Name:	Pingchuan Ma
Student number:	3526400
Date:	August 19, 2018

Seminar: Explainable Machine Learning
Lecturer: PD Dr. Ullrich Köthe
SS 2018

Network Dissection: Quantifying Interpretability of Deep Visual Representation

Seminar Report

Pingchuan Ma

Contents

1 Introduction

The chosen paper is "Network Dissection: Quantifying Interpretability of Deep Visual Representations." from B. Zhou et. al. [2] of Topic 3: Feature Visualization of the seminar. The paper didn't propose some fancy network structure but conduct empirical studies on common properties of neural networks. It tried to explain the function of each unit with the help of a broadly densely labeled dataset, as they measure the alignment between units and images in this dataset.

Over last few years deep convolutional neural networks (CNNs) have been used widely in computer vision especially for visual recognition and they have achieved the state-of-the-art performance, which accelerates the research in this field greatly. However, being one of the most powerful and sophisticated tool we have, CNNs were often criticized as black boxes for their lack of interpretability, since they have millions of unexplained model parameters, which we can't find a proper way to explain as we explain those early engineered representations such as Harris Corner, SIFT and HOG. Although people may still be able to explain how does the network works in the first few layers for some units,they can't manage to do for all the units inside a CNN.

1.1 Related Work

To understand what happened inside CNNs, some previous methods have been developed. The behavior of a CNN can be visualized by sampling image patches that maximize the activation of hidden units [7] [8]. The earlier layers capture the simple features such as edge, corner or texture, while the later layers capture more complicated and semantically meaningful concepts such as dog head, or truck. Variants of backpropagation are also be used to identify or generate salient image features [4]. They back project the weights learned by our networks back onto the image space to see what have been learned insides. The discriminative power of hidden layers of CNN features can also be understood by isolating portions of networks, transferring them or limiting them, and testing their capabilities on specialized problems [1]. The visualization approaches digest the mechanisms of a network down to images which themselves must be interpreted, and motivate to match representations of CNNs with labeled interpretations.

Most relevant to this paper's work are explorations of the roles of individual units inside neural networks. [8] shows that individual units behave as object detectors in a scene-classification network determined by human evaluation. [5] automatically generated prototypical images for individual units by learning a feature inversion mapping, which contrast the approach of automatically assigning concept labels from this paper.

They have revealed that CNNs may be learning spontaneously the disentangled representation, which aligns its variables with a meaningful factorization of the underlying problem structure, for example, object detector units emerge within network trained to recognize places, and part detectors emerge in object classifier network, without any explicit instruction telling them to understand the task in any interpretable way.

1.2 Motivation

To quantify the interpretability, an usual way used by neuroscientists to solve similar problems in biological neurons is to explain it with some ideas we already know, which includes three steps:

1. Identify a broad set of human-labeled visual concepts. *(Broden Dataset)*
2. Gather hidden variables' response to know concepts. *(Distribution of individual unit activation beyond a certain threshold)*
3. Quantify alignment of hidden variable-concept pairs. *(Calculate the Intersection over Union - IoU value between them, namely individual hidden units in network and single concepts in Broden dataset)*

Past experience shows that an emergent concept will often align with a combination of several hidden units or vice versa - partly disentangled [1] [3], while we want to assess how well a representation is disentangled. Therefore, the authors define the interpretability in terms of the alignment between single units and single interpretable concepts.

2 Dataset & Methods

Following the three steps mentioned above. The authors draw concepts from a new broadly and densely labeled image dataset that unifies labeled visual concepts from a heterogeneous collection of labeled data sources, described in subsection 2.1. They then measure the alignment of each hidden unit of the CNN with each concept by evaluation the feature activation of each individual unit as a segmentation model for each concept. They define the interpretability of a certain layer by counting the the number of unique visual concepts aligning with a unit in this layer, as detailed in subsection 2.2

2.1 Broden: Broadly and Densely Labeled Dataset

To be able to ascertain alignment with bot low-level concepts such as colors and higher-level concepts such as objects, a new heterogeneous dataset has been assembled.
The Broadly and Densely Labeled Dataset *(Broden)* unifies several densely labeled image datasets, see Table 1. These datasets contain examples of a broad range of objects, scenes, object parts, textures, and materials in a variety of contexts, and most of them are segmented down to pixel level except textures and scenes which are full-image concepts.

