Project results

Kamil Jaśkiewicz, Michał Piwowarczyk

January 2026

1 Project goal

The objective of this project was to investigate the extent to which deep learning models may expose sensitive information under a white-box attack scenario. In particular, we focused on reconstruction attacks based on gradient information and embedding optimization techniques. The experiments were conducted using several commonly used datasets, including MNIST, CIFAR, and the Yale Faces dataset.

Additionally, we evaluated the impact of different convolutional neural network architectures on the amount of information leakage. Specifically, we analyzed how the embedding dimensionality and the number of neurons in individual layers influence the susceptibility of a model to white-box attacks and the quality of reconstructed inputs.

2 Results and observations

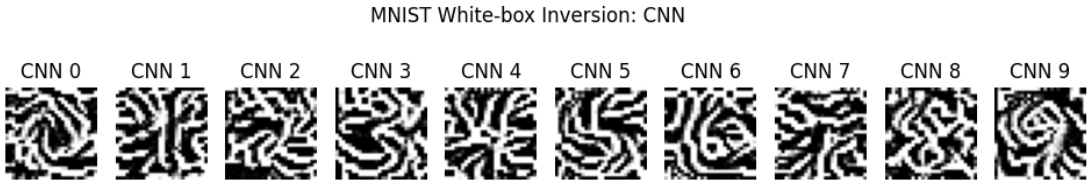


Figure 1: Logit inversion

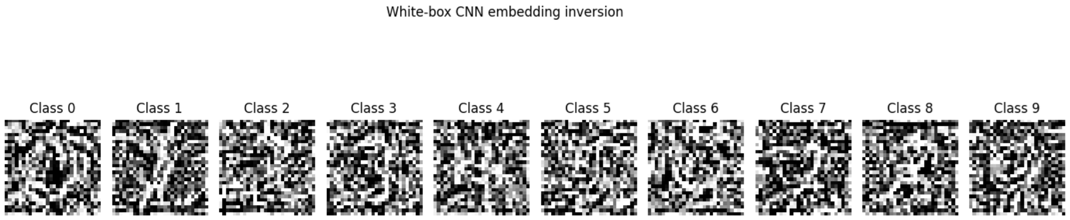


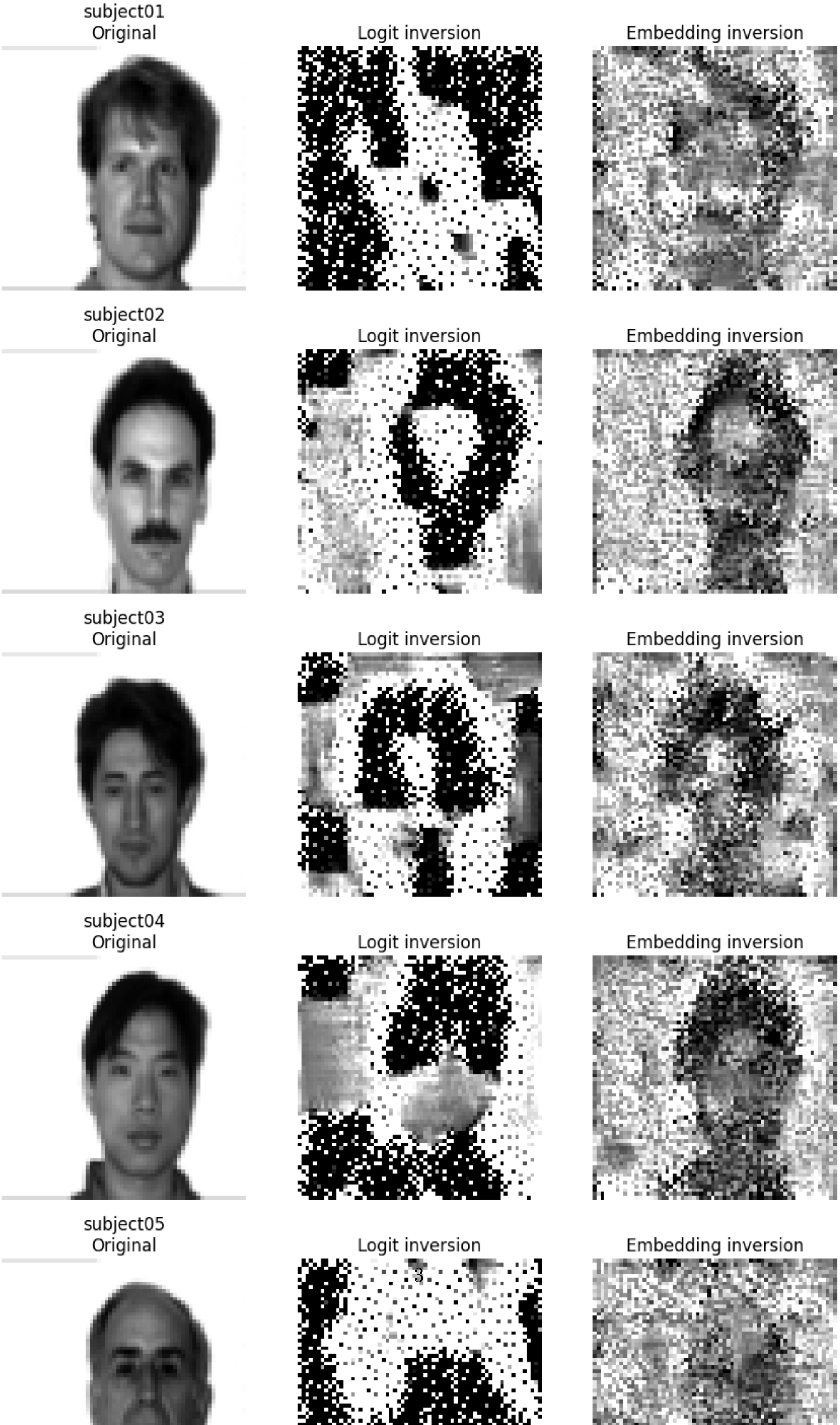
Figure 2: Embedding inversion

At the initial stage of the experiments, we implemented a convolutional neural network that served as the victim model and trained it on the MNIST dataset. The dataset consists of ten classes, each corresponding to a distinct digit from 0 to 9. The primary objective of this experiment was to determine whether representative visual features of each digit class could be recovered using a white-box attack.

The first reconstruction attempt was performed using gradient optimization based on the model logits. The optimization process aimed to generate input samples that maximize the activation of a selected output class. As shown in Figure 1, the resulting reconstructions exhibit recognizable digit-like structures located near the center of the images. The shapes are particularly distinguishable for digits such as *1* and *6*, while the surrounding regions are dominated by high levels of noise.

In the subsequent experiment, the attack was shifted to an earlier stage of the network by performing optimization in the embedding space. The reconstructed samples obtained using this approach present similar overall characteristics; however, the digit shapes are generally less distinct and more difficult to separate from the background noise. This suggests that embeddings, while still containing class-related information, provide a slightly weaker reconstruction signal compared to direct logit-based optimization.

