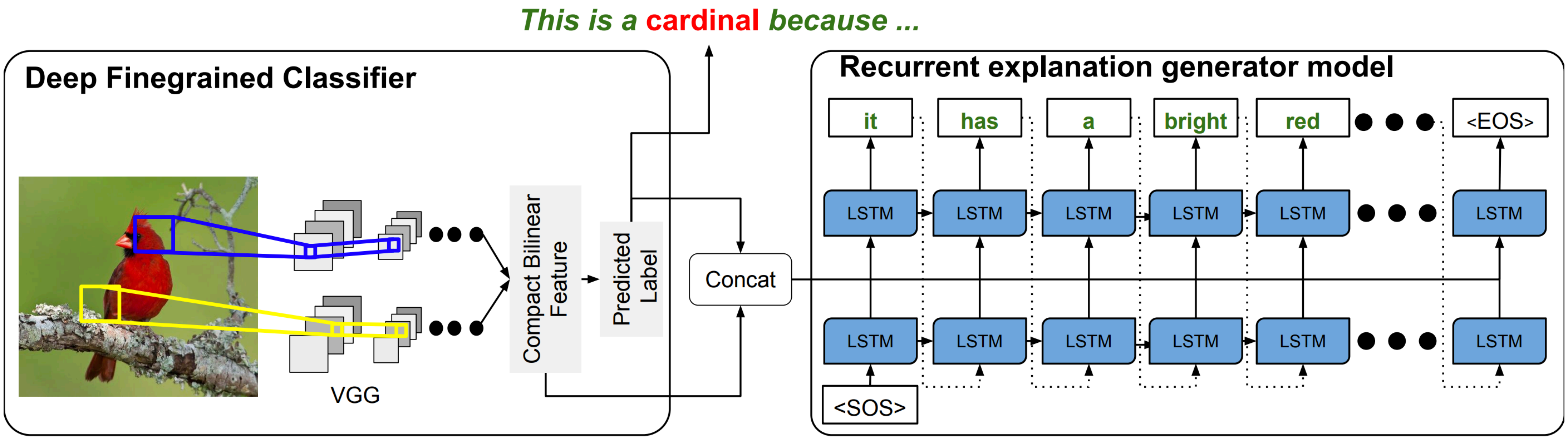
Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck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

- What makes a visual explanation? Should be relevant to the class and the image
- Are these features *really* what the model used?



Hendricks et al. (2016)

