Toward Interpretable Image Recognition

A mini-review

Amir Rahnama

KTH Royal Institute of Technology

March 17, 2020

Interpretability

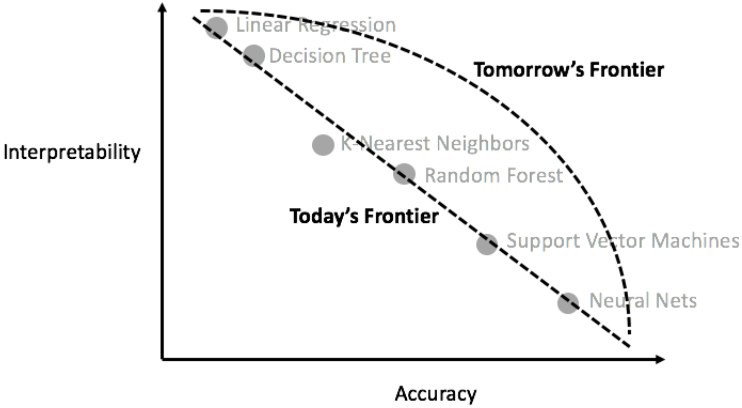
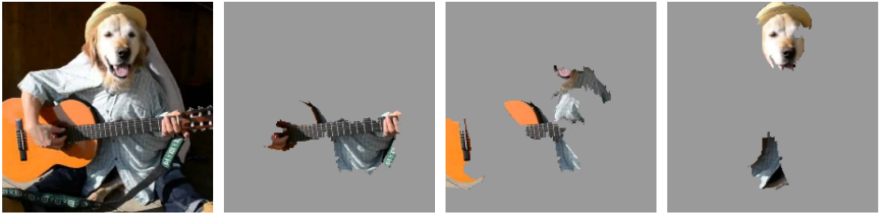


Figure: Interpretability vs. Accuracy

Post-hoc interpretability



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure: Predictions of an original image with Google's Inception V3

Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains its Predictions

Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin

AAAI-18

Problems with Post-hoc Interpretability

- Explanations are based on the explanation model used
- Explanations are separate models that are trained separately and cannot really be trusted as representing the model we are trying to explain

How can we make vision models interpretable?

- Models should classifying on the basis of visual feature prototypes
- Prediction of images should be at each of the levels in a taxonomy and not based on dataset labels
- Ability to detect never-seen-before subclasses within a taxonomy

Proposed Method

- Authors propose a *Prototype classifier* in which observations are classified based on their proximity to a prototype
- Methods like nearest centroid classifier or nearest prototype classifier that perform something similar
- Prototype theory is a mode of graded categorization in cognitive science, where some members of a conceptual category are more central than others.
- In this theory, any given concept in any given language has a real world example that best represents this concept.

Model Architecture

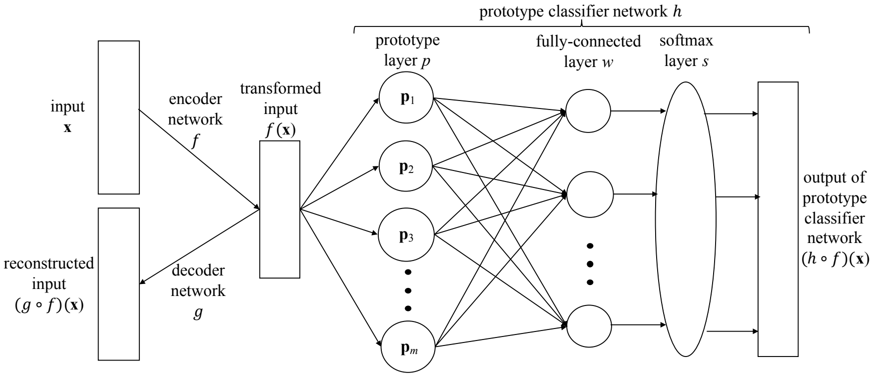


Figure: Network Architecture

Cost Function

$$
\begin{aligned}
L((f, g, h), D) = E(h \circ f, D) + \lambda R(g \circ f, D) \\
+\lambda_1 R_1(p_1, ..., p_m, D) \\
+\lambda_2 R_2(p_1, ..., p_m, D)
\end{aligned} \tag{1}
$$

