

Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, Benjamin I. P. Rubinstein

School of Computing and Information Systems
The University of Melbourne, Victoria, Australia

{ruihanz, pmathugama}@student.unimelb.edu.au, {tmiller,kehinger,benjamin.rubinstein}@unimelb.edu.au

Abstract

Convolutional neural network (CNN) models for computer vision are powerful but lack explainability in their most basic form. This deficiency remains a key challenge when applying CNNs in important domains. Recent work on explanations through feature importance of approximate linear models has moved from input-level features (pixels or segments) to features from mid-layer feature maps in the form of concept activation vectors (CAVs). CAVs contain concept-level information and could be learned via clustering. In this work, we rethink the ACE algorithm of Ghorbani et al., proposing an alternative invertible concept-based explanation (ICE) framework to overcome its shortcomings. Based on the requirements of fidelity (approximate models to target models) and interpretability (being meaningful to people), we design measurements and evaluate a range of matrix factorization methods with our framework. We find that *non-negative concept activation vectors* (NCAVs) from non-negative matrix factorization provide superior performance in interpretability and fidelity based on computational and human subject experiments. Our framework provides both local and global concept-level explanations for pre-trained CNN models.

Introduction

Deep learners such as convolutional neural networks (CNNs) (He et al. 2016) are widely used across important domains like computer vision due to demonstrated performance in numerous tasks. However, when applying to critical domains like medicine, justice, and finance, explainability has become a key enabler and mitigation for applications. While commentators like Rudin (2019) argue that deep learning approaches should not be used for these risky domains, using deep learning to discover features for more ‘interpretable’ models requires explainability to determine what features have been discovered.

Recent CNN explanation methods attempt to quantify the importance of each feature. Feature importance usually corresponds to a linear approximation of highly complex models. Linear models are simple and explainable with understandable features and weights. Methods like CAM (Zhou et al. 2016), LIME (Ribeiro, Singh, and Guestrin 2016) and saliency maps (Bach et al. 2015) use input-level features.

However, these methods only point out the important areas in an image, rather than, for example, identifying key concepts. To make explanations more human understandable, more recent work uses features from within the CNN models with higher-level information (Kim et al. 2018; Bau et al. 2017; Zhou et al. 2018) under the assumption that these are closer to human-relevant concepts than areas of pixels. This work uses supervised learning to classify feature maps with concepts and use weights to indicate the directions of certain concepts called Concept Activation Vectors (CAVs) in the feature map space. To reduce the labeling workload, Automated Concept-based Explanations (ACE) (Ghorbani et al. 2019) apply unsupervised learning to segments of images from certain classes and generate clusters of image segments as CAVs.

ACE is powerful for learning features (CAVs) using only images from target classes without any labels but has several drawbacks. First, learned concept weights are inconsistent for different instances. This is a common problem for all linear approximation explanation methods as they are all approximations to the target CNN models. Second, it is difficult to measure the performance of learned sets of CAVs. ACE is a one-way explainer from the target model to explanations and irrelevant segments are discarded. Transforming an explanation back to the target model’s prediction is difficult with segments. However, how much information is lost in the explanation? Information could be lost in unused segments and distances between used segments and their cluster centroids.

Olah et al. (2018) introduces an interpretable interface with many methods to display what’s inside the CNN model for a given image to achieve interpretability. Feature maps can be shown using different grouping methods (axes). They also introduce a method using non-negative matrix factorization (NMF) on feature maps to gather interpretable information. This shares a similar idea with ACE as the clustering methods like k -means used in ACE could be considered as a matrix factorization (dimensionality reduction) method. In addition to reducing the dimensionality, matrix factorization also analyzes the information loss with inverse functions. The feature maps from segments used in ACE are replaced by spatial activations from the feature map. This replacement is also used in CoCoX (Akula, Wang, and Zhu 2020).

In this paper, we modify the ACE framework using ma-

trix factorization for feature maps instead of clustering segments’ feature maps as in ACE to overcome its shortcomings. Targeting the last feature map of CNN models, our framework could provide consistent CAVs weights for different instances (in most CNN models). Through the inverse function of matrix factorization, the information lost (fidelity, differences in prediction) in the explanations can be measured. Given performance requirements (e.g. accuracy of approximations), explainers (hyper-parameters) could then be learned automatically. Also, inverse functions allow us to analyze the distributions of contributions from learned CAVs to provide detailed local explanations for given instances. ACE only provides global explanations for classes. Having this framework, we compare three different methods k -means, PCA and NMF. Sample local and global explanations are shown in Figure 1. Our contributions are:

- We propose a new concept-based explanation framework that provides consistent weights for features (under some limitations) and consistent fidelity measurement using non-negative matrix factorization.
- We propose new measurements of fidelity for CAV-based interpretability methods. This measures how accurate the explanations are to the original CNN model. Hyper-parameters can be learned through this measurement.
- We propose a new human subject experiment design for measuring the interpretability of concept-based interpretability methods. Unlike previous work, we measure interpretability in a scientific manner using human studies, rather than relying on “intuition” and selected examples to illustrate interpretability (Leavitt and Morcos 2020).

