Perceptual Similarity

Agha Mara, Dudau Claudia
Supervisor: Emanuela Haller

Introduction

Perceptual similarity is defined as the assessment of how well one image matches another, which forms a critical component both of models of human visual processing and of many image analysis systems.

Given the new perceptual metric, LPIPS, how will the resulting model perform in comparison to the established cosine distance?

We systematically evaluate deep features across different architectures and tasks and compare them with classic metrics and in relation to the human perceptual judgements.

1) Collect a large-scale perceptual similarity dataset
2) Deep features across training objectives outperform widely-used perceptual metrics (e.g., SSIM)
3) Train new metric (LPIPS) on perceptual judgments

Loss function

Learned Perceptual Image Patch Similarity (LPIPS)

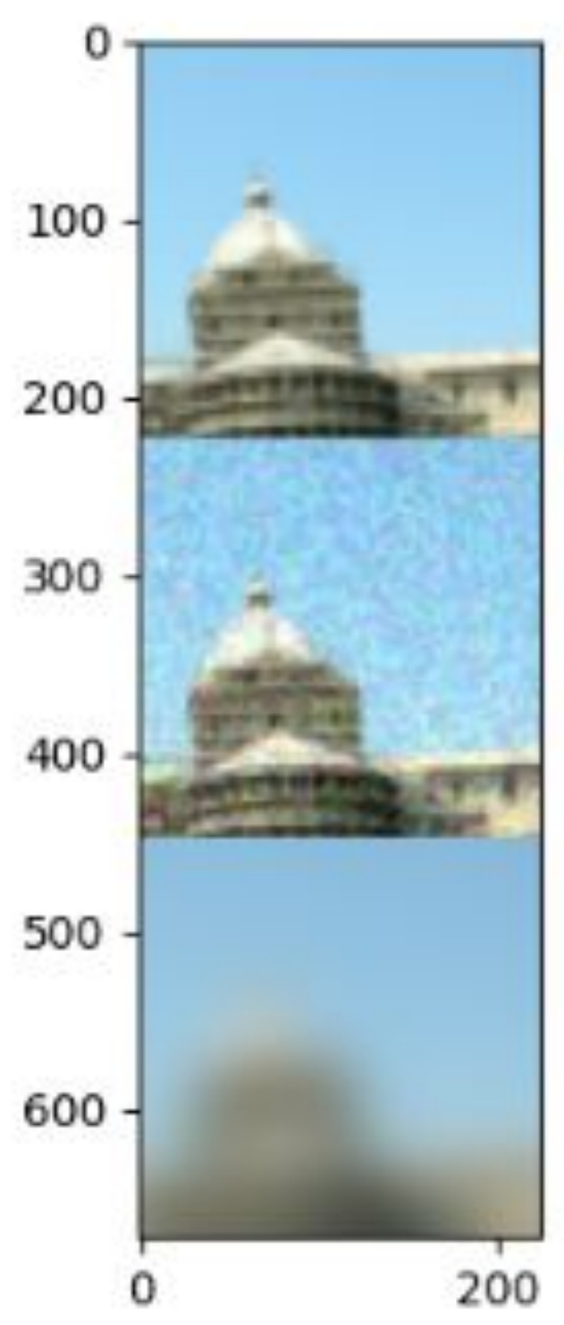
$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

It is based on an existing network as it uses internal activation of networks which are trained for image classification tasks. This amounts to the use of pretrained weights when training the network, to aim for perceptuality.

Dataset

2AFC consists of a patch triplet (1 reference + 2 distorted). Each 2AFC subdirectory contains the following folders:

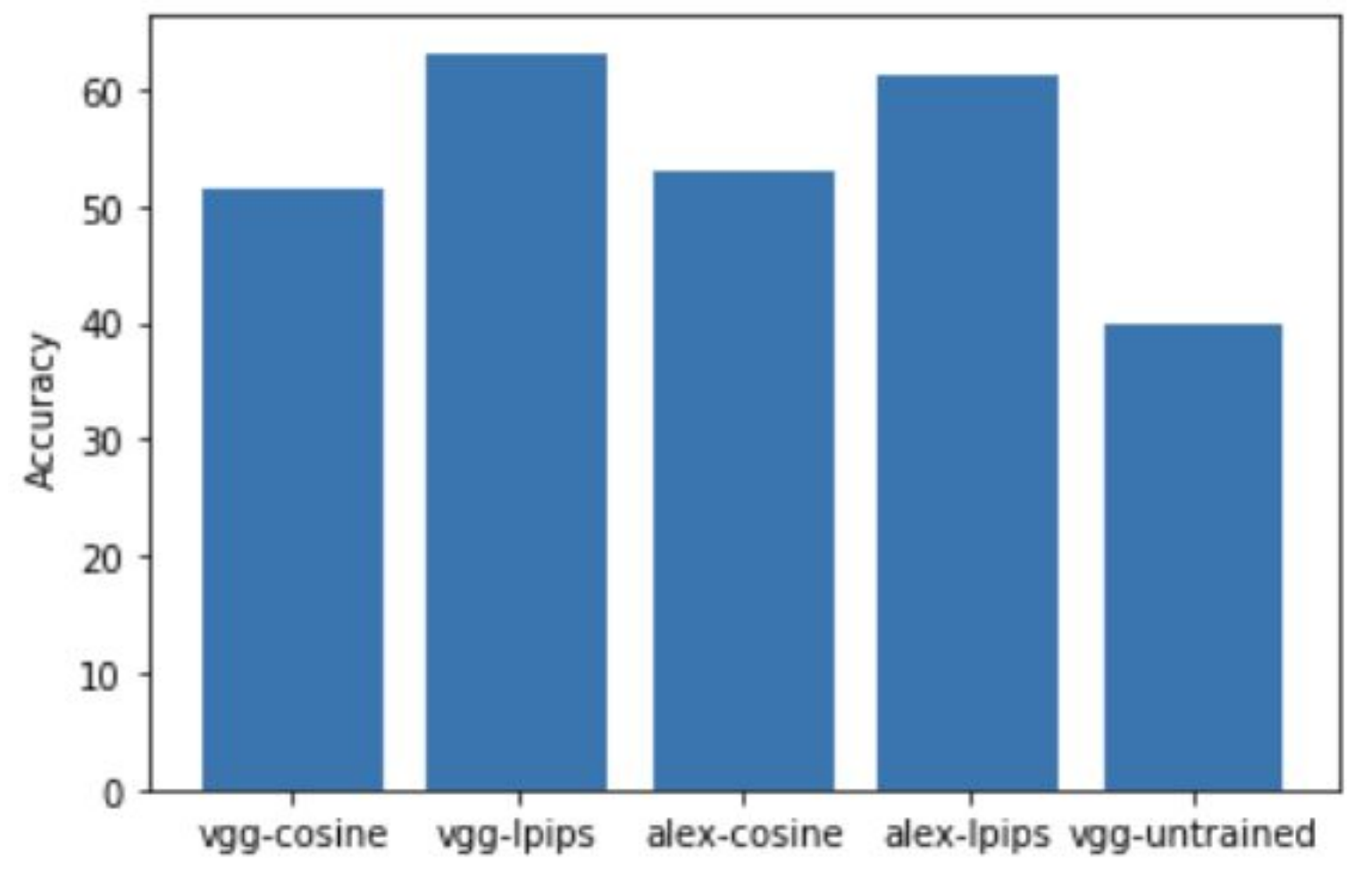
- ref: original reference patches
- p0, p1: two distorted patches
- judge: human judgments - 0 if all preferred p0, 1 if all humans preferred p1