Transposed weight matrix

	0	1	2	3	4	5	6	7	8	9
	-0.07	7.77	1.81	0.66	4.01	2.08	3.11	4.10	-20.45	-2.34
	2.84	3.29	1.16	1.80	-1.05	4.36	4.40	-0.71	0.97	-18.10
	-25.66	4.32	-0.23	6.16	1.60	0.94	1.82	1.56	3.98	-1.77
	-1.22	1.64	3.64	4.04	0.82	0.16	2.44	-22.36	4.04	1.78
	2.72	-0.27	-0.49	-12.00	2.25	-3.14	2.49	3.96	5.72	-1.62
	-5.52	1.42	2.36	1.48	0.16	0.43	-11.12	2.41	1.43	1.25
	4.77	2.02	2.21	-13.64	3.52	-1.32	3.01	0.18	-0.56	-1.49
	0.52	-24.16	2.15	2.63	-0.09	2.25	0.71	0.59	3.06	2.00
	0.56	-1.28	1.83	-0.53	-0.98	-0.97	-10.56	4.27	1.35	4.04
	-0.18	1.68	0.88	2.60	-0.11	-3.29	-11.20	2.76	0.52	0.75
	5.98	0.64	4.77	-1.43	3.13	-17.53	1.17	1.08	-2.27	0.78
	1.53	-5.63	-8.78	0.10	1.56	3.08	0.43	-0.36	1.69	3.49
	1.71	1.49	-13.31	-0.69	-0.38	4.55	1.72	1.59	3.18	2.19
	5.06	-0.03	0.96	4.35	-21.75	4.25	1.42	-1.27	1.64	0.78
	-1.31	-0.62	-2.69	0.96	2.36	2.83	2.76	-4.82	-4.14	4.95

Figure: Transposed weight matrix (every entry rounded off to 2 decimal places) between the prototype layer and the softmaxlayer.

Prototype Examples with R1 and R2

Figure: Prototype Examples with R1 and R2

Prototype Examples with R1 Only

Figure: Prototype Examples with R1 Only

Prototype Examples with R2 Only

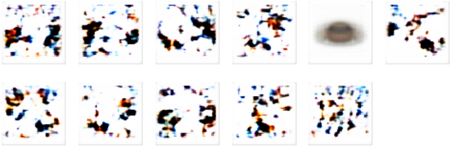


Figure: Prototype Examples with R2 Only

Prototype Examples with neither R1 or R2

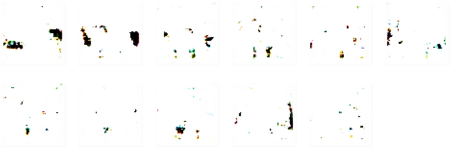


Figure: Prototype Examples with neither R1 and R2

Discussion: Deep Learning for Case-based reasoning

- The authors have fomrulated the problem and their solutions very well
- The experiments are very thorough and convincing
- The paper lacks proper details in case you are interested in implementation
- The paper argues against the "interpretability-accuracy" trade-off which is a common misconception

Interpretable Image Recognition with Hierarchical Prototypes

Peter Hase, Chaofan Chen, Oscar Li, Cynthia Rudin

HCOMP-19

How can we make vision models interpretable?

- Models should classifying on the basis of visual feature prototypes
- Prediction of images should be at each of the levels in a taxonomy and not based on dataset labels
- Ability to detect never-seen-before subclasses within a taxonomy

Why taxonomies?

- We organize the world into "inductively rich" categories that relate to each other
- Humans "explain" their visual judgment by pointing to prototypes with regards to a class, e.g. a certain animal is a tiger because it is a large cat with black stripes
- These explanations vary across each level and has multiple layer of abstractions

Benefits of Taxonomy

- Make the trade-off between information gain and accuracy explicit
  - Useful when policy repsponses do not change after a level of specifity, e.g. when a model cannot tell whether an image is of a rifle or a pistol, but still can detect that it is of the type gun
- Explanations will be taylored to a specific taxonomical level
  - Within the taxonomy of cars, a siren installed on a van can classify an ambulance, however a siren on a car can classify a police car
- In cases of never-seen-before taxonomical types, model can still show the broader class of the taxonomy (novel class detection)

Related works

- Using class attention maps: identify subsections of an image that are important to the classification Video: How Class attention maps work
- Feed only a portion of the image that is selected in a supervised or unsupersived way

Saliency maps

Figure: Saliency maps show where the model is looking, but they dont tell why a model classifies an image

Hierarchical Classifications

- Hierarchical classification have been explored with models such as:
  - SVMs
  - Bayesian Graphical Models
  - CNNs
  - CNN + RNN
- Usually the problem is solved in supervised way, however in some studies infering the tree has been done in an unsupervised manner as well.
- Many studies construct predictions using only one CNN, however in other studies branch their network to find representations for each sub-classification task.

How this work is different?