A Framework for Concept-based Explanations

Concept-based explanations may be approached as linear approximations for separate CNN models as follows. First, we separate the CNN classifier into a concept extractor and classifier from a single CNN layer, and the explanations are based on the feature maps from that layer. Input-level explanations like LIME (Ribeiro, Singh, and Guestrin 2016) and saliency maps (Shrikumar, Greenside, and Kundaje 2017; Smilkov et al. 2017) skip this step and use input pixels as features for approximate models directly. Second, we apply matrix factorization to the feature maps to provide CAVs for the next step. A reducer is trained with a target concept-related dataset. Note that the middle-layer feature maps may contain too many dimensions, and information in each dimension is not enough to be meaningful. Therefore, the reducer may gather information separated in all dimensions to provide CAVs and reduce the complexity of the approximate model. For the final step, we build a linear approximation to the classifier and estimate the concept importance for each CAV. The explanation is based on the learned reducer and estimated weights for each CAV. For explanations of new inputs, reducers provide meaningful concept descriptions and concept scores from the feature maps. A diagram of the framework is shown in Figure 2.

The selection of the target layer is important. Higher layers focus more on concepts (high-level features) and lower layers focus more on edges and textures of the image (low-level features) (Zeiler and Fergus 2014). If the reduced concepts are to be meaningful, the selection of a higher layer seems more suitable. Higher layers usually mean a classifier with fewer layers (simpler) and a concept extractor with higher-level concepts. This provides more accurate estimated weights. One special case is when using feature maps from the last layers when they are a Global Average Pooling (GAP) layer and a dense layer, the classifier under explanation will reduce to a simple linear model. Estimated weights are accurate as they are constant at any position under any CAVs. The previous layers’ weight estimates take the average weights of all instances. Weights could vary for different inputs. In this paper, we assume that the last layer is the most suitable for concept-based explanations, so we use this as the target layer.

A benefit of using reducers instead of clustering methods is that reducers provide scores for concepts as outputs instead of predictions of clusters’ centroids. The reduced concept scores can be applied to the approximate model to analyze the contribution distribution for each feature more accurately. In ACE, concept scores can only be binary. Thus, reducers provide better fidelity when inverting the reduction process. This could help when evaluating the fidelity of the learned CAVs and also when applying the concepts in a larger explainability framework.

Ribeiro, Singh, and Guestrin (2016) claim two important requirements for linear approximations: interpretability and fidelity. **Interpretability** means that the feature representation used in the approximate model needs to be meaningful to human observers. **Fidelity** prescribes that the approximate model should make similar predictions to the model under explanation. This clearly introduces a dual-objective problem: increasing fidelity can decrease interpretability, or vice-versa. In this paper, we evaluate different matrix factorization methods on these two measures.

Methodology

Given a pre-trained CNN classifier F with n training images I , the prediction process will be $F(I) = Y$. We remove any final softmax layer (if present) so that each y is a scalar instead of a probability. Let A be the feature map from the target layer l , then F can be separated into two parts, feature extractor $E_l(I) = A_l$ and classifier $C_l(A_l) = Y$. Feature map A should be of shape $n \times h \times w \times c$ where h and w reflect the feature map size and c is the number of channels. A is assumed to be non-negative as most recent CNN models use the *relu* activation. Let $a^{(i,j)}$ be a vector from A at position (i, j) , $\{0 \leq i < h, 0 \leq j < w\}$. CNN models share weights, so vector a at each position in A could be considered as a vector description of the original images but with different correlated receptive fields after equivalent processing.

Non-Negative Concept Activation Vectors Feature map A can be flattened to $V \in R^{(n \times h \times w) \times c}$. Non-negative matrix factorization (NMF) reduces the channel dimensions of non-negative matrix V from c to c' . Here V is reduced to