Category	Classes	Sources	Avg sample
scene	468	ADE	38
object	584	ADE, Pascal-Context	491
part	234	ADE, Pascal-Part	854
material	32	OpenSurfaces	1,703
texture	47	DTD	140
color	11	Generated	59,250

Table 1: Statistics of each label type included in the dataset

The purpose of Broden is to provide a ground truth set of exemplars of normalized and cleaned visual concepts, which means every class corresponds to an English word based on merging synonyms, disregarding positional distinctions such as ’left’ and ’top’ and avoiding a black list of 29 overly general synonyms (such as “machine” for “car”). Besides, multiple Broden labels can apply to the same pixel. For example, a blue pixel of a sea scene can has a color label "blue" as well as a scene label "sea".

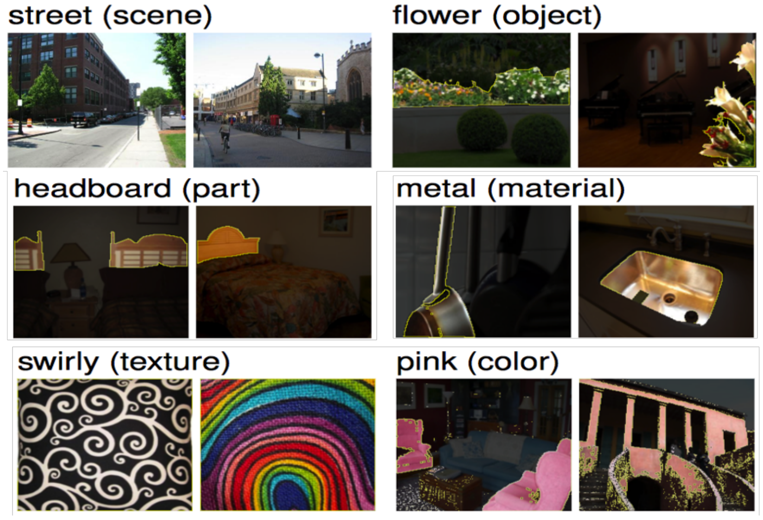


Figure 1: Samples from the Broden Dataset. The ground truth for each concept is a pixel-wise dense annotation.

2.2 Scoring Unit Interpretability

The proposed Network Disscetion method compares every single unit inside the given CNN with every visual concept in Broden as a binary segmentation task.

The activation map $A_k(x)$ of every internal convolutional unit k is collected for every input image x in the Broden dataset. For each unit k, the top quantile level T_k is determined such that $P(a_k > T_k) = 0.005$ over every spatial location of the activation map in the dataset. $A_k(x)$ is then thresholded into a binary segmentation: $M_k(x) \equiv A_k(x) \geq T_k$, selecting all regions for which the activation exceeds the threshold T_k. All these segmentation are evaluated against every concept c in the data set by computing intersections $M_k(x) \cap L_c(x)$, for every (k, c), unit-concept pair.

The score of each unit k as segmentation for concept c is reproted as a data-set-wide *intersection over union score*

$$IoU_{k,c} = \frac{\sum |M_k(x) \cap L_c(x)|}{\sum |M_k(x) \cup L_c(x)|} \tag{1}$$

The value of $IoU_{k,c}$ is the accuracy of unit k detecting concept c, if it exceeds a threshold. Although the number of units as detectors is affected by different threshold, the the relative ordering remain stable. To quantify the interpretability of a layer, the authors count the number of unique concepts aligned with units inside this layer, which is called *unique detectors*.