In addition to MNIST, we conducted experiments using the CIFAR-10 dataset. Unlike MNIST, CIFAR-10 contains highly diverse and complex images within each class, such as different types of cars. As a result, the reconstruction attacks produced abstract and noisy patterns that lacked interpretable semantic structure. Due to the limited visual interpretability of the reconstructed samples, we decided to skip further analysis.


subject01
Original
Logit inversion
Embedding inversion
subject02
Original
Logit inversion
Embedding inversion
subject03
Original
Logit inversion
Embedding inversion
subject04
Original
Logit inversion
Embedding inversion
subject05
Original
Logit inversion
Embedding inversion

Finally, we conducted experiments using the Yale Faces dataset, which proved to be the most relevant for the intended use case of this project. The dataset consists of facial images of 15 individuals, captured with varying facial expressions and lighting conditions. Based on observations from earlier experiments, we applied both logit-based and embedding-based inversion attacks.

As shown in the corresponding figures, the obtained results are visually meaningful. Reconstructions based on logits allow for approximate localization of the face within the image and reveal coarse facial contours. In particular, hair shape and general head outline are often clearly visible. However, finer facial details remain difficult to distinguish.

Better results were achieved when performing inversion in the embedding space. In this case, reconstructed images exhibit sharper facial features and more clearly defined structures. Hair and face shape are generally easier to recognize compared to the logit-based approach, indicating that embeddings contain richer information about the input images.

To quantitatively compare different convolutional neural network configurations, we introduced a simple evaluation metric based on mean squared error (MSE) computed between the reconstructed image and the original input. The results were summarized using a confusion matrix, which allowed us to measure how often embedding inversion produced the lowest MSE for the correct class. In our experiments, this occurred in approximately 80% of the cases, suggesting a strong correspondence between reconstructed samples and their original classes.