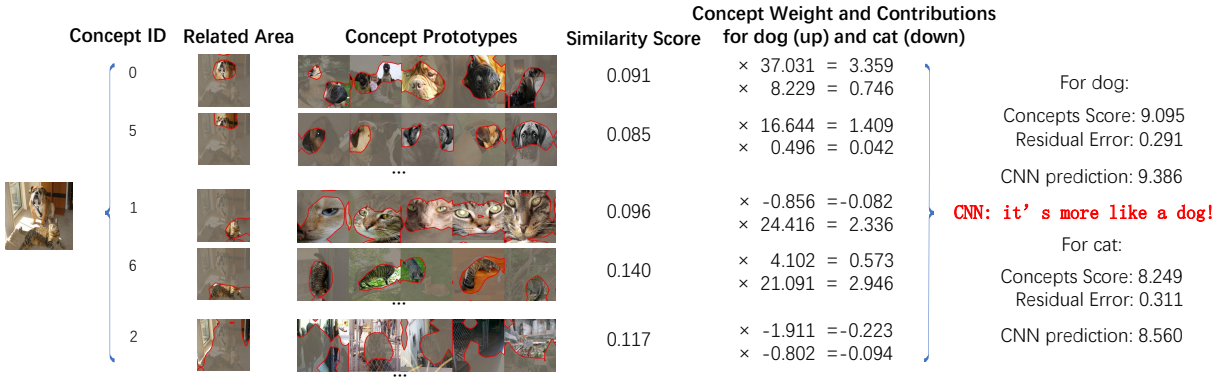


Figure 1: Explanation for an image with a dog and cat. For each concept, the framework provides prototypes based on the training set and correlates areas as global explanations. The explainer decomposes the final prediction score to concept scores and weights through a linear model to explain locally. The explanation is based on the ResNet50 CNN model from TorchVision. The full version for this explanations is shown in the appendix

feature score $S \in R^{(n \times h \times w) \times c'}$ and feature direction $P \in R^{c' \times c}$ as $V = SP + U$. The aim is to minimize the residual error U . It is given formally as $\min_{S,P} \|V - SP\|_F$ s.t. $S \geq 0, P \geq 0$. Having images with the same label or concept, most vectors a with different correlated receptive fields in A can be considered as a vector description of a target concept part (e.g. eye, mouth or ear of a dog). Factorization on these vectors can disentangle important and frequently-appearing directions for target concepts.

P , the meaningful NCAVs in the feature map dimension space, is a fixed parameter for the explainer after being trained with A from some images. Given the explainer, for feature maps from new images, we can apply NMF with P to get S . P is a vector basis and S is the length of projections of a on these directions. S can be considered as feature scores for feature directions in P as it's the degree to which vector a is related (similar) to the NCAVs in P .

Weight Estimation For weights or concept importance, other interpretability methods such as saliency maps use the gradients of output scores for some classes with respect to individual input features. Feature importance of classifier C can be estimated using the method in TCAV (Kim et al. 2018). Given a learned NCAV p_l in target layer l , the estimated feature importance for a classifier targeting class k for given feature map A_l is $\frac{\partial C_{l,k}}{\partial p_l} = \frac{1}{n \times h \times w} \sum_{a \in A_l} \lim_{\epsilon \rightarrow 0} \frac{h_{l,k}(a + \epsilon p_l) - h_{l,k}(a - \epsilon p_l)}{2\epsilon}$. Classifier C is usually a non-linear function. Estimated weight is an average derivative over the area sampled by A , and A should come from images related to the target concept to reduce estimate error.

Only a linear classifier has a consistent derivative over the whole input space, so the best choice for target layer is the last layer before a GAP layer and a dense layer. Having weights $W \in R^c$ with bias b from the last dense layer for target class k and learned NCAV parameter P , the classifier

C will be:

$$C_{k,l}(A) = GAP(A)W + b \\ = GAP(S)PW + GAP(U)W + b$$

The feature importance for NCAV P will be PW . This is independent of the input feature map A .

Vector Visualization There are many ways to visualize a vector from a layer. For instance, having a vector in a middle layer, Deep Dream concept vector visualization (Olah, Mordvintsev, and Schubert 2017) provides a pattern-enhanced image based on gradients for the vector. In this work, we use the method of prototypes (Kim, Khanna, and Koyejo 2016), choosing images containing target concepts and highlighting these concepts. Applying GAP to the decomposed feature maps, we can provide a score for each concept. Images with high concept scores are taken as the prototypes. Previous work shows that middle-layer feature maps have spatial correlations with input images, as was used in image segmentation to replace input masks with feature map masks (Dai, He, and Sun 2015). Decomposed feature maps for a single CAV could be presented as heatmaps for target concepts. Combining a heatmap and an image, we can apply a threshold for the heatmap and highlight only areas with high concept value in the image. In this paper, the threshold is taken to be 0.5, and only regions with values higher than 0.5 (after a min-max normalization) are considered to be related. Concept prototypes from Figure 1 are visualized in this way.

Evaluation

Following the desiderata of this work, we aim to measure both fidelity and interpretability. Fidelity is measured computationally while for interpretability, we propose a new task that uses human subject experiments.

For both the computational and human subject experiments, we use well-known CNN models for image classification. We consider two different datasets: ILSVRC2012 (ImageNet) (Russakovsky et al. 2015) and CUB (Wah et al.