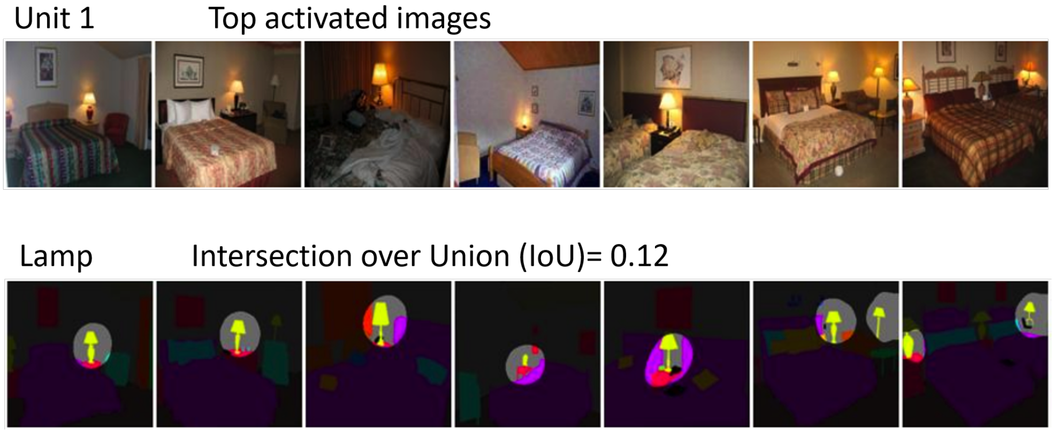


Figure 2: Demonstration of how the IoU value is evaluated for a certain unit.

Given the IoU as an objective confidence score for interpretability, we are able to compare the interpretabilities of different representations across different layers as well as different networks, which also conducts the structure of the following experiment section. However, we must notice that, since the Broden dataset is a dataset pre-defined by the authors, if a unit matches a human-understandable concept that is absent in Broden, then it will not score well for interpretability. Future expansion of Broden may ease this drawback.

3 Experiments and Results

The authors test the Network Disscetion on a collection of CNN models with different network structures and supervision of primary tasks, see table below.

Training	Network	Dataset or task
none	AlexNet	random
Supervised	AlexNet	ImageNet, Places205, Places365, Hybrid.
	GoogLeNet	ImageNet, Places205, Places365.
	VGG-16	ImageNet, Places205, Places365, Hybrid.
	ResNet-152	ImageNet, Places365.
Self	AlexNet	context, puzzle, egomotion, tracking, moving, videoorder, audio, crosschannel, colorization, objectcentric.

Table 2: Tested CNN models. **Hybrid** contains both part of ImageNet and part of Places365. **Self** represent self-supervised learning, there is a really great video by Alexei Efros introduced it. https://www.youtube.com/watch?v=YhYsvD6IfKE last accessed 18,Aug,2018

Outline of Experiments. They begin with human evaluation as a validation to the method. Then they test whether it is meaningful to assign an interpretable concept to an individual unit, which they found it's not and concluded that interpretability is not an inevitable result of the discriminative power of a representation. By analyzing all the convolutinoal layers of two AlexNets as trained on ImageNet and as trained on Places, the hypothesis that detectors for higher-level concepts at higher layers and lower-level concepts at lower layers was confirmed, besides, more detectors for higher-level concepts emerged under scene training compared to object training. After that, they showed that different network architectures such as AlexNet and VGG-16 and different super- and self-supervised training yield various interpretability. Finally, they showed the impact of different training conditions, examined the relationship between discriminative power and interpretability and investigated possible way to improve the interpretability of CNNs by increasing their width (units number).

3.1 Human evaluation

They use Amazon Mechanical Turk (AMT) to evaluate the results found by Network Disscetion. Raters were shown 15 images with highlighted patches showing the most activated regions selected by each unit, at the meantime they were ask a yes/no question, that whether a given phrase generated by Network Disscetion describes most of the image patches. Results are the table below.

	conv1	conv2	conv3	conv4	conv5
Interpretable units	57/96	126/256	247/384	258/384	194/256
Human consistency	82%	76%	83%	82%	91%
Network Dissection	37%	56%	54%	59%	71%

Table 3: Human evaluation of Network Disscetion. Interpretable units are those where raters agreed with ground-truth interpretations, namely the number of units that the rater think they represent the image. Within this set the portion of interpretations assigned by Network Dissection were rated as descriptive, based on IoU score. Human consistency is based on a second group evaluation of ground-truth labels.