Explanations of Bird Classification

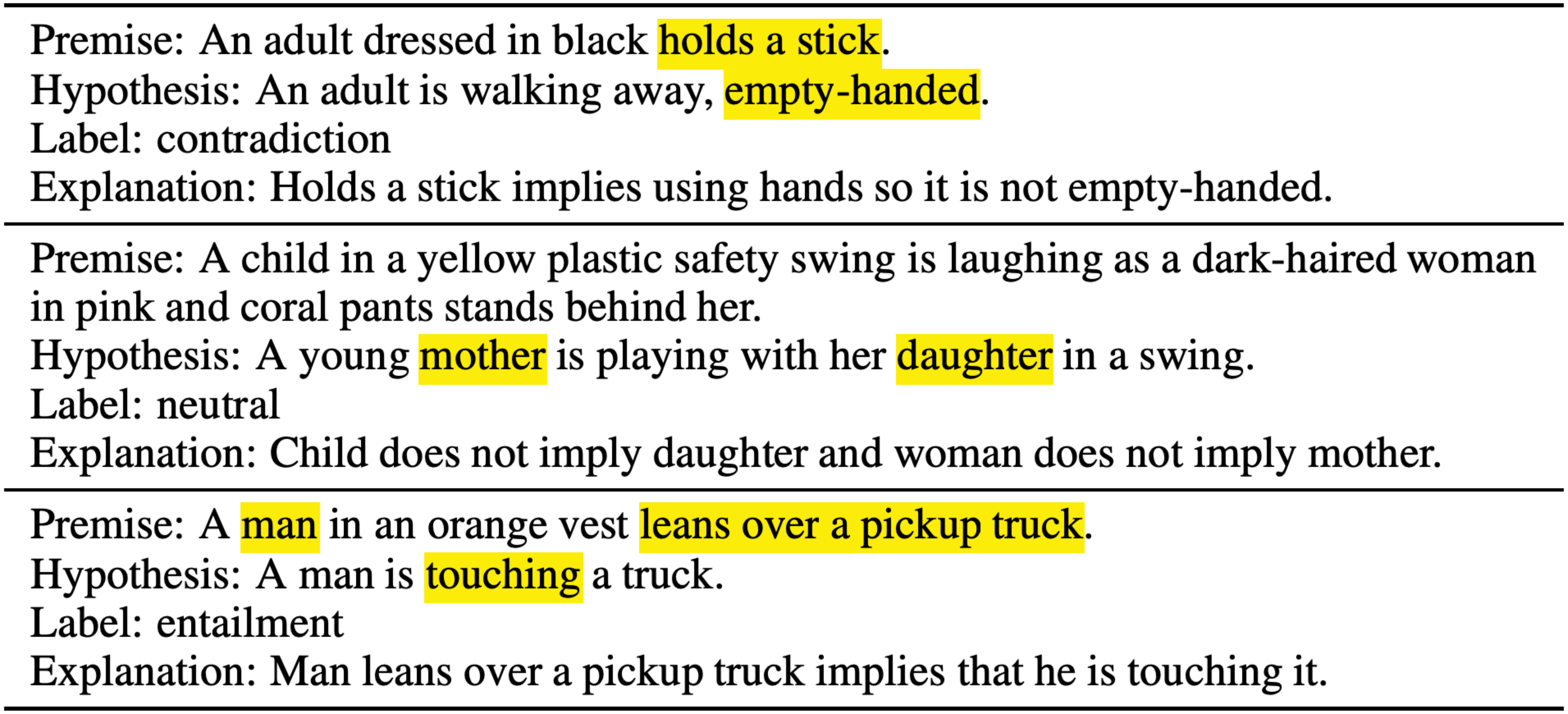

This is a cardinal because ...
Deep Finegrained Classifier
VGG
Compact Bilinear Feature
Predicted Label
Concat
Recurrent explanation generator model
it
has
a
bright
red
<EOS>
LSTM
<SOS>

- Are these features *really* what the model used? The decoder looks at the image, but what it reports may not truly reflect the model's decision-making

- More likely to produce plausible (look good to humans) but unfaithful explanations!

Hendricks et al. (2016)

e-SNLI



Premise: An adult dressed in black holds a stick.
Hypothesis: An adult is walking away, empty-handed.
Label: contradiction
Explanation: Holds a stick implies using hands so it is not empty-handed.