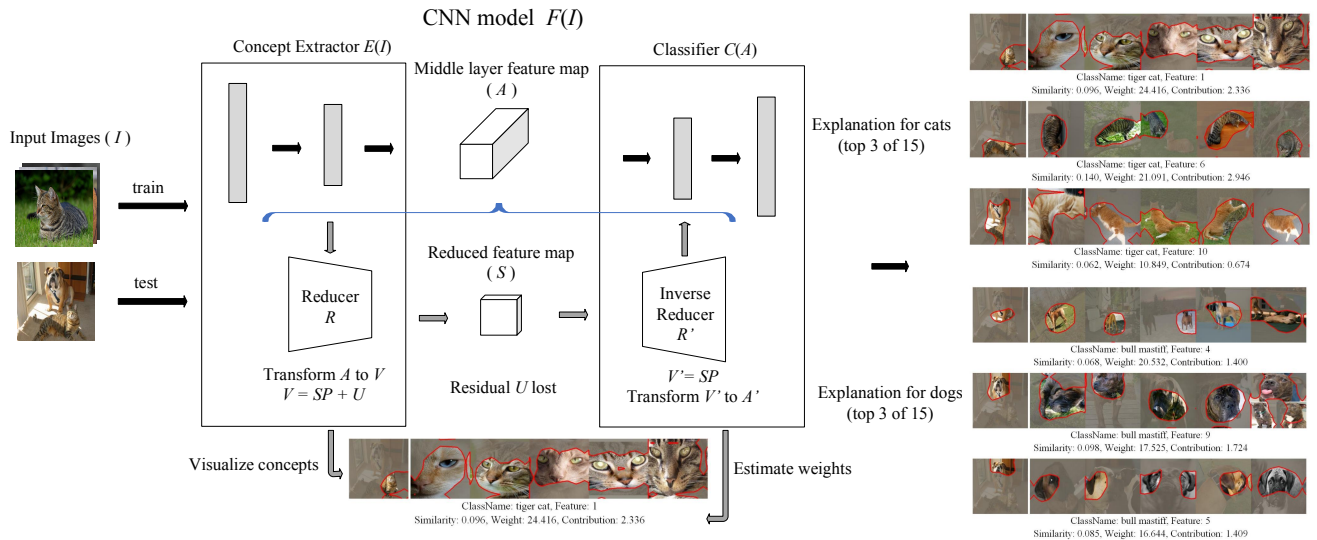


Figure 2: A depiction of our framework. The CNN model is separated into concept extractor and classifier by chosen middle layer with a reducer. The concept extractor provides concept presentations, instance correlated areas and feature scores. The classifier provides concept weights and generates linear approximations as an explanation.

2011). The implementation is based on PyTorch and scikit-learn. CNN models used for the ILSVRC2012 dataset are from torchvision pre-trained models. The top1 error of ResNet50 on ILSVRC2012 is 23.85% and that of Inception-V3 on ILSVRC2012 is 22.55%. For the CUB dataset, we use the ResNet50 (He et al. 2016) architecture and apply fine-tuning based on ImageNet pre-trained weights. The top1 error is 15.81%. For comparison with NMF, we choose the baseline of clustering (from ACE) and PCA (a popular matrix factorization method). Reducers are trained based on the training set and evaluated on the test set or the validation set.

Fidelity for Approximate Models

In this section, we compare the fidelity of approximate models using the three different matrix factorization methods. We evaluate the fidelity for CNN pre-trained models with different c' for each matrix factorization method using both classification and regression measurements.

For fidelity, we measure the difference between the approximate and original model predictions, but the measurements for classification and regression problems are different. Classification models only focus on predicting labels, so errors that do not change the predicted labels are ignored. For regression, any difference in approximate models will affect the performance based on the loss function measurement.

Measures Given the original model F and an approximate model \hat{F} , Craven and Shavlik (1996) measure the fidelity of approximate classification models as the 0-1 loss. This targets the difference in accuracy through predictions. For regression, Ribeiro, Singh, and Guestrin (2016) measure fidelity as the squared error $(F(X) - \hat{F}(X))^2$. While the squared error is appropriate as a loss function during training, for evaluation, relative error (RE) is more easily inter-

pretable, being scale-free. Given F , \hat{F} and a set of images I , the measurement for classification and regression models based on the dataset will be:

$$Fid_{c,F,\hat{F}}(I) = \frac{\#\{i \in I \mid F(i) = \hat{F}(i)\}}{\#\{I\}} \quad (1)$$

$$Fid_{r,F,\hat{F}}(I) = \frac{\sum_{i \in I} |F(i) - \hat{F}(i)|}{\sum_{i \in I} |F(i)| + \epsilon} \quad (2)$$

Having a trained reducer R and its inverse function R' for layer l , the approximate model is given by $\hat{F}_l(I) = C_l(R'(R(E_l(I))))$.

Experimental Setup Our experiment is based on a ResNet50 pre-trained model for ImageNet from torchvision and a ResNet50 pre-trained model for CUB. Both of them use the feature maps from *layer4*'s output. The parameter c' is evaluated from 5 to 50, in steps of 5. The model can be considered as both a classification and regression (score for every single class) model. So both fidelity measures can be evaluated. Here we trained reducers for all 1,000 classes in ILSVRC2012 and all 200 classes in CUB. Reducers are trained with images from a single class. For classification methods, only the top 5 classes are considered as candidates. For regression, only ground truth classes are tested, calculating the RE for the approximate models' outputs. We take the mean RE for all classes as the final result.