The result suggests that humans are better at recognizing and agreeing upon high-level concepts such as objects and parts, rather than the shapes and textures that emerge at lower layer.

3.2 Axis-Aligned Interpretability

Two possible hypotheses may explain the emergence of interpretability in individual hidden layer units:

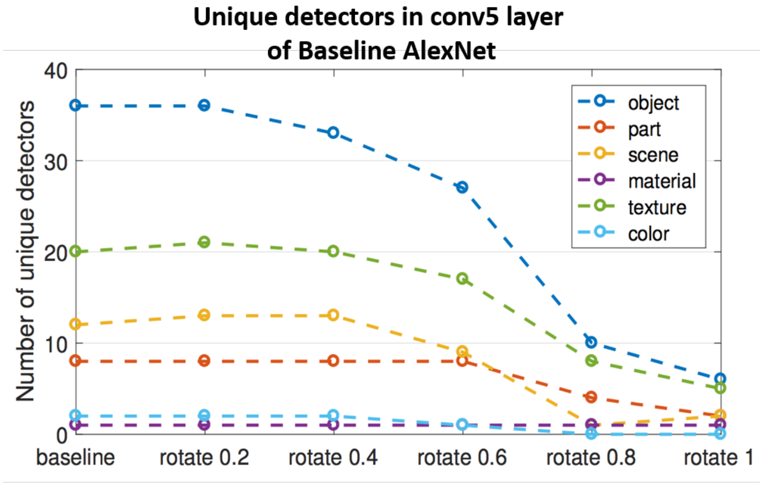
1. The overall level of interpretability should not be affected by a change in rotation in feature space. This reveals that interpretations of single units in the natural basis may not be a meaningful way to understand a representation.
2. The overall level of interpretability is expected to drop under this change, which means that the natural basis represents a meaningful decomposition learned by the network.

Hypothesis 1 is the default assumption supported by [6]. Under this hypothesis, if we change the basis to a representation learned by CNN, the overall level of interpretability should not be affected or at least not that much.

The team tested this by applying a random orthogonal transformation Q on the $conv5$ layer of an AlexNet trained on Places205. The rotation Q is drawn uniformly from $SO(256)$ by applying Gram-Schmidt on a normally-distributed $QR = A \in \Re^{256^2}$ with positive-diagonal right-triangular R.

They found that the number of unique detectors, namely the interpretability of a layer they defined, in $Qf(x)$, is 80% fewer than it in $f(x)$, where $f(x)$ represents above-mentioned conv5 of AlexNet. The finding rejects hypothesis 1 and is consistent with hypothesis 2.

In addition, they've also tried the rotation with a factor α on Q, where $0 \leq \alpha \leq 1$, to see how the intermediate rotations affect the interpretability.



Unique detectors in $Qf(x)$ is much fewer than in $f(x)$

However each rotated representation has exactly the same discriminative power as the original one.

Figure 3: Interpretability over changes in basis of the representation of AlexNet conv5 trained on Places. The vertical axis shows the number of unique interpretable concepts that match a unit in the representation. The horizontal axis shows α, which quantified the degree of rotation.

However, each rotated representation has exactly the same discriminative power as the original layer. They concluded that the interpretability of CNNs is not an axis-independent property, and it is neither an inevitable/ necessary result of the discriminative power of a representation, nor is a prerequisite to discriminative power. Instead, the interpretability is more likely to be a different quality from discriminative power that must be measured separately to be understood.

