FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

STUDENTSKÁ
VĚDECKÁ KONFERENCE 2021

# Vision Features in Artificial Neural Networks

Tomáš Zítka[1]

## 1 Introduction

Deep neural networks (DNN) have become prominent in almost all branches of machine learning. However, despite their widespread use, we still lack a robust understanding of the algorithms they implement. Particularly interesting is the process by which the algorithm in a given DNN evolves during optimization. To explore this we study SEResNeXt model [2] in three stages: randomly initialized, trained on ImageNet dataset and fine-tuned for a domain-specific classification task with five classes. We demonstrate that new, useful insights into the training process might be possible using the technique of feature visualization [1] combined with dimension reduction.

## 2 Vision Features in Artificial Neural Networks

Given a convolutional neural network $N$ composed of $L$ convolutional layers each with $C(l)$, $l = 1, 2, ..., L$ convolution kernels , we denote $N_l^i$ the $i$-th convolution filter in $l$-th layer (commonly referred to as neuron or unit) and $N_l^i(x)$ its output before applying non-linear activation function, called pre-activation, for image $x$ . In feature visualization algorithm we aim to find an image that maximizes pre-activation of selected neuron,

$$x = \arg\max_{x \in \mathbb{R}^2} N_l^i(x).$$

To improve the robustness of the optimization process, we optimize over frequency domain [4] and apply various transformations (e.g., jitter, random scale, or rotation) to the intermediate images during the optimization process. Figure 1 shows examples of feature visualizations of units in different layers from different models. Interpreting feature visualization is an open problem, however, by projecting the feature visualization vectors to low dimensional spaces e.g. 2D, we can gain some insights. For this we sampled $64$ neurons from each convolutional layer of our three SeResNeXt50 models and used UMAP algorithm [3] to project them. As can be seen in Figure 2 feature visualizations cluster by layer and are differently distributed in each stage of the training.

## 3 Conclusion

In this work, we created a dataset of feature visualizations for a deep convolutional network in three different stages of training and demonstrated the different distribution of learned features. In future work, we intend to further study the dataset as well as explore more representative and hopefully more computationally effective methods for feature representation.

[1] student navazujícího doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Multimodální interakce člověka-stroj, e-mail: zitkat@students.zcu.cz

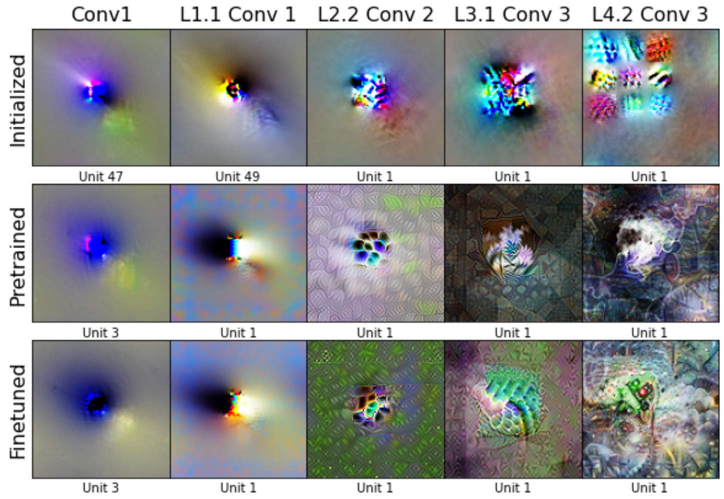

**Figure 1:** Feature visualizations for selected units from different layers in randomly initialized, pre-trained and fine-tuned SeResNeXt50 network.

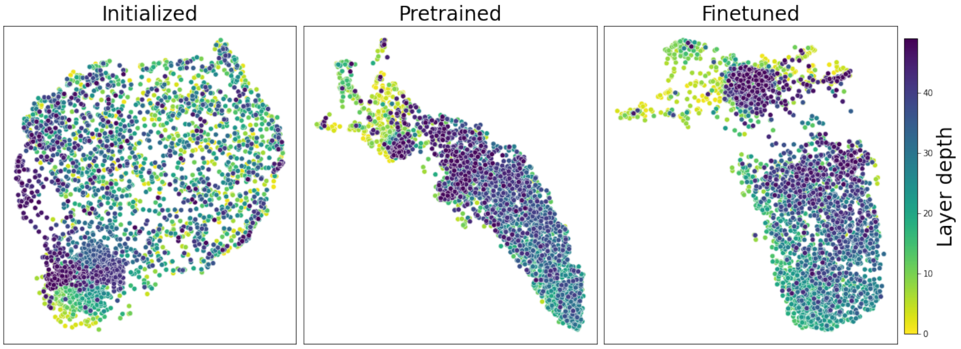

**Figure 2:** Projected features in randomly initialized, pre-trained and fine-tuned model.

### Acknowledgement

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2019-027. Computational resources were supplied by the project ”e-Infrastruktura CZ” (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

## References

[1] Dumitru Erhan, Y. Bengio, Aaron Courville, and Pascal Vincent. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univeristé de Montréal*, January 2009.

[2] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-Excitation Networks. *arXiv:1709.01507 [cs]*, May 2019.

[3] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:1802.03426 [cs, stat]*, September 2020.

[4] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. `https://distill.pub/2017/feature-visualization`, November 2017.