Results

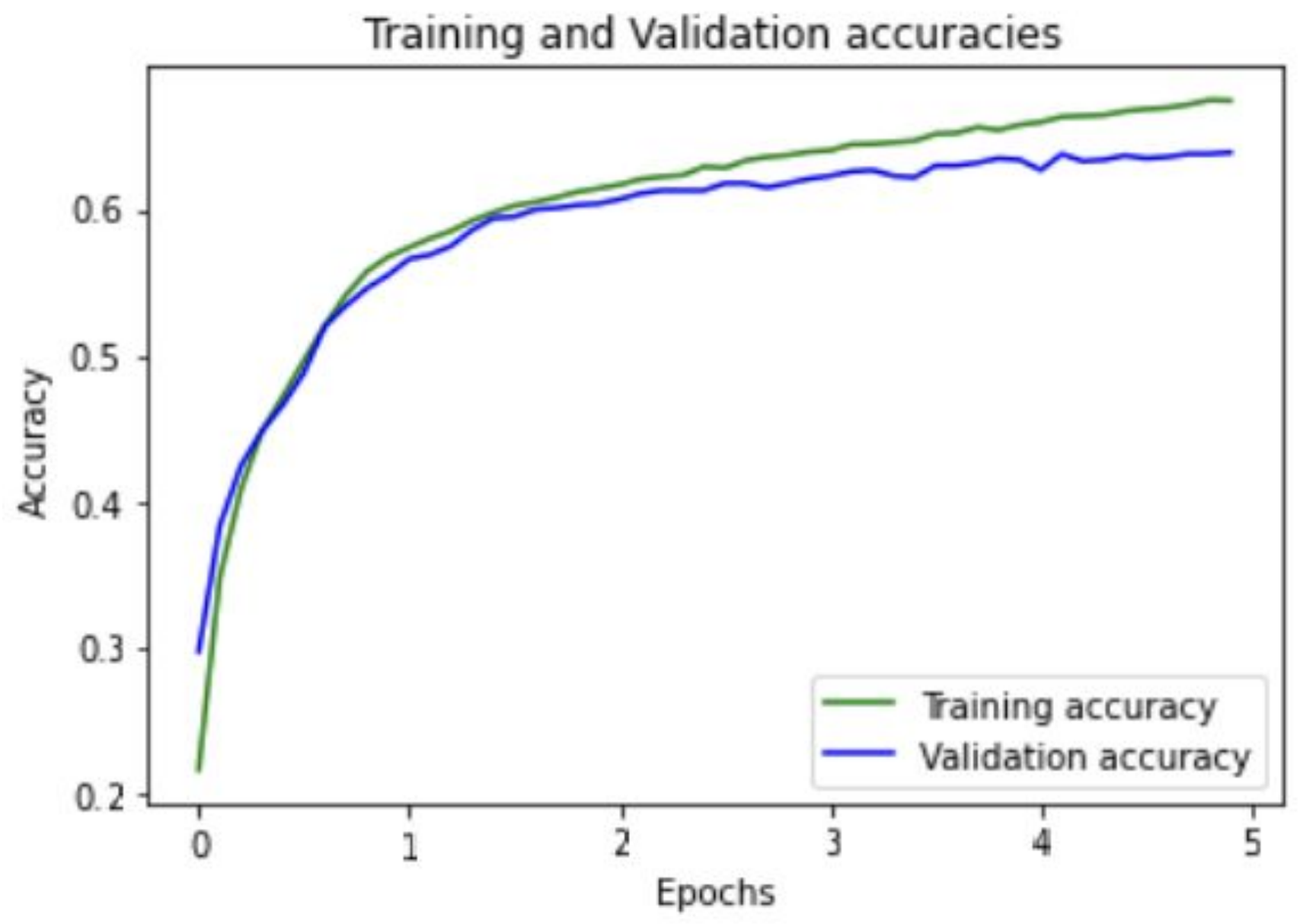
We compared the results from different models and architectures:

- VGG16 pretrained (2 convolutional layers) + cosine distance
- VGG16 pretrained (2 convolutional layers) + LPIPS distance + small fully connected network
- AlexNet pretrained (4 convolutional layers) + cosine distance
- AlexNet pretrained (4 convolutional layers) + LPIPS distance + small fully connected network
- VGG16 untrained network + cosine distance