3.3 Layer levels

Using network dissection, they analyzed and compared the interpretability of units within all the conv layers of AlexNet trained on Places and ImageNet. The results are summerized in Figure 4

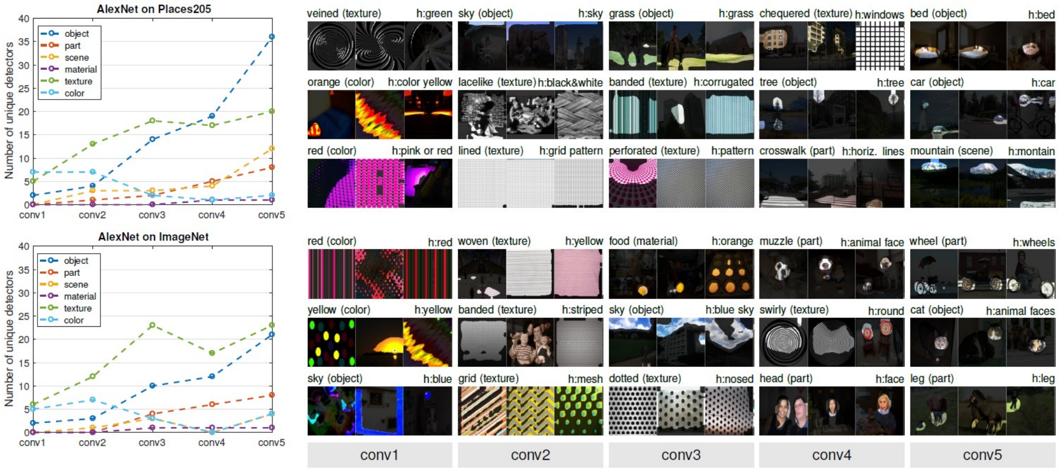


Figure 4: A comparison of the interpretability of all five convolutional layers of AlexNet, as trained on classification tasks for Places (top) and ImageNet (bottom). At right, three examples of units in each layer are shown with identified semantics. The segmentation generated by each unit is shown on the three Broden images with highest activation. Top-scoring labels are shown above to the left, and human-annotated labels are shown above to the right. Some disagreement can be seen for the dominant judgment of meaning. For example, human annotators mark the first conv4 unit on Places as a 'windows' detector, while the algorithm matches the 'chequered' texture.

We can see that the predicted label match the human annotation well, though sometimes capture different description of a visual concept. The result confirmed the intuition, color and texture concepts dominate at lower layers such as conv1 and conv2, while more object and part detectors emerge in higher-level layer conv5.

3.4 Architectures and supervisions

How do the representations of different network architectures and training supervisions affect the interpretability of the learned representations? The author tested Network Dissection on different models listed in Table 2. For simplicity they compared only the last conv layer of each models, since here is where semantic detectors emerged Figure 5.

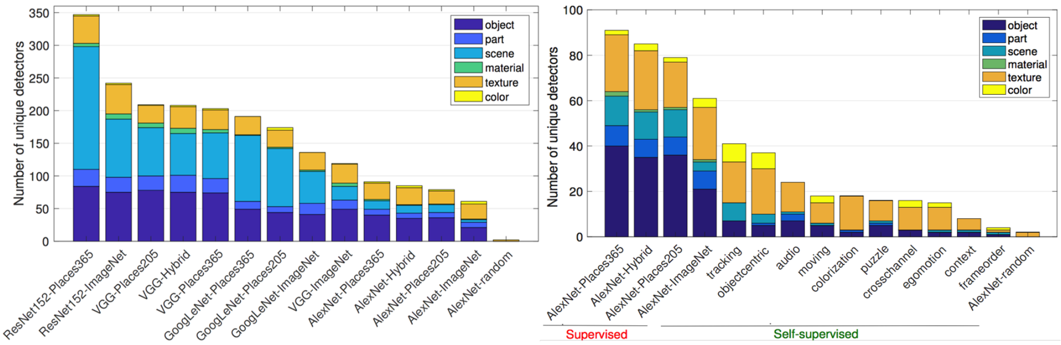


Figure 5: **a).**Interpretability of ResNet > VGPlaces205 G > GoogLeNet > AlexNet, and in terms of primary training tasks, we find Places365 > > ImageNet.
b). Interpretability varies widely under a range of self-supervised tasks, and none approaches interpretability from supervision by ImageNet or Places.