Experimental Results Figure 3 shows the accuracy (left) and fidelity for different c' with Fid_c (middle) and Fid_r (right). PCA provided the best fidelity result for both regression and classification. NMF's result is close to PCA's but diverges as c' increases. PCA is a popular and efficient matrix factorization method. Compared to PCA, NMF has two limitations: non-negativity and no introduction of extra bias. Also, NMF finds new vector bases to achieve a

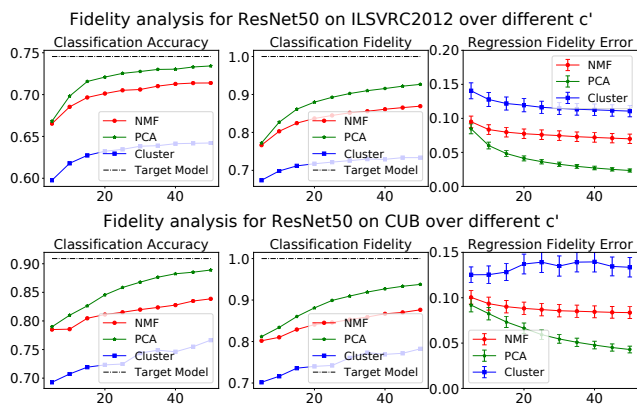


Figure 3: Average accuracy and fidelity for approximate linear models of two ResNet50 models over different number of concepts. Left figures show the accuracy if the approximate models are directly used as classifiers. For classification, higher means better. For regression, lower means better.

new balance for each vector every time c' increases, while PCA simply seeks a new basis vector iteratively, based on variance maximization. Clustering showed the worst performance. Clustering methods can be considered as matrix factorization methods, but they only provide one-hot vectors as centroid predictions offering the least information. When c' increases, approximate models provide more faithful predictions for both classification and regression. Around 1,000 compute (8 core CPU, v100 GPU) hours are needed for the evaluation.

The fidelity results demonstrate the limitation of approximate models: explanations will be incorrect some of the time. However, the classification accuracy on the left shows a potential trade-off for domains where accurate explanations are important: we can throw away the final layers of the target model and use the approximate model instead. This will result in a 2-5% drop in accuracy for NMF and PCA (30+ concepts), but the fidelity of the approximate model against itself is 100%.

Interpretability via Human Subject Experiments

In this section, we present a new type of experiment for measuring the interpretability of concept-based explanations, and use this to evaluate the interpretability of approximate models based on NMF, clustering, and PCA. We agree with the claims by Leavitt and Morcos (Leavitt and Morcos 2020) that interpretability should be measured in a falsifiable manner. In this paper, this includes measuring how well people can interpret and understand the concepts, rather than relying on “intuition” based on selected examples. These experiments provide a better basis for model comparison, compared to simply demonstrating interpretability via a handful of selected examples. To that end, we present a human-subject experiment design that measures the interpretability of concepts by asking study participants to provide free-text labels for concepts and measuring the consistency of labeling between participants.

We hypothesized that NCAVs learned from NMF are more interpretable than CAVs learned from clustering and PCA. We assume that CAVs learned from segments and feature maps using k -means are similar and consider CAVs from k -means clustering method on feature maps as ACE baselines, as both cluster feature maps and provide CAVs as outputs.

Methodology: We use *task prediction* (Hoffman et al. 2018, p. 11) and the Explanation Satisfaction Scale (Hoffman et al. 2018, p. 39) for evaluation. The assumption behind task prediction is that the ability to predict the output of a model implies a better understanding of how the model works.

Experimental Design: The experiment has two phases. In Phase 1, at each trial, participants are given one image with one concept highlighted and five concept explanations from the same class as candidates. Participants are asked to predict which of the candidate explanations matches the highlighted concept. An example is shown in Figure 4. Then, for each concept candidate, participants are asked to provide a 1-2 word description of the concept. All participants are given five images as a training phase followed by 15 testing images. They can move back to a training example at any time in the test phase. In Phase 2, participants are asked to complete an explanation quality survey to self-report their opinion about explanations in the form of Hoffman et al.’s explanation satisfaction scale. The experiment was implemented in a web-based environment on Amazon Mechanical Turk, a crowd-sourcing platform popular for obtaining data for human-subject experiments (Buhrmester, Kwang, and Gosling 2011).

We used a between-subject design: participants were randomly assigned into one of nine groups (3 scenarios and 3 types of reducers). There were a total of 157 participants who completed the survey. Participants with a prediction accuracy lower than 20% (random choice) were excluded from the results. Each experiment ran for approximately 30 minutes. We compensated each participant with \$5USD and an extra bonus of \$1USD for participants with high accuracy. 65% of participants were male, 34% were female and 1% specified their own gender. Participants’ reported ages were between 23 and 70 ($\mu = 38.9$).

