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-
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 activation vector to provide approximate models directly. For the final step, we build a linear approximation to the feature maps to provide CAVs 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).

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 $c\prime$. Here $V$ is reduced to
feature score \( S \in \mathbb{R}^{(n \times h \times w) \times c'} \) and feature direction \( P \in \mathbb{R}^{c' \times c} \) 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 \( \frac{\partial C_{l,k}}{\partial p_l} = \frac{1}{n \times h \times w} \sum_{a \in A_l} \lim_{a \to 0} \frac{h_{l,k}(a+\epsilon) - h_{l,k}(a-\epsilon)}{2\epsilon} \). 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 \( W \in \mathbb{R}^{c} \) 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.
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 interpretable, 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:

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

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

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

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 or label) 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 $Fid_{c,F}$ (middle) and $Fid_{c,R}$ (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
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 (Buhmester, 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,
Experimental Parameters: To validate the consistency of our AI model, we obtained ethics approval from The University of Melbourne Human Research Ethics Committee (ID 1749428).

Explanations are important. 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).

The University of Melbourne Human Research Ethics Committee (ID 1749428).

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 significant ($p > 0.05$ 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
inside the CNN models can also help explain the process of
and Zhu 2018; Chen et al. 2019). Visualization of the values
in interpretability (Hendricks et al. 2016; Zhang, Nian Wu,
works modify the structure of CNN models to provide built-
clusters. Learned concepts take the form of
discrimination method (Lee and Seung 1999),
a feature map. NMF is a common interpretable matrix fac-
dermatrices from CNN weights or instance feature maps
different layers, different axes for target layer matrices, different
CA Vs from NMF can provide overall better expla-
show that having requirements of fidelity and interpretability
In this work, NMF is introduced to gather interpretable information from
CNN models based on the post-training explanation method

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