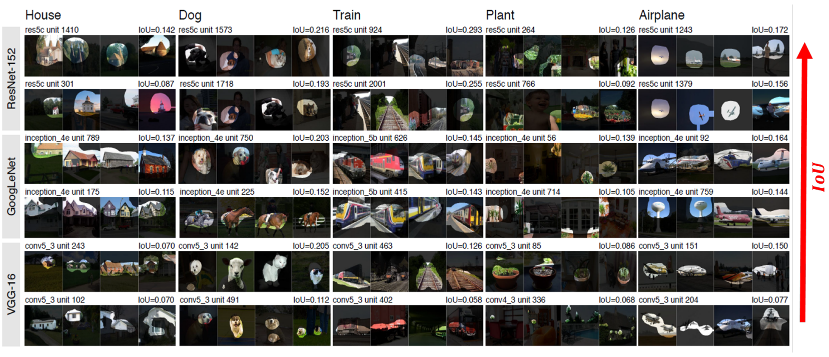


Figure 6: The two highest IoU scored units are shown. Deeper networks also have a higher IoU score. Besides, the detectors in deeper network detect more compact and generalized concepts than those shallower networks. Although the IoU scores are pretty low compared to those supervised object detectors, they are spontaneous learned by our intermediate hidden units. So it's pretty impressive.

Conclusions. Networks with complex architecture have usually more interpretability. And models trained on Places have more unique detectors than those trained on ImageNet. Because scene is composed of multiple objects, it may be beneficial for more object detectors to emerge in CNNs trained to recognize scenes. There also more low-level concepts emerge in self-supervised model. Apparently supervision from a self-taught primary task is much weaker at inferring interpretable concepts than supervised learning on a large annotated dataset.

3.5 Training conditions

Different training conditions such as number of training epoch, dropout, batch-normalization, and different way to initialize, are known to affect the representation learned by CNN. In this subsection the baseline model Places205-AlexNet was compared with its variants.

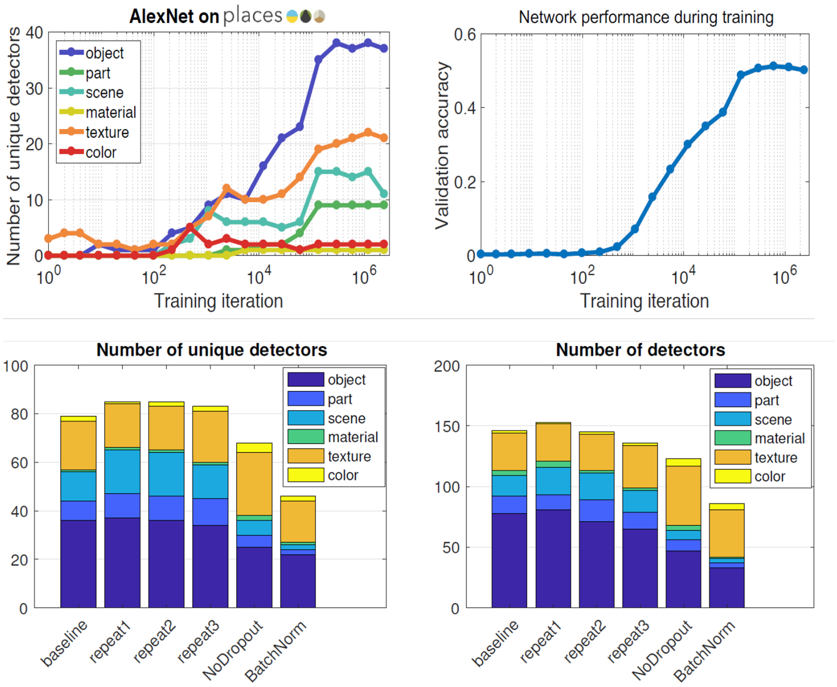


Figure 7: Results of different training conditions.

