TUKL - INDUCING INTERPRETABILITY IN CNN

Machine Learning Group, Technische Universität Kaiserslautern
Saurabh Varshneya, Antoine Ledent, Marius Kloft, Steffen Reithermann
Hochleistungsrechner: **ELWETRITSCH HPC CLUSTER (DGX-2)**, Nutzer: 1

Project goals

Deep neural networks are powerful models that achieve excellent results in the domain of machine learning problems such as image classification, speech recognition, and natural language processing. CNNs (Convolutional Neural Networks) in particular achieve state of the art results for several vision-related tasks such as object detection, image classification or pixel-wise segmentation. While they achieve state of the art results in various domains, they also tend to learn complex solutions via uninterpretable feature extraction in the hidden layers. These uninterpretable solutions many times lead counter-intuitive properties in neural networks and create a need for interpreting hidden layers of a deep neural network.

The interpretation of hidden representations in a CNN is a significant research field in the area of deep learning. Besides being theoretically interesting, CNNs with reasonable interpretations are often much more desirable than black box models in critical applications, where they can also help users gain trust in their predictions. Application areas like medicine or self-driving cars demand the deep learning models focusing on correct features to make critical decisions, which indeed can be fulfilled only when interpretability of hidden representations can be improved.

The interpretability of any neural network, however, is difficult to measure and define in terms of quality and quantity. In most of the cases, the interpretability of a CNN is estimated qualitatively by visualizing its hidden representations. *"Interpretability"* of deep visual representations can also be defined as the degree of alignment with certain human-interpretable visual concepts. The interpretability of a hidden layer in a CNN can be quantified by counting the number of convolutional filters which are well-aligned with one of the visual concepts. An interpretable CNN then will be the one where all its filters uniquely align to one of the human-readable concepts. It has been observed that the convolutional filter show some alignment to these visual concepts where higher layer filters learn higher-level visual concepts such as objects and parts, and lower-level filters learn lower-level concepts such as color and textures. Our goal is to apply special regularizers while training the neural network, which constrains the parameters of a CNN in a soft manner, such that its filters tend to forms groups in a layer with similar representations, and only a few groups become active in each layer throughout the network. Figure 1 shows the idea of a regularization framework which induces interpretability in a fully unsupervised manner.

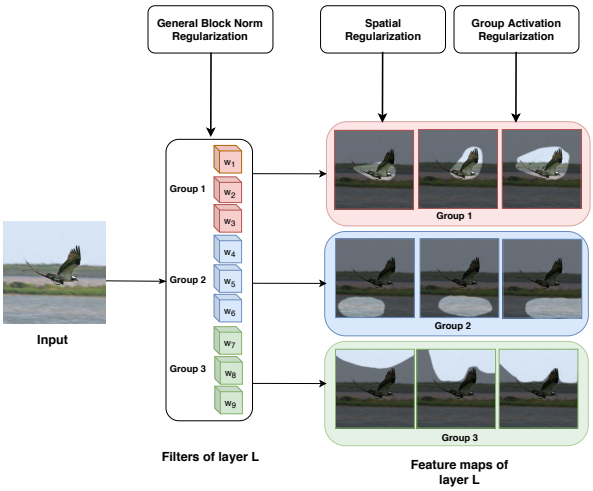

Fig. 1 An illustration of our training algorithm for a layer, with general block norm inducing a group structure where only a few groups of filters are active. The activation regularization further enhance the group structure where the filters of a specific group tend to be active in a similar region of an image and the spatial regularization constrain the activation to be concentrated in a single area of an image.

Current status

Typically, a CNN is regularized with a weight-decay (l^2 norm) regularizer over all the filter weights $\mathbf{w}$. The regularizer $\Omega(\mathbf{w})$ can be described as:

$$\Omega(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2 \,, \qquad (1)$$

which doesn't exploit the hierarchical group structure of CNNs. We induce a group structure in the filters of CNN by dividing filters in each layer into fixed predefined number of groups G and extending the regularizer for layer l as :

$$\Omega(\mathbf{w}^l) = \left\| \left(\|\widetilde{\mathbf{w}}_1^l\|_p, ..., \|\widetilde{\mathbf{w}}_G^l\|_p \right) \right\|_q, \qquad (2)$$

where,

$$\forall g \in \{1, ..., G\}, \;\; \widetilde{\mathbf{w}}_g^l = (\mathbf{w}_{ig}^l)_{i=1}^{N_g^l}.$$

Here, $N_g^l \in \mathbb{R}$ denotes the number of filters in group g of layer l and $\mathbf{w}_{ig}^l$ denotes the i^{th} filter in group g of layer l. Equation 2 denotes a generic group-norm type regularizer, parameterized by $1 \leq p, q \leq \infty$. For comparison, we quantitatively measure the interpretabilty score for both the models trained with both the regularizers and also apply visualization techniques on the filters of the trained CNN model. To quantify the interpretability of a trained CNN, we use the evaluation metrics which quantify the alignment of each convolutional filter with one of the six defined human-interpretable visual concepts. These visual concepts are *objects, object parts, colors, scenes, materials* and *textures*. Evaluation is performed on a densely labeled dataset called *Broden*, which contains pixel level annotations of each image. Each pixel of an image belongs to one or more classes. Further, these classes are categorized into above mentioned visual concepts.

To quantify a filter of a trained network, its activation map is computed in the forward pass and is thresholded to a binary mask $B_i(x)$. This selects the maximally activated region, of a filter, for an image. The alignment of this filter to a concept c is then calculated by computing the Intersection over Union (IoU) score between the obtained binary activated region $B_i(x)$ and the corresponding pixel-level annotation $G_c(x)$ for concept c from the Broden dataset as:

$$\text{IoU}_{w_i,c} = \frac{\sum_{x \in X} |B_i(x) \cap G_c(x)|}{\sum_{x \in X} |B_i(x) \cup G_c(x)|}, \qquad (3)$$

The i^{th} filter is considered to learn a concept c if the above IoU score exceeds a certain threshold.

We tested our network on the state-of-the-art deep learning network Alexnet training on the popular image dataset, such as Imagenet and Places365. For evaluation purpose, we use two versions of Alexnet, one without bath normalization layers and one with a batch-normalization layer after every convolutional layer, we call it as Alexnet-B. 1 shows the comparison of the interpretability score for the common weight-decay regularizer against the general block norm regularizer.

Tab. 1 Table shows the sum of the number of unique detectors in each layer while training CNNs with different Datasets. We achieve a better interpretability score only with our regularizers replacing the weight-decay regularizer

Dataset	Model	Interpretability Score	
		Weight-Decay	General block Norm
Places365	Alexnet	130	187
Places365	Alexnet-B	148	166
Places501	Alexnet	119	142
Places501	Alexnet-B	105	115
Places502	Alexnet	117	130
Places502	Alexnet-B	106	111

Publications

1 Karen Simonyan and Andrea Vedaldi and Andrew Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Iclr (2014).

2 Bolei Zhou and David Bau and Aude Oliva and Antonio Torralba, Interpreting Deep Visual Representations via Network Dissection, IEEE transactions on pattern analysis and machine intelligence (2018)

3 Marius Kloft and Ulf Brefeld and Sören Sonnenburg and Alexander Zien, Lp-norm multiple kernel learning, Journal of Machine Learning Research (2011)

Future plans

In the future, we want to extend our regularizer to a complete regularization framework. Our aim will be to implement new regularization strategies on the activation of the CNN which act as inductive biases to enhance the group structure in the hidden representation of a CNN. Further, we want to compare the regularization strategies on latest neural network architectures with skip connections, such as, ResNet and DenseNet.