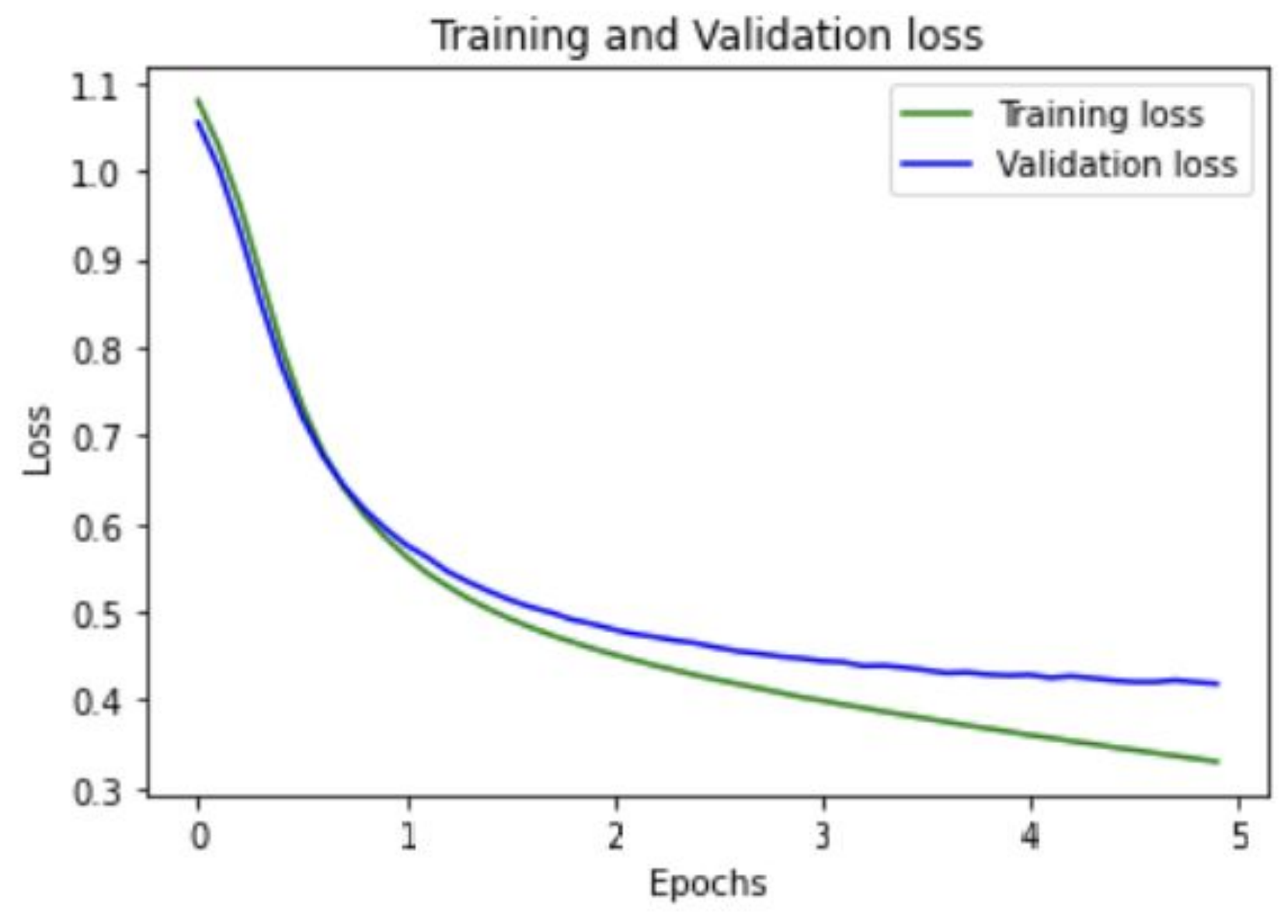


As expected, the models using the LPISP distance outperform the ones using the cosine distance. Another notable result is that the vgg model outperforms the alexnet model when using the LPIPS distance, but, when using the cosine distance, the alexnet model is the one with better results.