Additionally, we analyzed the influence of embedding dimensionality and the number of neurons in individual layers on reconstruction quality. As illustrated in the figures below, embedding dimension had the most significant impact on the results. Larger embedding sizes generally led to lower reconstruction error, with performance improvements saturating around an embedding dimension of 128. Interestingly, the number of neurons in the network layers had a relatively minor effect on reconstruction quality, provided that the embedding dimension was sufficiently large.

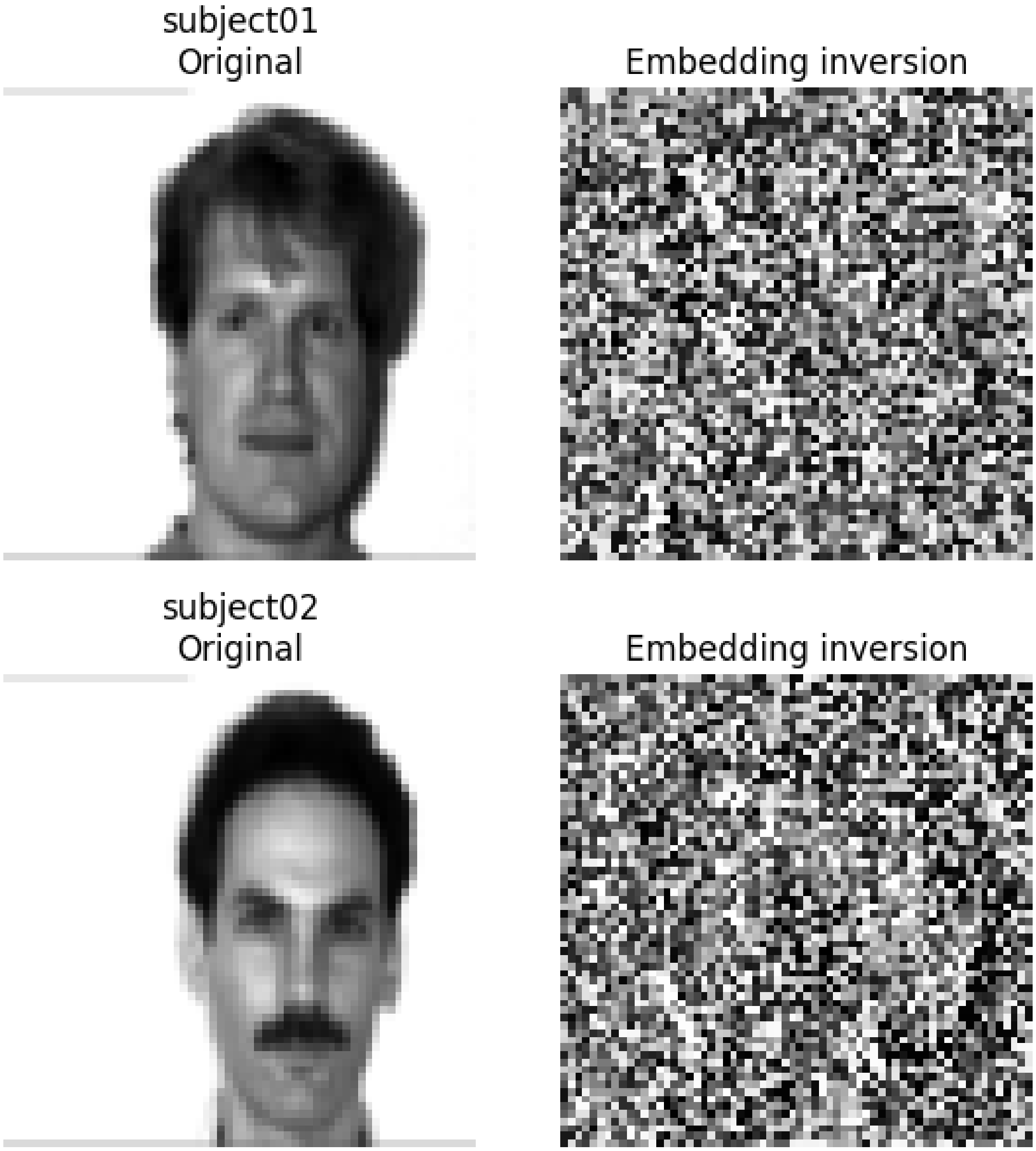


Figure 3: Embedding dim: 8

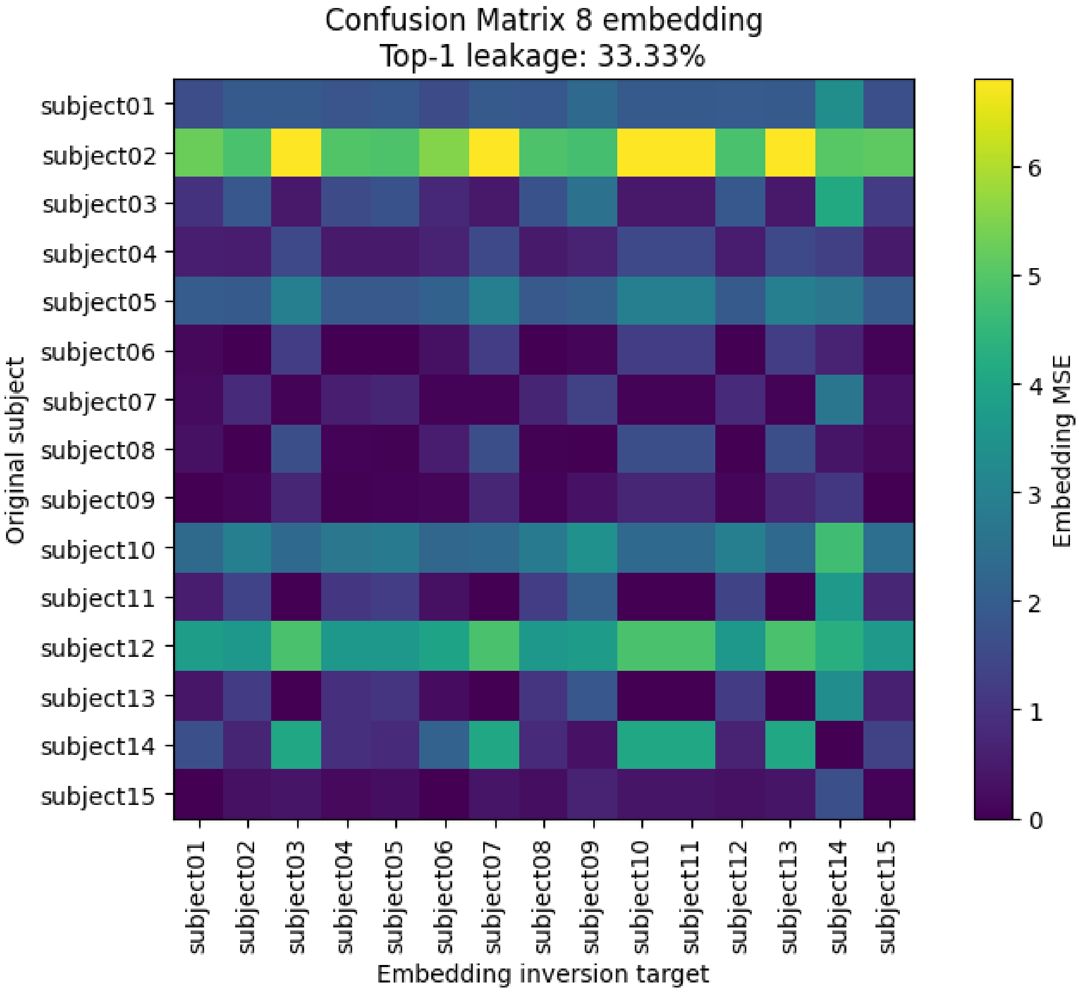


Figure 4: Embedding dim: 8