Our measure for the task prediction is the percentage of correctly identified concept explanations. This assumes that better concept explainers allow participants to better match the highlighted concept to the correct explanation. For the 1-2 word descriptions, participants should have similar descriptions for the concepts that correspond to clear meanings. We use GloVe (Pennington, Socher, and Manning 2014) pre-trained word vector representations for each description, then use the average pairwise cosine similarity to measure the similarity of the concept descriptions across participants. We chose the free-text response task over providing pre-defined concept labels because using pre-defined labels could result in making concepts appear more interpretable than they are. If participants are uncertain, choosing the closest available label from a limited list makes the task about discrimination, rather than interpretation. Finally,

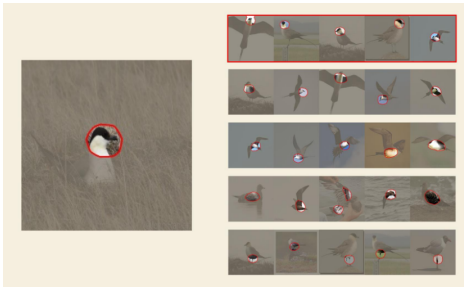


Figure 4: Survey trial sample in the prediction phase using NMF reducer. Participants instructions: “Each row on the right reflects some concepts in the form of image samples generated by our AI model. Look carefully at the image on the left and select the concepts on the right which it most likely belongs to.” The first row is the ground truth.

we measure the participants’ satisfaction with the explanations in terms of confidence, understanding, satisfaction, sufficiency and completeness (Hoffman et al. 2018). Explanations that are easy to interpret but perceived to be poor are unlikely to be used, so measuring subjective perception of explanations is important. We obtained ethics approval from The University of Melbourne Human Research Ethics Committee (ID 1749428).

Experimental Parameters: To validate the consistency of results, we include three different scenarios: ResNet50 (*layer4* as the target layer) for ILSVRC2012 (scenario RI), Inception-V3 (*Mixed_7c* as the target layer) for ILSVRC2012 (scenario II) and ResNet50 (*layer4* as the target layer) for CUB (scenario RC) as target CNN models. The three methods NMF, Clustering and PCA are applied individually in each scenario.

The 20 images were drawn from 20 random classes chosen from all classes within a given dataset. For each class, we train an explainer with ℓ' of 10, and only the top 5 CAVs with the highest weights are chosen as selection candidates. One of these CAVs is randomly selected as the target. The concept in the target CAV is identifiable only if the sample image highly activates that CAV. So each target image is chosen from the top 10% images in the test set which activates the target CAV most strongly (with high feature scores). Thus, we avoid using images in which the concept is absent (e.g., the *tail* concept may be considered absent when only the upper part of a dog is shown in the image). Each CAV is visualized using the 5 prototype samples. All explainers were trained with images from one class. The 20 classes were the same for different models for the same dataset, but candidate concepts and target instances were different.

Results Table 1 shows the results of the human-subject experiment. We ran an ANOVA to evaluate the results and used pairwise t-tests to identify differences between each pair of reducers. In the prediction task, NCAVs from NMF are more interpretable than CAVs from PCA ($p < 0.001$), and CAVs from clustering are more interpretable than CAVs from PCA ($p < 0.001$). For description similarity, results are not sig-

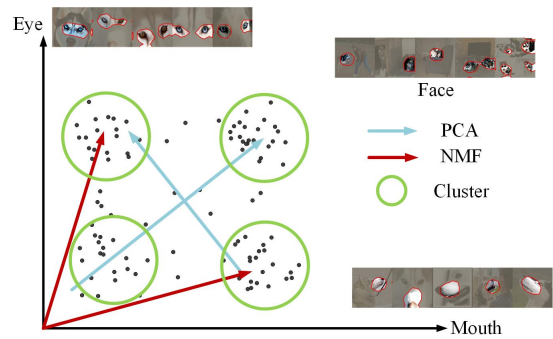


Figure 5: Having two concepts of ‘mouth’ and ‘eyes’ measured in two dimensions (only positive values reflect concepts), different reducers provide different directions to represent concepts. PCA learns less meaningful but efficient directions. Clustering methods could provide meaningful centroids’ center directions but are the least efficient. NMF may provide meaningful directions with fewer dimensions.

nificant ($p > 0.05$ level in most cases). Most CAVs contain some meaningful information; participants are confident about their choice. There is no significant difference in confidence and quality scores in most cases ($p > 0.05$). We conclude that NCAVs from NMF are more interpretable than CAVs from PCA. NCAVs are at least equally interpretable to CAVs from clustering.

We observe from our experiments that reducers can help generate meaningful concepts from feature maps, but fidelity and interpretability measure different aspects of performance. Here we propose an explanation for this phenomenon based on the differences between the three reducers. Figure 5 shows a distribution of some concept instances. Each dot reflects an instance with some concept scores. Due to the *Relu* activation in CNNs, we assumed that only positive values make sense in CNN models. The X axis could contain the concept of ‘mouth’ and Y axis may reflect ‘eyes’. PCA has a new intersection of dimensions (bias) other than the root so one of the dimensions is meaningless (points to negative values). Clustering methods provide correct concept directions (from the root to the center of each cluster). But it may require more clusters for the same fidelity, since clustering is based on data clusters but not directions. Also, it may provide some similar concepts (bottom left and upper right clusters have similar directions). Clusters may also be influenced by some isolated instances and provide meaningless concepts. But for NMF, it provides correct concept directions in an efficient way if only positive values reflect meaningful concepts.