Conclusions. 1.for different random initializations, the models converge to similar levels of interpretability, both in unique number and total number, see repeat 1 3 in Figure 7. 2. For dropout, there are more texture and color detectors emerge but fewer object detectors. 3. And batchnorm seems to decrease interpretability significantly. This may give us a hint that the discriminative power may not the only thing we need to measure for a representation. And the loss may caused by the reason that batchNorm smooths out scaling issues and allows a network to easily rotate axes of intermediate representations during training. While it speeds up the training, it may also have an effect similar to random rotations analyzed before, that destroy interpretability. How to find a balance between them should be an important further work.

3.6 Discrimination vs. Interpretability

People easily treat discriminative power and interpretability as the same thing. However, in this section they showed actually we'd better take them as two different qualities of a CNN into account.

Activations from the higher layers of CNNs are often used as generic visual features, showing great discrimination and generalization ability. For this task, they took several networks, train them on standard image classification datasets for the discriminative ability. And then extract the highest conv layer, and train a linear SVM on dataset action40 for action recognition task, and compute the classification accuracy.

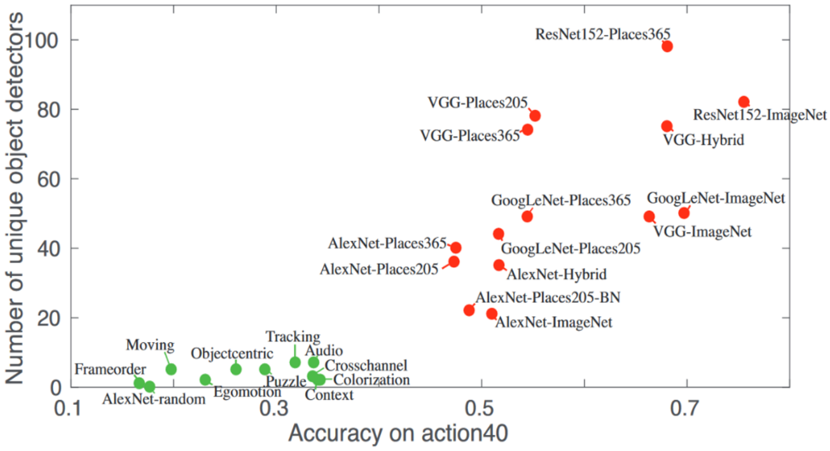


Figure 8: There is a positive correlation between them. Thus the emergence of detectors may also improve discriminative power. Interestingly, the best discriminative representation for action40 is the representation from ResNet152-ImageNet, which has fewer unique object detectors compared to ResNet152-Places3365. The possible explanation maybe that the accuracy depends not only on the number of detectors but also on whether these detectors are suitable for the primary task.

3.7 Layer Width vs. Interpretability

From AlexNet to ResNet, CNNs have grown deeper in the quest for higher accuracy. Depth has been shown to be important to high discriminative power, and we have seen before the interpretability increase as well. However the width (units per layer) has been less explored. Increasing width-significantly increase computational cost, while brings only small improvement in accuaracy. To explore how it affects our interpretability, they build another version of AlexNet - AlexNet-GAP-Wide, as they removed FC-layers, tripled the number of units in conv5, i.e. 256 to 768 units, finally put a global average pooling layer after conv5 and fully connect the pooled 768-features activations to the final class prediction.

