Learning Explainable Embeddings for Deep Networks

Zhongang Qi, Fuxin Li
School of Electrical Engineering and Computer Science
Oregon State University
qiz@oregonstate.edu, lif@eecs.oregonstate.edu

Abstract

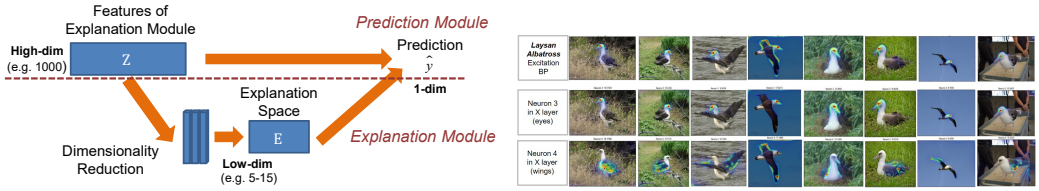
We propose a novel explanation module to explain the predictions made by deep learning. Explanation module works by embedding a high-dimensional deep network layer nonlinearly into a low-dimensional explanation space while retaining faithfulness, so that the original deep learning predictions can be constructed from the few concepts extracted by the explanation module. We then visualize such concepts for human to learn about the high-level concepts that deep learning is using to make decisions. We propose Sparse Reconstruction Autoencoder (SRAE) for learning the embedding to the explanation space. SRAE aims to reconstruct part of the original feature space while retaining faithfulness. The proposed method is applied to explain CNN models in image classification tasks, and several novel metrics are introduced to evaluate the performance of explanations quantitatively without human involvement. Experiments show that the proposed approach could generate better explanations of the mechanisms CNN to use for making predictions in the task.

1 Introduction

With all its incredible advances, the usage of deep learning in real applications still must overcome a trust barrier. Imagine scenarios with a doctor facing a deep learning prediction: this CT image indicates malignant cancer, or a pilot facing a prediction: make an emergency landing immediately. These predictions may be backed up with a claimed high accuracy on benchmarks, but it is human nature not to trust them unless we are *convinced* that they are reasonable for each individual case. The lack of trust is worsened because of known cases where adversarial examples can fool deep learning to output wrong answers [8, 3]. In order to establish trust, human needs to understand how deep learning makes decisions. Such understanding could also help the human to gain additional insights into new problems, potentially improve deep learning algorithms, and improve human-machine collaboration.

People like explanations of the form “A is something because of B, C, and D", e.g. this is a bird because it has feathers, wings and a beak. This type of explanation is concise – there are not a hundred different reasons that add up to explain that A is something. Besides, it relies on high-level concepts B, C, and D. Both are often at odds with deep learning predictions, which are combinations of outputs from thousands of neurons in dozens of layers. Approaches have been proposed to visualize each of the filters [10] and for humans to name them [1], but it is difficult for this approach to obtain a concise representation. On the other hand, many other approaches generate attention maps that backtrack a decision to specific important areas in the original image [7, 2, 12, 11]. These are often nice and quite informative, but they work on individual images and do not provide any high-level concept that can be broadly applicable to many images simultaneously, nor can we believe they are complete explanations so that we can trust them.

In this paper, we make an attempt to reconcile these explanation approaches by extracting several high-level concepts from deep networks to aid human understanding. Our model attaches a separate explanation network to a certain layer in the deep network to reduce the network to a few human-understandable concepts, from where one can generate predictions similar to the original deep network



(a) (b)

Figure 1: (a) The explanation module is a dimensionality reduction mechanism so that the original prediction $\hat{y}$ can be reproduced from this low-dimensional space. Here we focus on 1-dimensional outputs. A multi-class understanding in principle can be built up from separate understandings of one-against-all classifiers; (b) Comparison of ExcitationBP and x-features.

(Fig. 1(a)). We focus on making those concepts to be *faithful*, that the deep learning predictions can be faithfully approximated from those few concepts; and *local*, that the concepts are relatively spatially localized in images so that the human can understand them.