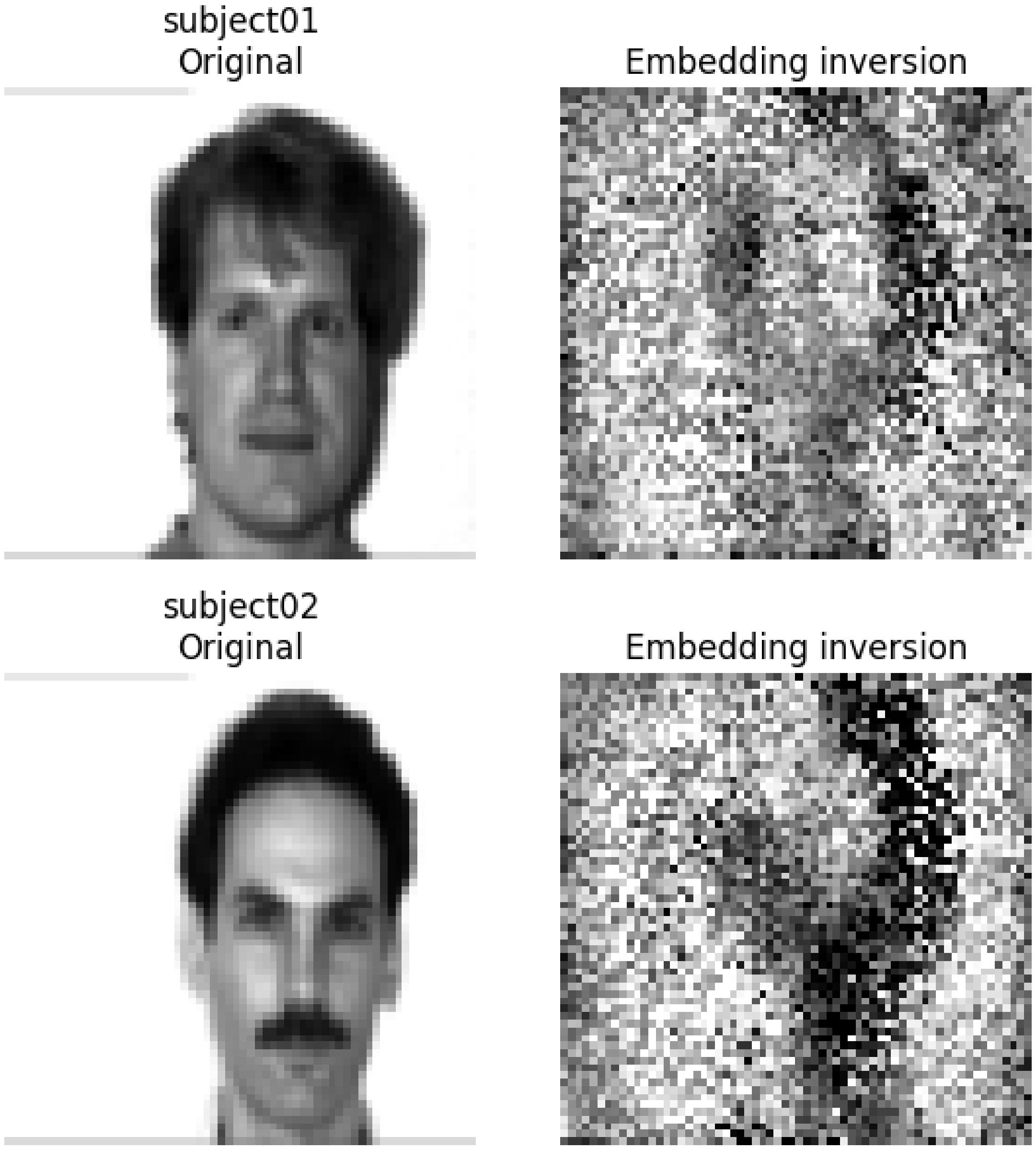


Figure 5: Embedding dim: 32

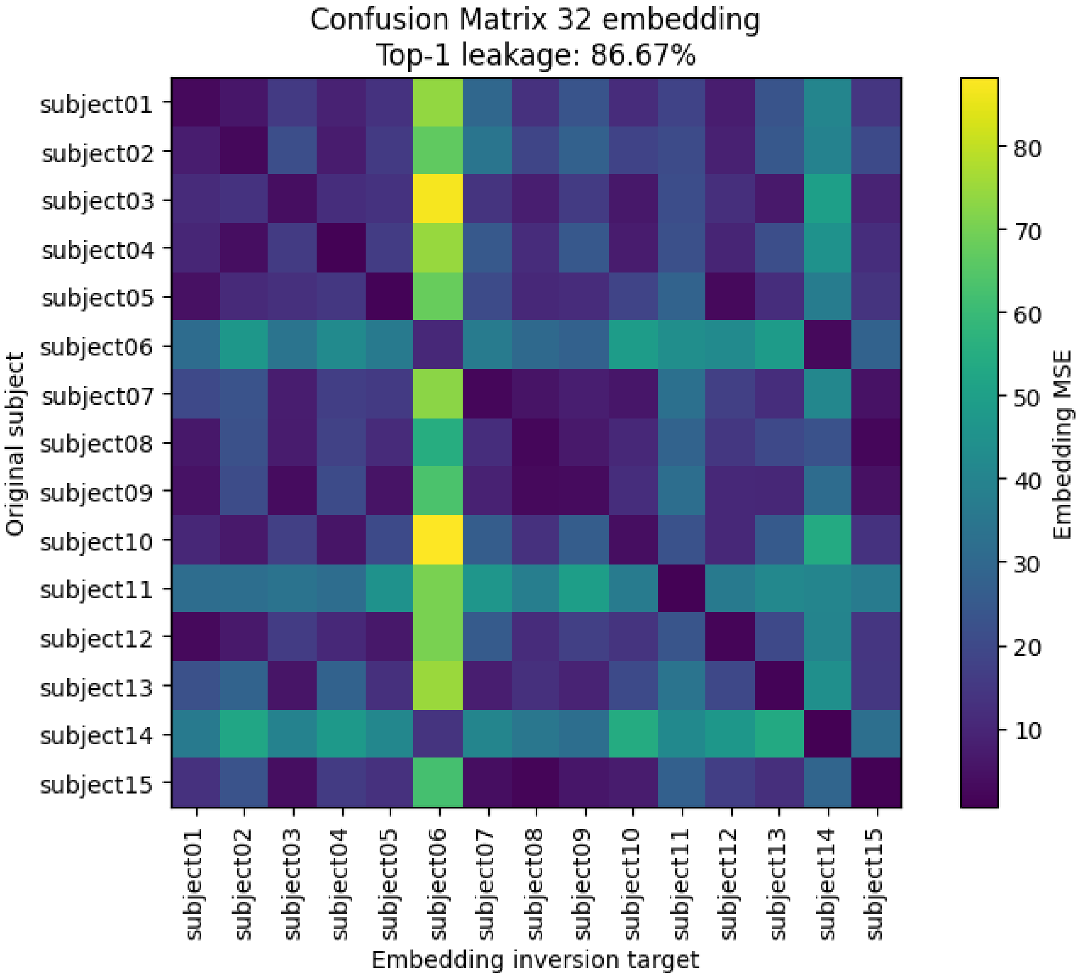


Figure 6: Embedding dim: 32

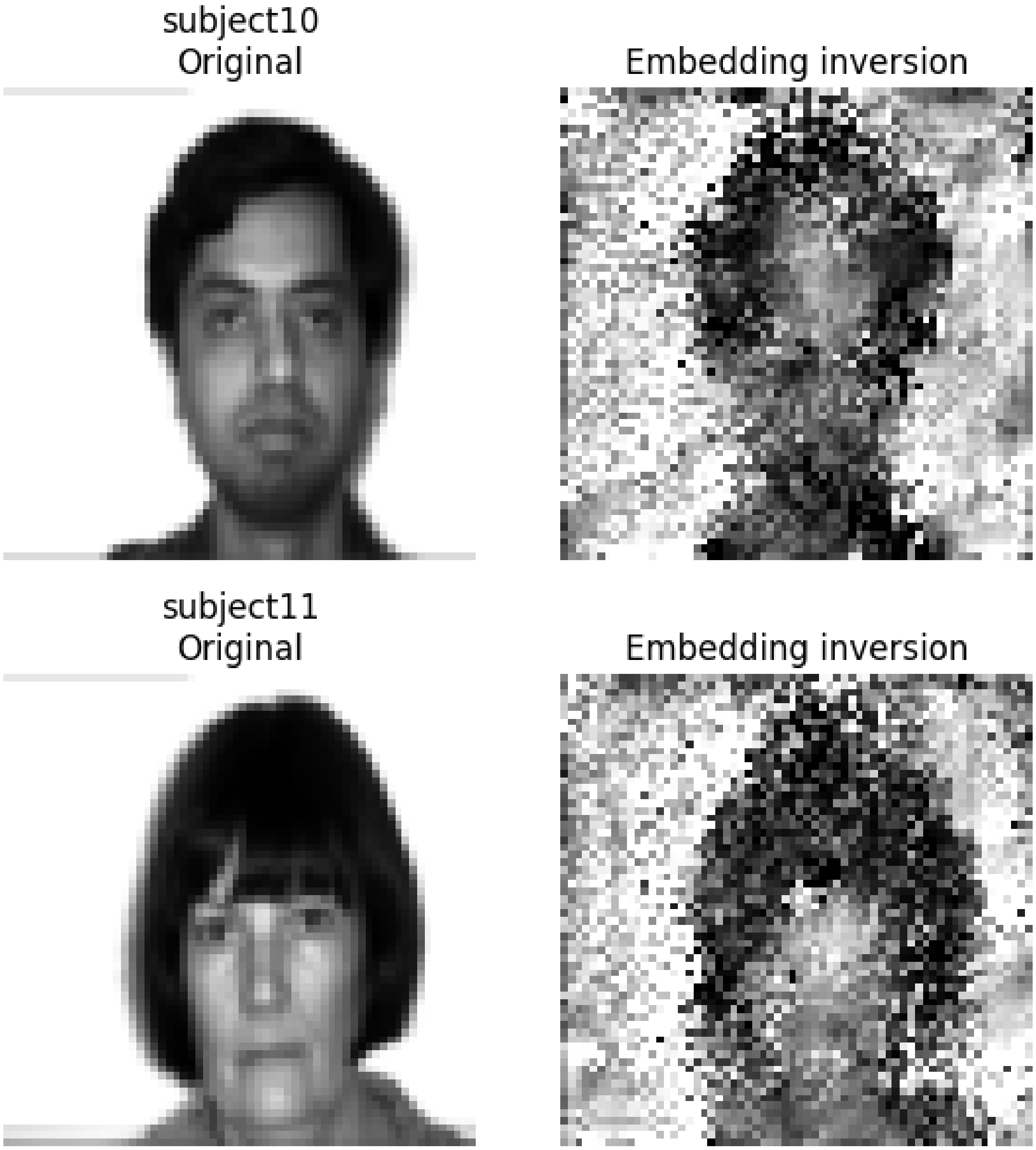