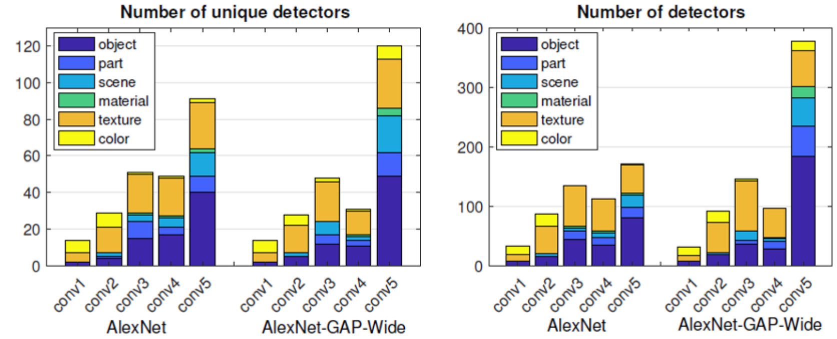


Figure 9: The impact of network width on interpretability. Accuracy is similar, but many more concept detectors emerged, both in terms of unique detectors and total number. After that they also tried increase to 1024 and 2048 at conv5, but the number did not change significantly, this may indicated that a limit on the number of disentangled concepts that are helpful to solve the primary task, which is consistent with the explanation from Figure 8

4 Conclusion

The paper proposed a general framework - Network Dissection, based on a densely labeled dataset (Broden) to quantify the interpretability (based on the definition given by the authors) of any CNNs. The related experiments reveal that:

- Interpretability is not an axis-independent phenomenon, which is consistent with the hypothesis that interpretable units indicate a partially disentangled representation.

- Deeper CNNs architectures appear to allow a greater interpretability. Besides, the interpretability also depends on the concepts in the dataset used for training the model.

- Representation at different layers of CNNs disentangle different categories of meaning. In other word, the lower/higher the layer level is, the lower/higher level of concepts it contains. For example, detectors such as edge, corner and texture tend to emerge at some earlier layers, while concrete concepts of object and parts tend to emerge at later layers.

- Different training techniques and conditions lead to a significant change of interpretability of representation learned by hidden units. For example, a self-supervised learning model has less more interpretability than a supervised learning model.

- Interpretability and discriminative power are two qualities that need to be measured separately, though they have a positive correlation.

They defined the interpretability as the number of unique detectors, and then based on this assumption performed different experiments, which I consider may not that convincing, because of the self-defined standard. Nevertheless, personally I think the paper is quite a good example to demonstrate how to conduct a systematic analysis work. It didn't contain any fancy method, but a more intuitive approach.

References

1. P. Agrawal, R. B. Girshick, and J. Malik. “Analyzing the Performance of Multilayer Neural Networks for Object Recognition”. *CoRR* abs/1407.1610, 2014. arXiv: 1407.1610. URL: http://arxiv.org/abs/1407.1610.
2. D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. “Network Dissection: Quantifying Interpretability of Deep Visual Representations”. In: *Computer Vision and Pattern Recognition*. 2017.
3. A. Gonzalez-Garcia, D. Modolo, and V. Ferrari. “Do semantic parts emerge in Convolutional Neural Networks?” *CoRR* abs/1607.03738, 2016. arXiv: 1607.03738. URL: http://arxiv.org/abs/1607.03738.
4. A. Mahendran and A. Vedaldi. “Understanding Deep Image Representations by Inverting Them”. *CoRR* abs/1412.0035, 2014. arXiv: 1412.0035. URL: http://arxiv.org/abs/1412.0035.
5. A. M. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. “Synthesizing the preferred inputs for neurons in neural networks via deep generator networks”. *CoRR* abs/1605.09304, 2016. arXiv: 1605.09304. URL: http://arxiv.org/abs/1605.09304.
6. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus. “Intriguing properties of neural networks”. *CoRR* abs/1312.6199, 2013. arXiv: 1312.6199. URL: http://arxiv.org/abs/1312.6199.
7. M. D. Zeiler and R. Fergus. “Visualizing and Understanding Convolutional Networks”. *CoRR* abs/1311.2901, 2013. arXiv: 1311.2901. URL: http://arxiv.org/abs/1311.2901.
8. B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. “Object Detectors Emerge in Deep Scene CNNs”. *CoRR* abs/1412.6856, 2014. arXiv: 1412.6856. URL: http://arxiv.org/abs/1412.6856.