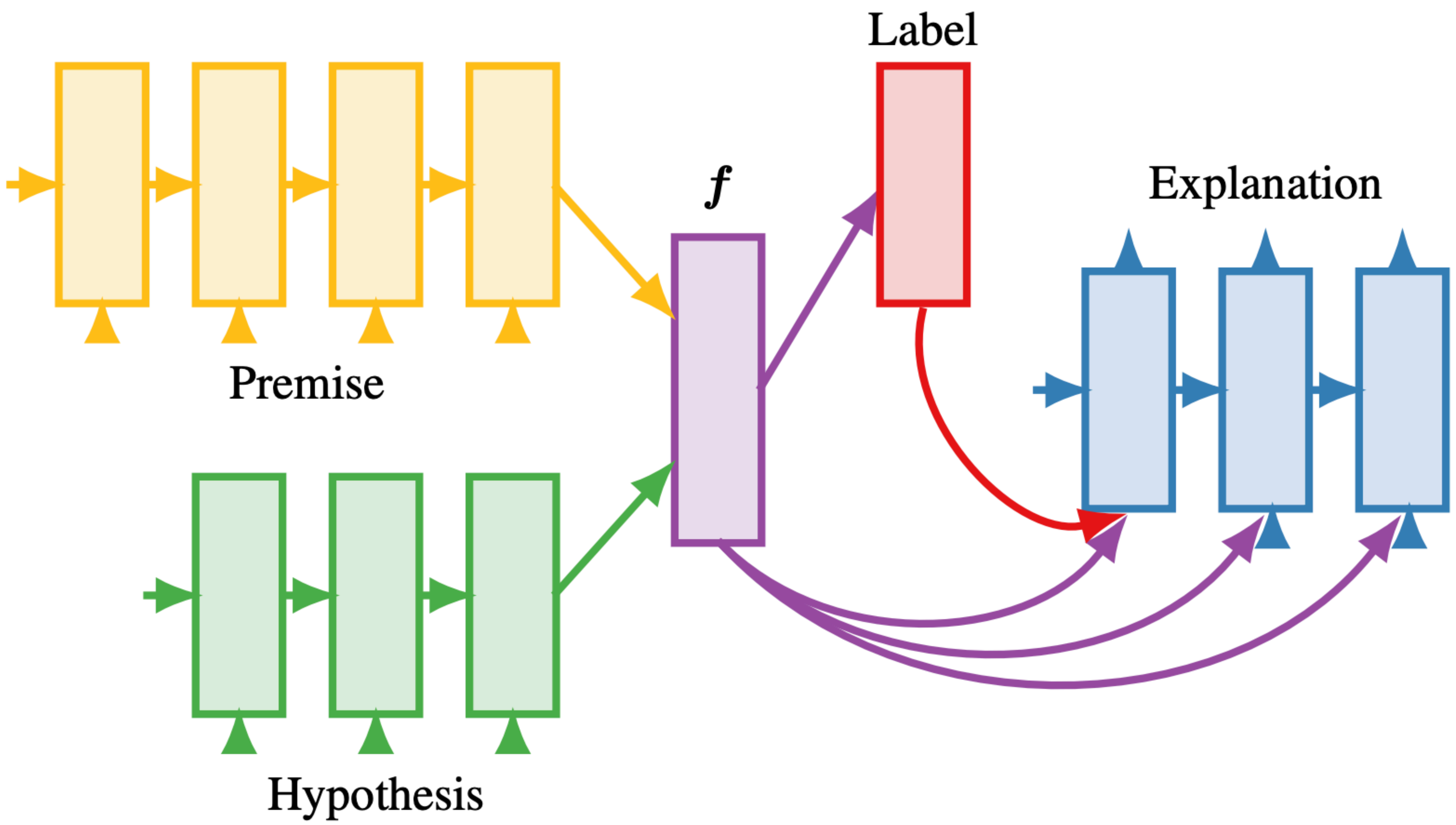
Premise: A child in a yellow plastic safety swing is laughing as a dark-haired woman in pink and coral pants stands behind her.
Hypothesis: A young mother is playing with her daughter in a swing.
Label: neutral
Explanation: Child does not imply daughter and woman does not imply mother.

Premise: A man in an orange vest leans over a pickup truck.
Hypothesis: A man is touching a truck.
Label: entailment
Explanation: Man leans over a pickup truck implies that he is touching it.

- e-SNLI: natural language inference with explanations

Camburu et al. (2019)