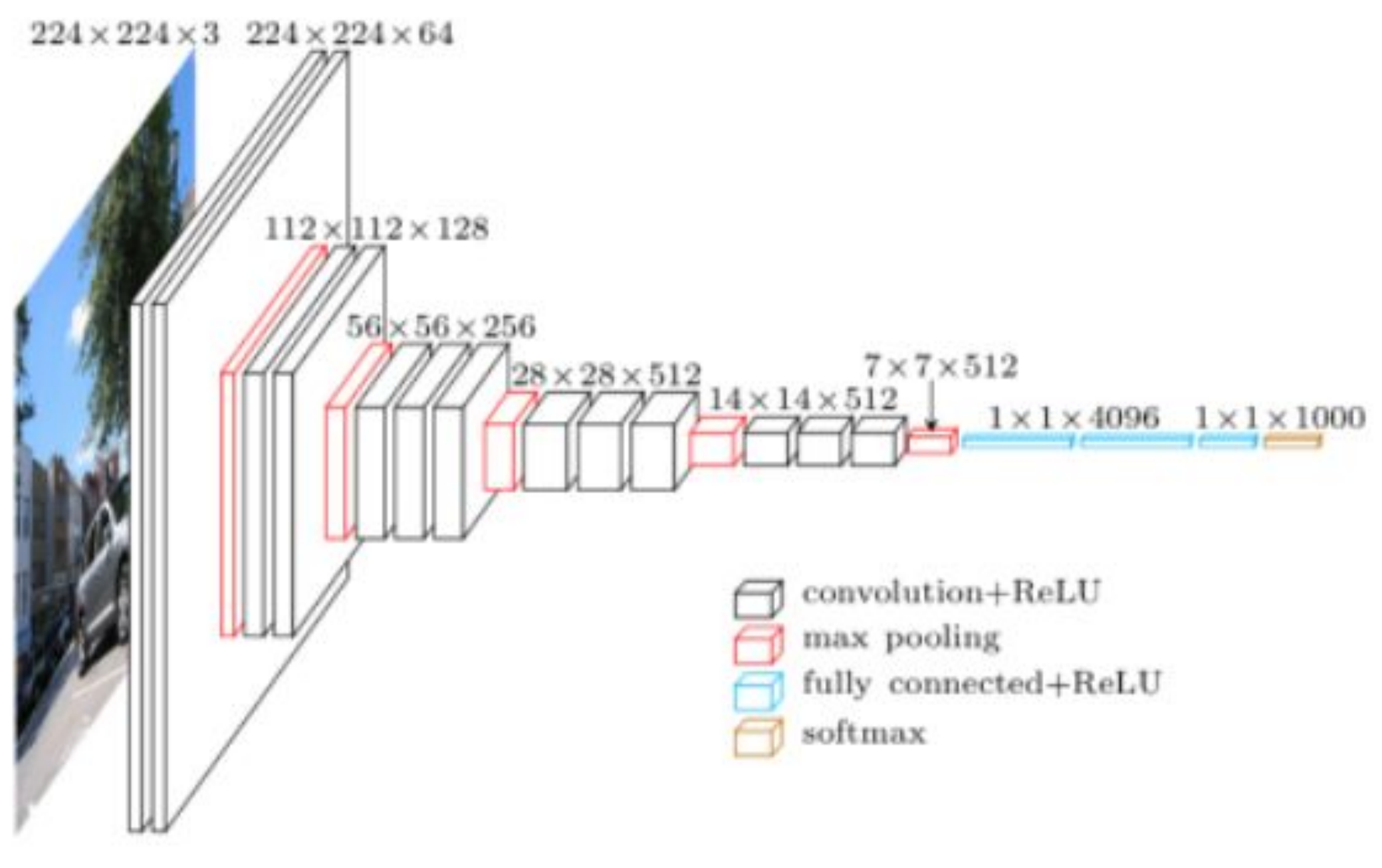
Grafic 1. Accuracy for best model



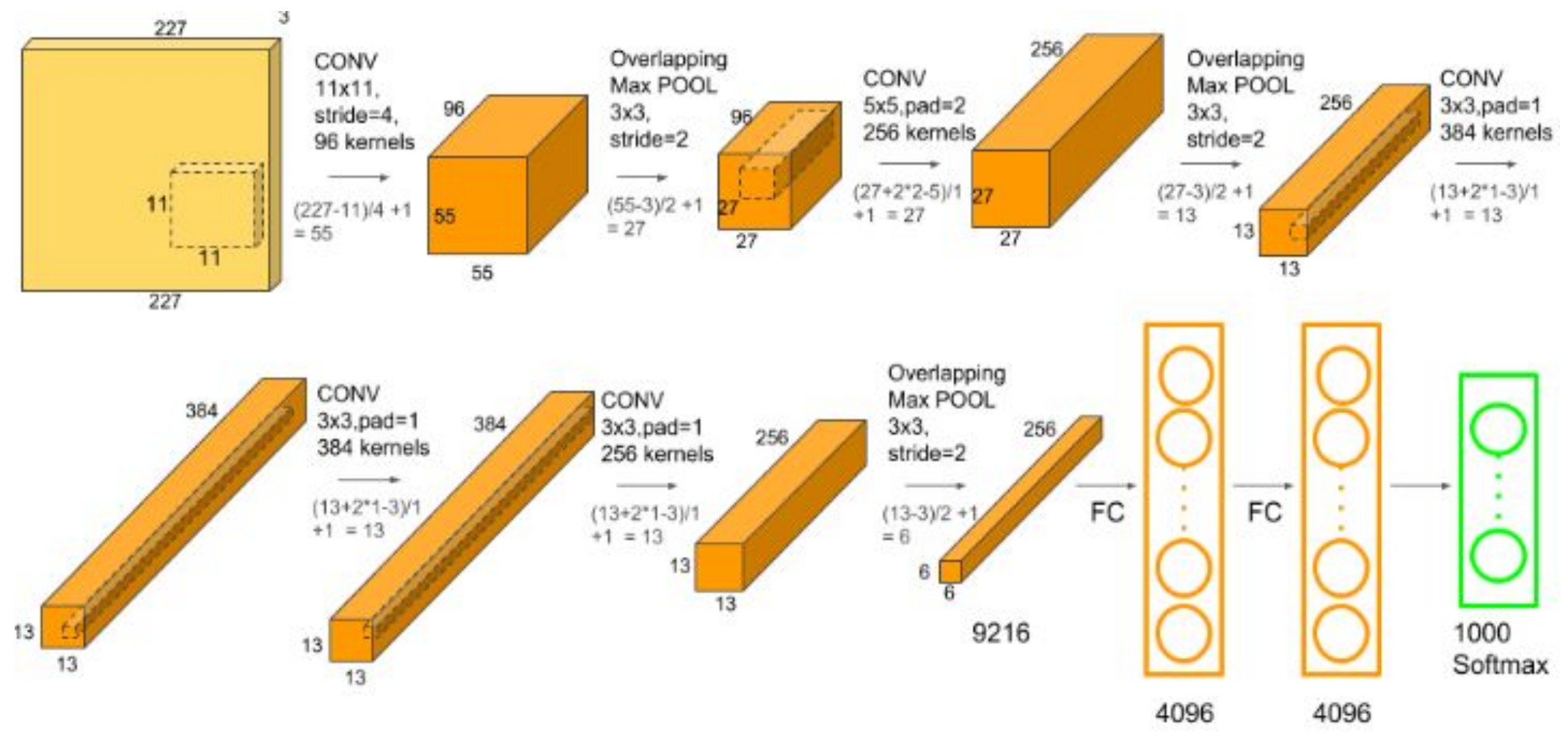
Grafic 2. Loss for best model



Model



Architecture of VGG16



Architecture of AlexNet

Conclusions

We find that deep features outperform all previous metrics by large margins on our dataset. More surprisingly, this result is not restricted to ImageNet-trained VGG features, but holds across different deep architectures and levels of supervision (supervised, self-supervised, or even unsupervised). Our results suggest that perceptual similarity is an emergent property shared across deep visual representations.

Contact Information

mara.agha@s.unibuc.ro
claudia.dudau@s.unibuc.ro

References

➔ https://arxiv.org/pdf/1801.03924v2.pdf
➔ https://github.com/richzhang/PerceptualSimilarity#2-berkeley-adobe-perceptual-patch-similarity-bapps-dataset