Related Work

This work focuses on explanations for pre-trained models. Common explanation methods provide explanations based on input level feature importance. Some methods provide model agnostic explanations based on importance for image segments (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017) by training linear approximate models on nearby

Scenario	Reducer type	Accuracy	Description Similarity	Confidence	Quality			
					Understand	Satisfaction	Sufficiency	Completeness
RI	NMF	74.4% ± 9.2%	0.59 ± 0.1	77.7% ± 13.0%	4.3 ± 0.6	4.1 ± 0.6	3.8 ± 0.8	3.7 ± 1.2
	Cluster	66.3% ± 13.8%	0.56 ± 0.08	75.6% ± 13.8%	4.2 ± 0.7	3.8 ± 1.0	3.9 ± 1.0	3.6 ± 1.1
	PCA	37.8% ± 5.9%	0.52 ± 0.08	78.3% ± 14.7%	4.0 ± 0.9	3.8 ± 1.1	3.8 ± 1.1	3.7 ± 1.2
II	NMF	62.6% ± 18.6%	0.57 ± 0.08	69.3% ± 13.2%	3.5 ± 1.0	3.4 ± 1.3	3.3 ± 1.1	3.4 ± 1.3
	Cluster	44.8% ± 13.2%	0.53 ± 0.09	75.1% ± 13.7%	3.9 ± 1.1	3.6 ± 1.2	3.6 ± 1.2	3.5 ± 1.4
	PCA	40.0% ± 8.6%	0.49 ± 0.08	76.0% ± 13.0%	3.8 ± 0.9	3.7 ± 1.1	3.4 ± 1.2	3.2 ± 1.3
RC	NMF	81.1% ± 8.4%	0.7 ± 0.04	79.5% ± 10.8%	4.1 ± 0.8	3.7 ± 0.9	3.4 ± 1.2	3.5 ± 1.1
	Cluster	78.6% ± 15.5%	0.7 ± 0.05	75.0% ± 18.7%	3.9 ± 1.0	4.1 ± 1.0	4.0 ± 1.0	3.9 ± 1.1
	PCA	57.0% ± 11.6%	0.59 ± 0.03	61.1% ± 17.6%	3.6 ± 1.0	3.0 ± 1.2	3.4 ± 1.2	3.2 ± 1.2

Scenario	Accuracy	Description Similarity	Confidence	Quality			
				Understand	Satisfaction	Sufficiency	Completeness
RI	<0.001	0.131	0.841	0.446	0.592	0.941	0.948
II	<0.001	0.064	0.304	0.493	0.752	0.690	0.844
RC	<0.001	<0.001	0.004	0.283	0.016	0.219	0.174

Scenario	Reducer Pair	Accuracy	Description Similarity	Confidence	Quality			
					Understand	Satisfaction	Sufficiency	Completeness
RI	NMF vs. Cluster	0.159						
	NMF vs. PCA	<0.001	/	/	/	/	/	/
	Cluster vs. PCA	<0.001						
II	NMF vs. Cluster	0.019						
	NMF vs. PCA	<0.001	/	/	/	/	/	/
	Cluster vs. PCA	0.754						
RC	NMF vs. Cluster	1.00	1.00	1.00		0.784		
	NMF vs. PCA	<0.001	<0.001	0.002	/	0.222	/	/
	Cluster vs. PCA	<0.001	<0.001	0.088		0.023		

Table 1: Top: Mean and standard deviation of prediction accuracy, description similarity, confidence and quality comparison for 9 different groups. Middle: ANOVA test p values for each scenario. Bottom: Bonferroni corrected T-test p values for each pair of reducers with significant ANOVA p value ($p < 0.05$)

datasets. Saliency maps provide pixel-level feature importance for images based on gradients (Shrikumar, Greenside, and Kundaje 2017; Bach et al. 2015; Smilkov et al. 2017). However, some papers point out the unreliability of saliency methods (Kindermans et al. 2017; Dombrowski et al. 2019). CAM is another type of approach, providing heatmaps to indicate where the image activates the target class most based on CNN weights from the last several layers (Zhou et al. 2016; Selvaraju et al. 2017). All these methods use features from input images.

Other than input level explanations, some papers build explanations from feature maps inside the CNN model and provide concept-level explanations based on supervised learning (Kim et al. 2018; Bau et al. 2017; Zhou et al. 2018). These methods build linear classifiers on target concepts and use weights as CAVs. ACE (Ghorbani et al. 2019) relaxes the limitation of the labeled dataset using unsupervised learning with k -means clustering. Learned concepts take the form of vectors from cluster centroids.