e-SNLI

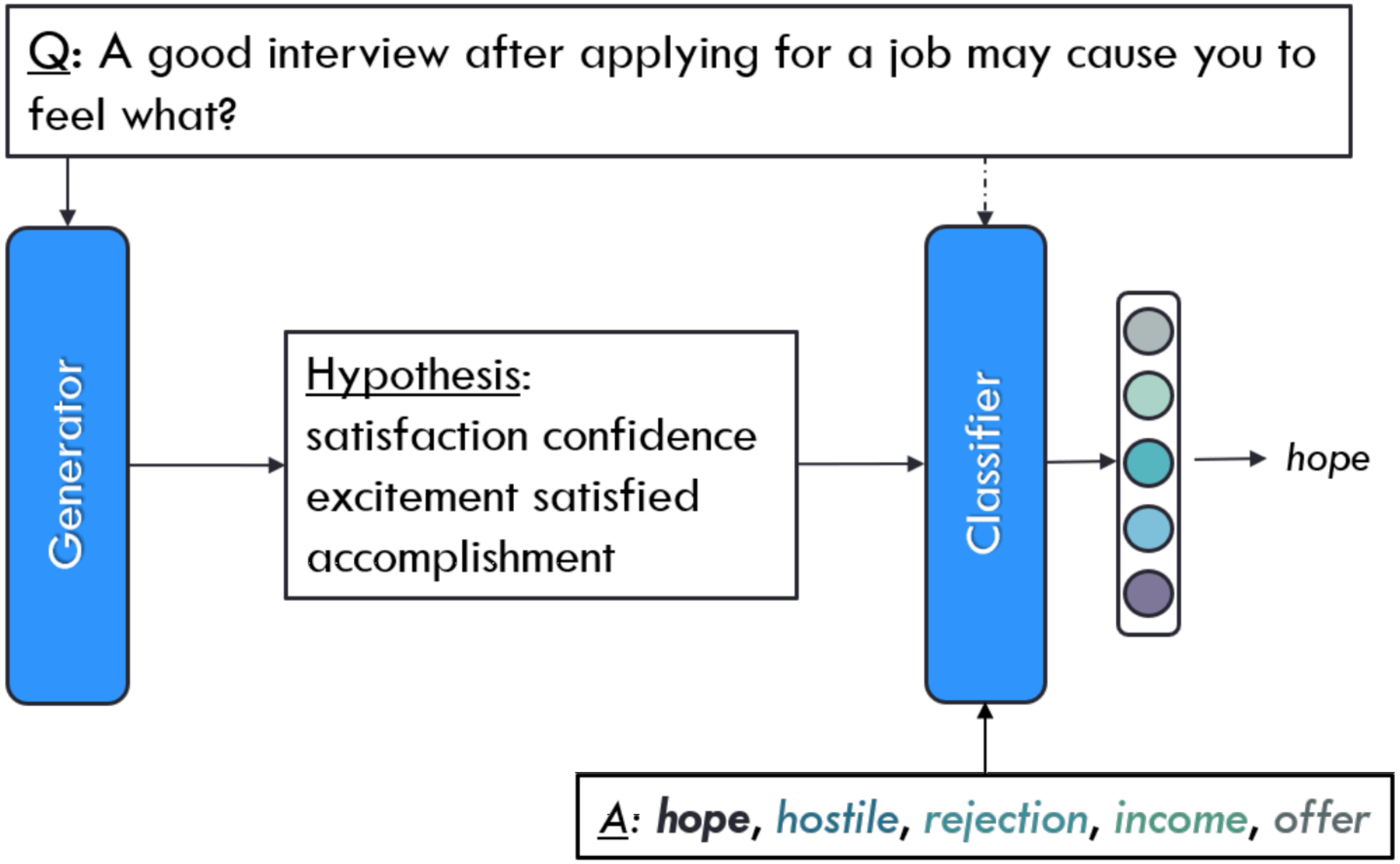


f = function of premise and hypothesis vectors

- Similar to birds: explanation is generated conditioned on the label and the network state *f*
- Information from *f* is fed into the explanation LSTM, but **no constraint that this must be used**. Explanation might be purely generated from the label

Camburu et al. (2019)

Latent Textual Explanations


Q: A good interview after applying for a job may cause you to feel what?
Generator
Hypothesis: satisfaction confidence excitement satisfied accomplishment
Classifier
hope
A: hope, hostile, rejection, income, offer

- Model generates text "hypothesis", which is completely latent
- Hypothesis isn't constrained to be natural language, ends up being keywords

Latcinnik and Berant (2020)