InterpNET: Neural Introspection for Interpretable Deep Learning (1710.09511)

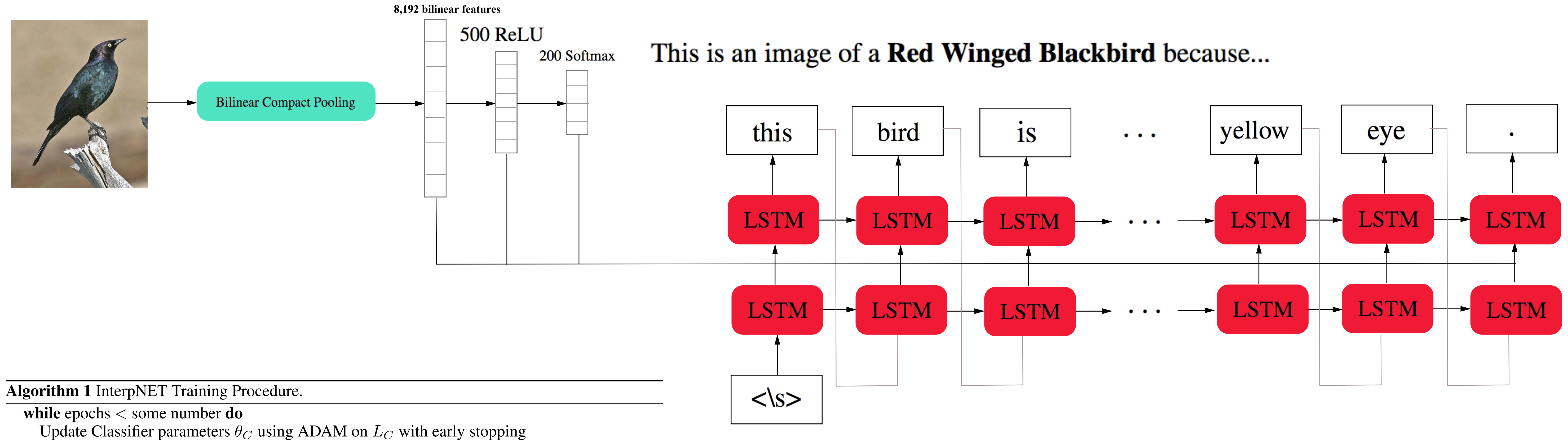
Shane Barratt

Department of Electrical Engineering, Stanford University

Abstract

Humans are able to explain their reasoning. On the contrary, deep neural networks are not. This paper attempts to bridge this gap by introducing a new way to design interpretable neural networks for classification, inspired by physiological evidence of the human visual system's inner-workings. This paper proposes a neural network design paradigm, termed InterpNET, which can be combined with any existing classification architecture to generate natural language explanations of the classifications. The success of the module relies on the assumption that the network's computation and reasoning is represented in its internal layer activations. While in principle InterpNET could be applied to any existing classification architecture, it is evaluated via an image classification and explanation task. Experiments on a CUB bird classification and explanation dataset show qualitatively and quantitatively that the model is able to generate high-quality explanations. While the current state-of-the-art METEOR score on this dataset is 29.2, InterpNET achieves a much higher METEOR score of 37.9. Source code is available online[2].

Architecture



```
Algorithm 1 InterpNET Training Procedure.
  while epochs $<$ some number do
    Update Classifier parameters $\theta_C$ using ADAM on $L_C$ with early stopping
  end while
  while epochs $<$ some number do
    Update Explainer parameters $\theta_D$ using ADAM on $L_E$ with early stopping
  end while
```

Main Idea

In principle, the layers of a neural network compute higher and higher-level representations of the parts of the input which are relevant to producing a class label [3]. Therefore, it is reasonable to assume that the relevant aspects to classification are contained in the internal activations of the network. For example, a single ReLU hidden-layer neural network computes the function

$$y = \text{softmax}(W_1 \text{relu}(W_2 x + b_2) + b_1)$$

and has internal activations f_1, f_2, f_3:

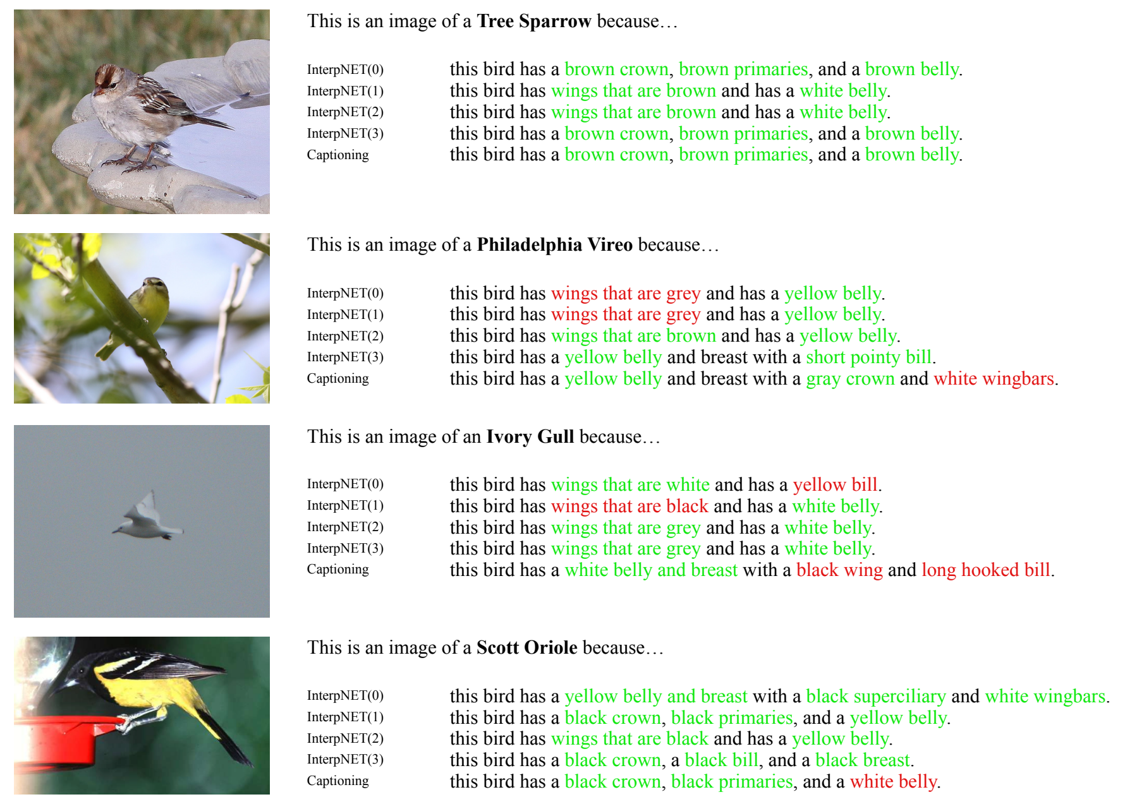
$$f_1 = x$$
$$f_2 = \text{relu}(W_2 x + b_2)$$
$$f_3 = \text{softmax}(W_1 \text{relu}(W_2 x + b_2) + b_1)$$

Building on this idea, the computation/reasoning of the network can be viewed as the internal activations of the network concatenated into a single feature vector, $r(x) = [f_1, f_2, f_3]$. Then, InterpNET uses this feature vector as input to a language-generating network and trains the language-generator in a supervised fashion to generate explanations $E(x, y)$. The next few sections go through the technical details to make this idea concrete on the problem of fine-grained bird classification.

Experimental Results

Table 1: Results. Explanation metrics and Classification Accuracy for a variety of models. InterpNET$_2$ achieves the highest metrics, except for classification accuracy. Higher is better for all metrics.

	METEOR	BLEU	CIDer	Classification Accuracy
InterpNET$_0$ (output only)	35.0	55.6	68.3	81.3%
InterpNET$_1$ (1 hidden layer)	36.1	58.7	73.5	**81.5%**
InterpNET$_2$ (2 hidden layers)	**37.9**	**62.3**	**82.1**	79%
InterpNET$_3$ (3 hidden layers)	31.7	47.3	54.3	76.7%
Captioning (input only)	32.2	48.2	55.5	81.3%
Generating Visual Explanations (baseline)	29.2	n/a	56.7	n/a



This is an image of a **Tree Sparrow** because…

InterpNET(0)	this bird has a brown crown, brown primaries, and a brown belly.
InterpNET(1)	this bird has wings that are brown and has a white belly.
InterpNET(2)	this bird has wings that are brown and has a white belly.
InterpNET(3)	this bird has a brown crown, brown primaries, and a brown belly.
Captioning	this bird has a brown crown, brown primaries, and a brown belly.

This is an image of a **Philadelphia Vireo** because…

InterpNET(0)	this bird has wings that are grey and has a yellow belly.
InterpNET(1)	this bird has wings that are grey and has a yellow belly.
InterpNET(2)	this bird has wings that are brown and has a yellow belly.
InterpNET(3)	this bird has a yellow belly and breast with a short pointy bill.
Captioning	this bird has a yellow belly and breast with a gray crown and white wingbars.

This is an image of an **Ivory Gull** because…

InterpNET(0)	this bird has wings that are white and has a yellow bill.
InterpNET(1)	this bird has wings that are black and has a white belly.
InterpNET(2)	this bird has wings that are grey and has a white belly.
InterpNET(3)	this bird has wings that are grey and has a white belly.
Captioning	this bird has a white belly and breast with a black wing and long hooked bill.

This is an image of a **Scott Oriole** because…

InterpNET(0)	this bird has a yellow belly and breast with a black superciliary and white wingbars.
InterpNET(1)	this bird has a black crown, black primaries, and a yellow belly.
InterpNET(2)	this bird has wings that are black and has a yellow belly.
InterpNET(3)	this bird has a black crown, a black bill, and a black breast.
Captioning	this bird has a black crown, black primaries, and a white belly.

This paper introduces a general neural network module which can be combined with any existing classification architecture to generate natural language explanations of the network's classifications provided one has supervised explanation data. InterpNET's classifications are highly accurate and interpretable at the same time as demonstrated by quantitative and qualitative analysis of experiments on a bird classification+explanation dataset. InterpNET achieves a METEOR score of 37.9 on the CUB test set, making it state-of-the-art in the visual explanation task. The model is able to use the information extracted from a trained classifier to produce excellent explanations and is a sizable step towards interpretable deep neural network models.