2 Model Formulation

We propose to learn an explanation module (Fig. 1(a)), a module that can be attached to any layer of a deep network. It attempts to learn an embedding that lowers the dimensionality of an intermediate layer and directly learn a mapping from such an explanation layer to mimic the output of the original deep learning network (DNN).

We denote the input feature space of explanation module as $\mathbf{Z}(\mathbf{x}; \mathbf{W})$, where $\mathbf{x}$ and $\mathbf{W}$ are the input features and parameters (from multiple layers) of the original DNN model, respectively, and $\mathbf{Z}$ represents the output of a particular intermediate layer in the network. The explanation module is used to embed $\mathbf{Z}$ to an explanation space, denoted as $\mathbf{E}(\mathbf{Z}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents parameters of the embedding that need to be learned. As a shorthand, we will also refer to the explanation space as an *x-layer*, and each dimension in the x-layer as an *x-feature*. Note that in the explanation, we do not attempt to change the parameters $\mathbf{W}, \mathbf{b}$ of the original DNN model. The explanation module can in principle be attached to any layer in the DNN, although the closer to the prediction, the higher level the concepts are and it becomes easier to mimic the prediction of DNN with a low-dimensional embedding. In this paper, we focus on attaching explanation modules to fully-connected layers.

Here we propose a novel network called Sparse Reconstruction Autoencoder (SRAE), which handles the objective as defined in (1). SRAE is also a neural network, hence can seamlessly combine with the original prediction DNN, making the following visualization process simple. The encoding layer in SRAE forms the explanation space $\mathbf{E}$. We utilize the encoding layer to mimic the predictions, and also to reconstruct part of the features in $\mathbf{Z}$ which are responsible for the prediction target.

$$\min_{\boldsymbol{\theta},\widetilde{\boldsymbol{\theta}},\mathbf{v}} \frac{1}{M}\sum_{i=1}^{M}\left\|\mathbf{v}^\top \mathbf{E}(\mathbf{Z}^{(i)};\boldsymbol{\theta}) - \hat{y}^{(i)}\right\|^2 + \frac{\beta}{S_z}\sum_{k=1}^{S_z}\log(1+q\cdot\frac{1}{M}\sum_{i=1}^{M}\left\|\phi^{-1}\left(\mathbf{E}(\mathbf{Z}^{(i)};\boldsymbol{\theta});\widetilde{\boldsymbol{\theta}}\right)_k - Z_k^{(i)}\right\|^2) \quad (1)$$

There are 2 terms in optimization (1), which are faithfulness loss and sparse reconstruction loss. The first item attempts to be faithful to the original DNN. The prediction result of SRAE $\bar{y}^{(i)}$ is $\mathbf{v}^\top \mathbf{E}(\mathbf{Z}^{(i)}; \boldsymbol{\theta})$, which is optimized to be similar to the original prediction $\hat{y}^{(i)}$ of DNN. Here $\boldsymbol{\theta}$ is the parameter for encoder; $\mathbf{v}$ is the parameter for prediction; M is the number of the training examples.

The second item of the optimization attempts to prevent degeneracy where $\hat{y}^{(i)}$ itself can be used as x-feature in the explanation space. Here S_z is the dimensionality of the original feature layer $\mathbf{Z}$; k is one dimension from $\mathbf{Z}$; $q > 0$ is a sparsity parameter for the log penalty [4]; ϕ^{-1} is the reconstruction mapping which maps the explanation space $\mathbf{E}$ back to $\mathbf{Z}$; $\widetilde{\boldsymbol{\theta}}$ is the parameter for reconstruction (decoder); $Z_k^{(i)}$ and $\phi^{-1}\left(\mathbf{E}(\mathbf{Z}^{(i)};\boldsymbol{\theta});\widetilde{\boldsymbol{\theta}}\right)_k$ are the k-th element of $\mathbf{Z}^{(i)}$ and $\phi^{-1}\left(\mathbf{E}(\mathbf{Z}^{(i)};\boldsymbol{\theta});\widetilde{\boldsymbol{\theta}}\right)$, respectively; β is the parameter for sparse reconstruction loss. The quadratic loss between $\phi^{-1}\left(\mathbf{E}(\mathbf{Z}^{(i)};\boldsymbol{\theta});\widetilde{\boldsymbol{\theta}}\right)_k$ and $Z_k^{(i)}$ measures the capability of reconstructing the k-th dimension in the space of $\mathbf{Z}$. With the sparsity term, optimization (1) enables the explanation space to only reconstruct features that are responsible for the prediction target, and refrain from reconstructing other irrelevant dimensions, which improves the interpretability of the Explanation Space.