Other than explanations for pre-trained models, some works modify the structure of CNN models to provide built-in interpretability (Hendricks et al. 2016; Zhang, Nian Wu, and Zhu 2018; Chen et al. 2019). Visualization of the values inside the CNN models can also help explain the process of

prediction. Optimization methods can visualize which patterns produce the highest activation within the CNN models (Olah, Mordvintsev, and Schubert 2017). There are many options for feature visualizations. These could choose different layers, different axes for target layer matrices, different matrices from CNN weights or instance feature maps and different methods of presentation to satisfy different interpretability requirements (Olah et al. 2018). In this work, NMF is introduced to gather interpretable information from a feature map. NMF is a common interpretable matrix factorization method (Lee and Seung 1999),

Conclusion

We propose a framework for concept-based explanations for CNN models based on the post-training explanation method ACE. By using feature maps inside the CNN model, we can gather some interpretable concept vectors to provide explanations and invert them back to predictions. We also show that having requirements of fidelity and interpretability, NCAVs from NMF can provide overall better explanations compared with clustering and PCA methods. PCA provides CAVs with better fidelity but lack interpretability, which makes PCA less suitable for explanations. CAVs from clustering methods are interpretable but lower in fidelity.

Acknowledgements

This research is supported by Australian Research Council (ARC) Discovery Grant DP190103414: *Explanation in Artificial Intelligence: A Human-Centred Approach*. The first two authors are supported by the University of Melbourne research scholarship (MRS) scheme. Experiments were undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

Ethics Statement

This work focuses on improving the interpretability of CNN models, helping human analysts understand models' decision processes. Accountability and transparency are two key areas of responsible artificial intelligence (Dignum 2019), and the ability to interpret models have a positive impact on both of these: accountability because making people accountable for decisions that are informed by machine learning is not reasonable if those people cannot gain insight as to why and how decisions are made; transparency because understanding the model helps to increase the transparency of decision making and to find inappropriate factors in the CNN models.

Despite this, there are potential negative impacts of this work. First, the explanations provided are approximate, and therefore can be incorrect in some cases. This has the potential to mislead people on the reasons for decisions. Second, examples are required as sample prototypes in the explanations. Inappropriate usage of training images may cause unintended and undesirable privacy disclosure if care is not taken to use public, consented sources or if privacy protections are not employed.

References

- Akula, A. R.; Wang, S.; and Zhu, S.-C. 2020. CoCoX: Generating Conceptual and Counterfactual Explanations via Fault-Lines. In *AAAI*, 2594–2601.
- Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS one* 10(7): e0130140.
- Bau, D.; Zhou, B.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*, 6541–6549.
- Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science* 6(1): 3–5. doi:10.1177/1745691610393980.
- Chen, C.; Li, O.; Tao, D.; Barnett, A.; Rudin, C.; and Su, J. K. 2019. This looks like that: deep learning for interpretable image recognition. In *NIPS*, 8928–8939.
- Craven, M.; and Shavlik, J. W. 1996. Extracting tree-structured representations of trained networks. In *NIPS*, 24–30.
- Dai, J.; He, K.; and Sun, J. 2015. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 3992–4000.
- Dignum, V. 2019. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing.
- Dombrowski, A.-K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.-R.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. In *NIPS*, 13567–13578. Curran Associates, Inc.
- Ghorbani, A.; Wexler, J.; Zou, J. Y.; and Kim, B. 2019. Towards automatic concept-based explanations. In *NIPS*, 9273–9282.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.
- Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *ECCV*, 3–19. Springer.
- Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.
- Kim, B.; Khanna, R.; and Koyejo, O. O. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*, 2280–2288.
- Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; et al. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*, 2673–2682.
- Kindermans, P.-J.; Hooker, S.; Adebayo, J.; Alber, M.; Schütt, K. T.; Dähne, S.; Erhan, D.; and Kim, B. 2017. The (Un)reliability of saliency methods. *NIPS workshop on Explaining and Visualizing Deep Learning*.
- Leavitt, M. L.; and Morcos, A. 2020. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*.
- Lee, D. D.; and Seung, H. S. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755): 788.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *NIPS*, 4765–4774.
- Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature Visualization. *Distill* doi:10.23915/distill.00007. <https://distill.pub/2017/feature-visualization>.
- Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The Building Blocks of Interpretability. *Distill* doi:10.23915/distill.00010. <https://distill.pub/2018/building-blocks>.
- Pennington, J.; Socher, R.; and Manning, C. D. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*, 1532–1543. URL <http://www.aclweb.org/anthology/D14-1162>.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 1135–1144. ACM.
- Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1(5): 206–215.

- Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* 115(3): 211–252. doi: 10.1007/s11263-015-0816-y.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.
- Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *ICML*, 3145–3153. JMLR. org.
- Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* .
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology.
- Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *ECCV*, 818–833. Springer.
- Zhang, Q.; Nian Wu, Y.; and Zhu, S.-C. 2018. Interpretable convolutional neural networks. In *CVPR*, 8827–8836.
- Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.
- Zhou, B.; Sun, Y.; Bau, D.; and Torralba, A. 2018. Interpretable basis decomposition for visual explanation. In *ECCV*, 119–134.