- While this work is using the idea of branching the network to solve many sub-classification tasks, but it is different in the sense that it uses prototypes in the latent space.
- The other approaches are using hierarchical class labelings
- Other works are using prototypes for Graphical Bayesian Models but in the pixel space

Ensure Interpretability

- Features corresponding to object properties (prototype in work of Bloom et al. , 2017)
- To produce measures of similarity between new instances and representatives of each class (Exemplar in the work of Bloom et al. (2017) ✓

Hierarchical Classification

- Predict an images class at each level of taxonomy tree
- each y_i has k elements and $y_i^{(k)}$ is the image's label at the kth level at the tree
- $y^{(0)}$ is the root and $y^{(1)}$ is the first node after the root and so on
- Not all branches need to be at the same legth, K

Hierachical Classification

- We learn a function f that approximates $P(Y|X)$ over paths in the tree in full labels {animal, dog} with impossible paths as being 0, like {animal, car}.
- We learn each of the factors of the probability:

$$P(Y|X) = P(Y^{(1)}|X) \times ... \times P(Y^{(K)}|Y^{(K-1)}, X) \tag{2}$$

- Each distribution represents the multinomial distribution over children classes $Y^{(K)}$

Novel Class Detection

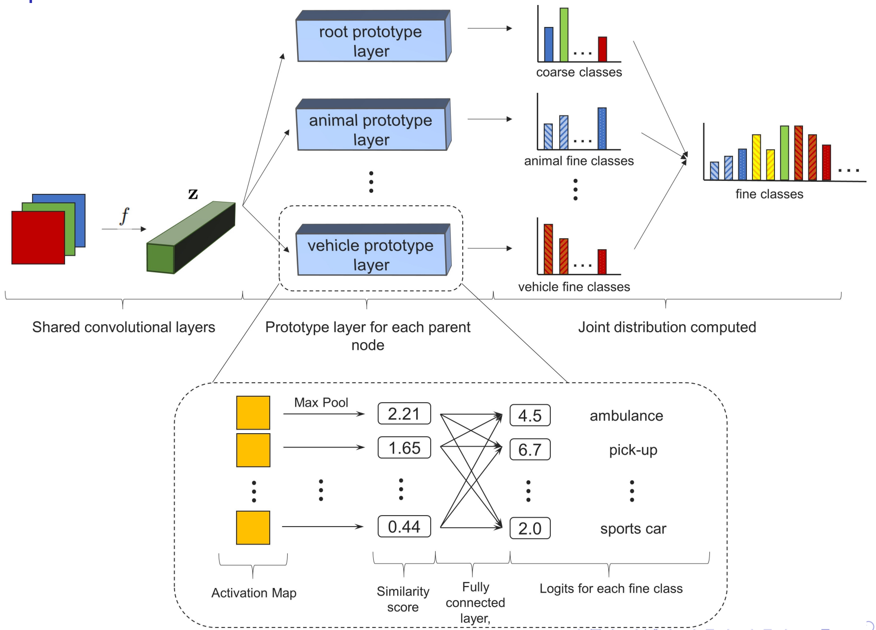
- How do we detect images from an unseen class?
- if one is willing and ableto set aside some data that is novel, while considering the remaining data to come from the known distribution
- Finding completely new classes at roots, or findings new classes in the children nodes
- Standard out-of-distribution detection is to estimate the probability:

$$P(Y^* \in Y^K | X) \tag{3}$$

- This is how it is defined in this study:

$$P(Y^* \in c^{(k)}_{\text{children}} | Y^* \in c^{(k)}, X) \tag{4}$$

Proposed Model



Proposed Model

- This model is an extension of Chen et al (2018) model
- VGG-16 (just the encoder network) model maps the images to the input space, namely $\tilde{z}$
- For each parent node, ther is a prototype layer that operates directly on $\tilde{z}$ and produces a similarity score where $\tilde{z}$
- During training, m prototypes are learned by mapping each instance and selecting for which prorotype the activation map is in its maximum:
- Before training, authors set pre-determined number of prototypes evenly to child class of $P^{c^{(k)}}$