Note that SRAE is different from conventional sparse autoencoder where the autoencoder activations in the hidden layers are constrained to be sparse. In SRAE, the sparsity constraint is on the amount of

Method		SRAE	NN	SAE	Z	ExcitationBP
F_{reg}	Training	0.0831	0.0657	0.0987	—	—
	Testing	0.1539	0.1170	0.1928	—	—
Locality		**1.9694**	2.3078	2.1492	**1.9623**	2.4934

Table 1: The Faithfulness and Locality for methods in 30 categories selected randomly. The column **Z** represents the average locality computed over all the dimensions of **Z**, the 4096-dimensional first fully-connected layer of the deep network. This is obtained by separately running ExcitationBP on each dimension of **Z** and evaluating the resulting heatmaps.

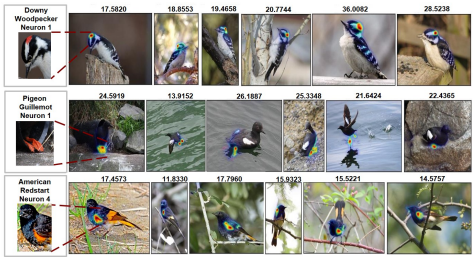


Figure 2: The most important x-feature for several categories. The weight above the feature is $v_n E_n$.

input dimensions to be reconstructed. In general, various sparsity functions can be used here such as the L_1 penalty, epsilon-L_1 penalty [4], the Kullback-Leibler divergence [5], etc. Here we choose the log penalty in our proposed model because it is differentiable, does not have singularities, and has a nice gradient form that is easy to solve with backpropagation. SRAE can be applied as a general method to the domains where input feature selection and feature coding are both needed.

3 Experiments

We utilize the CUB-200-2011 dataset [9] in the experiments. This is a task for fine-grained bird classification into 200 categories. The most challenging part in the experiments is to find objective metrics to evaluate the performance of the explanation module, since the explanation of images is a relatively subjective matter. Given image I_m, for each neuron n in the x-layer and each pixel (i,j) in I_m, we denote $S_{i,j}^{n,m} \triangleq P(\text{Pixel}_{i,j}^m|\text{Neuron}_n) = \frac{C_{i,j}^{n,m}}{\sum_{(i,j)\in I} C_{i,j}^{n,m}}$, where $C_{i,j}^{n,m}$ is the contrastive marginal winning probability (c-MWP) generated by ExcitationBP [11] for pixel (i,j) in I_m with neuron n in x-layer, (i,j) is the coordinate of the pixel. For the CUB dataset, since the given part label of each image is just one pixel in the middle of the part, and there is no extra information about the shape and the size of the part regions, we utilize the Voronoi diagram to partition the bounding box into 15 regions in which the nearest neighbor part annotation in each region would be the same, then compute the probability $S_p^{n,m} \triangleq P(\text{Part}_p^m|\text{Neuron}_n) = \sum_{(i,j)\in I_m} P(\text{Part}_p^m|\text{Pixel}_{i,j}^m)P(\text{Pixel}_{i,j}^m|\text{Neuron}_n)$.