Objective Function

$$\sum_{c^{(k)} \in C} [\sum_{c^{(k)}} \text{Cross Entropy}(h^{c^{(k)}} \circ g_{P^{c^{(k)}}} \circ f(x_i), y_i)$$
$$+\lambda_1 \text{Clust}(P^{c^{(k)}}, X, Y) + \lambda_2 \text{Sep}(P^{c^{(k)}}, X, Y) + \lambda_3 \text{Reg}(h^{c^{(k)}}) \tag{5}$$

$$\text{Clust}(P^{c^{(k)}}, X, Y) =$$
$$\sum_{i: y_i^{(k)}} c^{(k)} \min_{j: P_j \in P_{c_i^{(k+1)}}} \min_{\tilde{z} \in \text{patches}(f(x_i))} ||\tilde{z} - p_j||_2^2 \tag{6}$$

$$\text{Sep}(P^{c^{(k)}}, X, Y) =$$
$$-\sum_{i: y_i^{(k)}} c^{(k)} \min_{j: P_j \notin P_{c_i^{(k+1)}}} \min_{\tilde{z} \in \text{patches}(f(x_i))} ||\tilde{z} - p_j||_2^2 \tag{7}$$

Objective Function

- Cross entropy handles the accuracy of predictions over joint distribution of fine-grained data
- Clustering cost encourages the model to map at least one path vector of each image close to a prototype corresponding to its class
- Separation costs discourages the mapping of patches to different classes
- Regularization terms is both an L1 regularization term (to nullify the weights of the mapping between an image with different classes) while L2 regularization handles the weights for within-class weights

Experiments

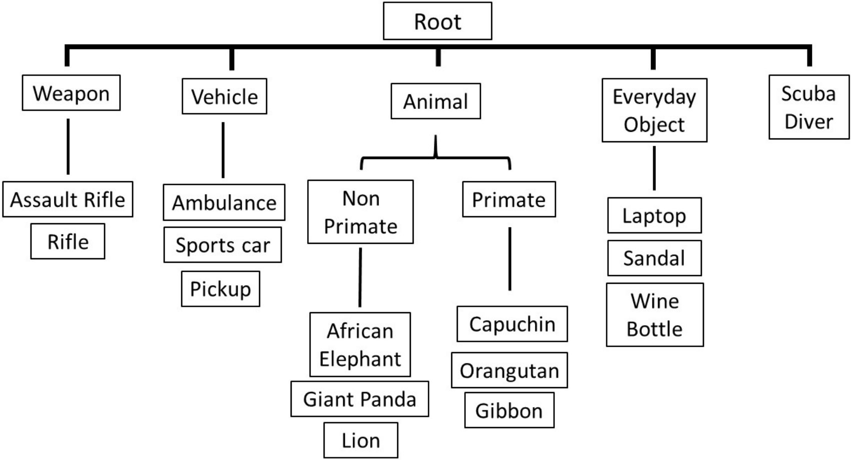


Figure: Training classes

Experiments

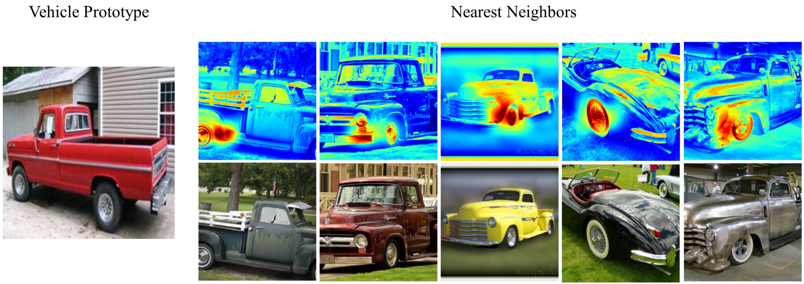
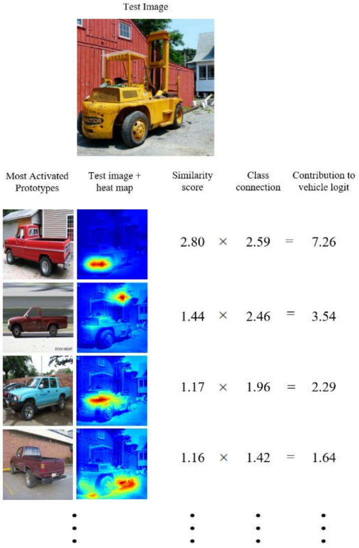


Figure: Similar images to vehicle prototypes

Experiments



$$P(c^{(1)} = \text{vehicle}|\mathbf{x}) = .999999$$

$$P(\mathbf{y}^* \notin \text{vehicles}|c^{(1)} = \text{vehicle}, \mathbf{x}) = .76$$

Discussion: Interpretable Image Recognition with Hierarchical Prototypes

- I found it extremely suprising that the paper "This looks like that" was completely a beginner version of this paper, however it was published after the latter paper
- The fact that the number of classes was very limited was indeed problematic
- The training of the model was set in a very ad-hoc manner without proper details
- It is intersting to see that we can somehow gain interpretable features with even black-boxes if we aim for it