We propose a couple of metrics to evaluate the performance of the explanation module: (1) **Faithfulness**: $F_{reg} = \frac{1}{M}\sum_m L(\bar{y}^{(m)} - \hat{y}^{(m)}) = \frac{1}{M}\sum_m |\bar{y}^{(m)} - \hat{y}^{(m)}|$, the mean absolute loss between $\hat{y}^{(m)}$ and its approximation $\bar{y}^{(m)}$; (2) **Locality**: For each x-feature n we have a p-dimensional histogram $\mathbf{S}_n$ whose element is $\bar{S}_p^n = \frac{1}{M}\sum_m S_p^{n,m}$. The locality for each x-feature is defined as the entropy: $H_n = -\sum_p \left(\frac{\bar{S}_p^n}{\sum_p \bar{S}_p^n} \cdot \log(\frac{\bar{S}_p^n}{\sum_p \bar{S}_p^n})\right)$.

The fine-tuned VGG19 model [6] for CUB-200-2011 birds is used as the prediction DNN to be explained. The explanation network is a 3 middle layers SRAE. We trained an explanation module on a random 30 of the 200-dimensional outputs of the DNN. For each category, we utilized 50 positive examples and $8,000$ negative examples as the training data; the remaining positive examples $(8-10)$ and $2,000$ negative examples as the testing data. The number of the x-features is set to 5, as our experiments showed that more x-features do not improve performance in this dataset and create x-features which have 0 weight in approximating $\hat{y}$, indicating that one one-against-all classifier of one bird does not depend on many high-level visual features. We compared the proposed SRAE with a fully-connected neural network (NN), a conventional stacked autoencoder (SAE), as well as directly performing ExcitationBP on the classification output $\hat{y}$ (ExcitationBP).

In Table 1, we summarize the results for different explanation embedding approaches with different parameters. Results show that we can achieve excellent faithfulness to the prediction. The F_{reg} in both training and testing are less than 0.2. Since $\hat{y}$ before softmax usually has a range in $[0, 50]$ and especially large in the positive examples, we consider the regression loss to be small. Our algorithm showed significant improvements over NN, SAE and ExcitationBP in terms of locality, indicating that we are capable of separating information that comes from different parts. The average locality of the x-features generated by SRAE are almost matching the average locality of features in **Z**. This means we are close to the limit of part separation on this layer: many of the features on the **Z** layer already represent multiple parts. In future work we plan to conduct more experiments explaining

earlier convolutional layers to see whether the locality could be further lowered while preserving faithfulness. We also show some qualitative examples from different categories in Fig.2 and Fig.1(b). Fig.2 shows the most important x-feature in several categories, where we can see that they fit our intuitions on the discriminative features of the birds. Fig.1(b) compares x-features with directly running ExcitationBP on $\hat{y}$. One can see x-features nicely separate different discriminative aspects of the bird while ExcitationBP sometimes focuses only on one part and miss others, and sometimes produces a heatmap that incorporates many parts simultaneously. Also, each x-feature seems distinct enough as a concept. Hence we believe they indeed provide more explanation on the decisions made by CNN algorithms.

4 Conclusion

In this paper we propose an explanation module, that can be attached to any layer in a deep network to compress the layer into several concepts which can approximate a 1-dimensional prediction output from the network. A sparse reconstruction autoencoder (SRAE) is proposed to avoid degeneracy and improve orthogonality. We also propose automatic evaluation metrics to evaluate the explanation on a fine-grained bird classification dataset. Quantitative and qualitative results show that the network can indeed extract high-level concepts from a CNN that make sense to human. We view this work as one of the first steps toward understanding deep learning and have many future plans to it, including performing more experiments on different kinds of data, including those without ground truth, and extending it to explain other types of neural networks, such as recurrent networks and convolutional-recurrent ones.

Acknowledgments

This work is partially supported by the Defense Advanced Research Projects Agency (DARPA) under Contract N66001-17-2-4030.

References

[1] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. *Advances in neural information processing systems*, 19:801, 2007.

[5] Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[6] Marcel Simon and Erik Rodner. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *International Conference on Computer Vision (ICCV)*, 2015.

[7] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

[8] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[9] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[11] Jianming Zhang, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer, 2016.

[12] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016.