Figure 7: Embedding dim: 128

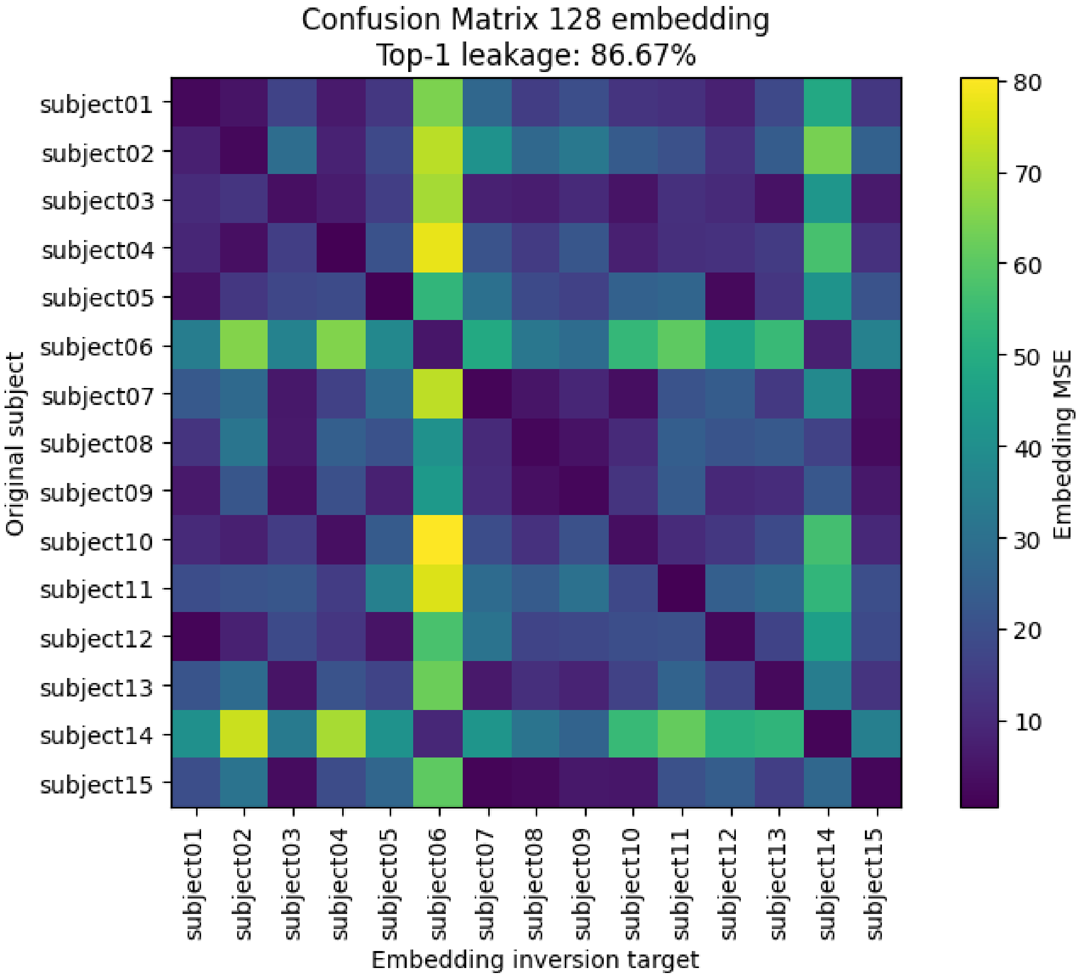


Figure 8: Embedding dim: 128

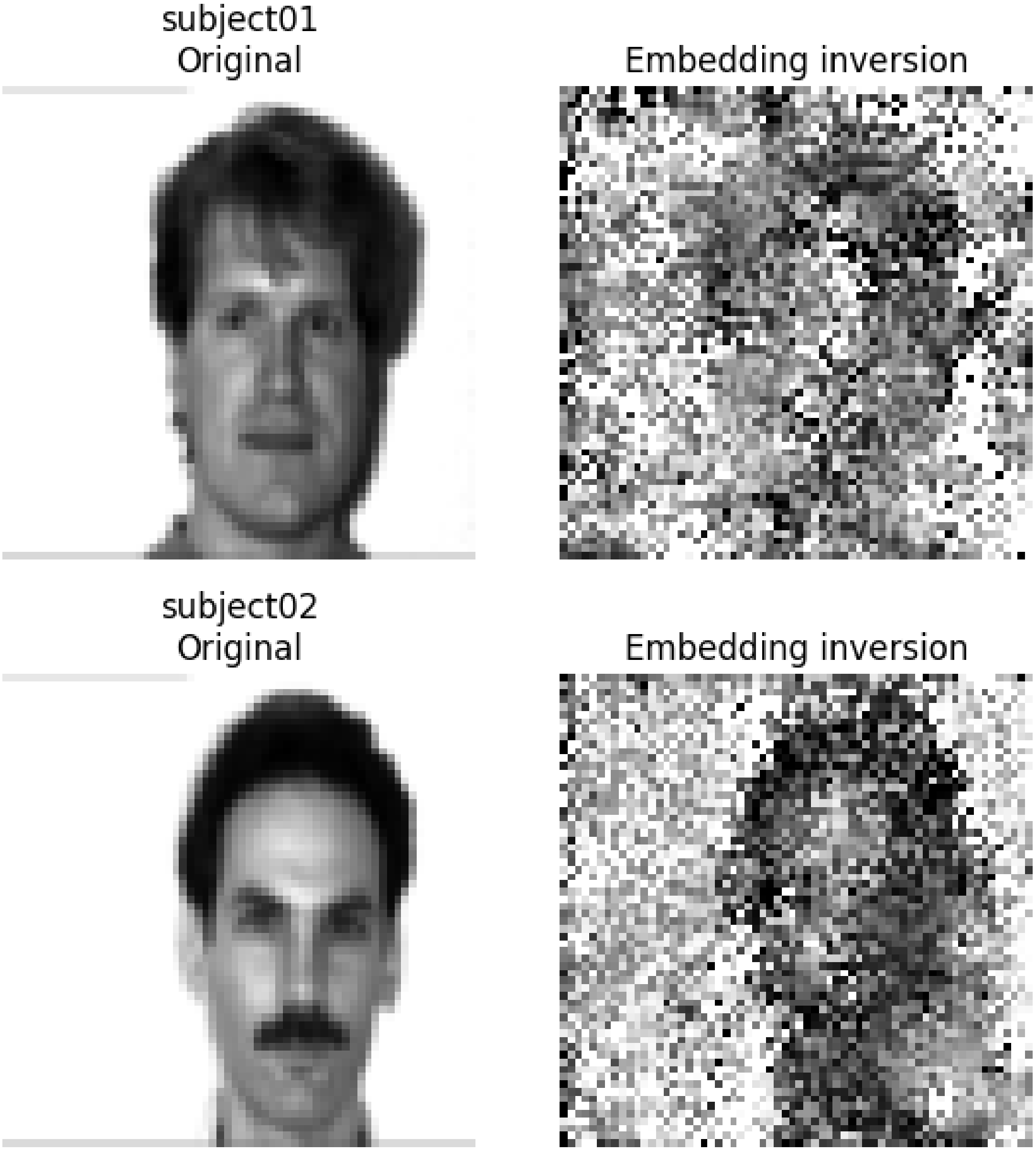


Figure 9: Neurons in layers: 32x64x128

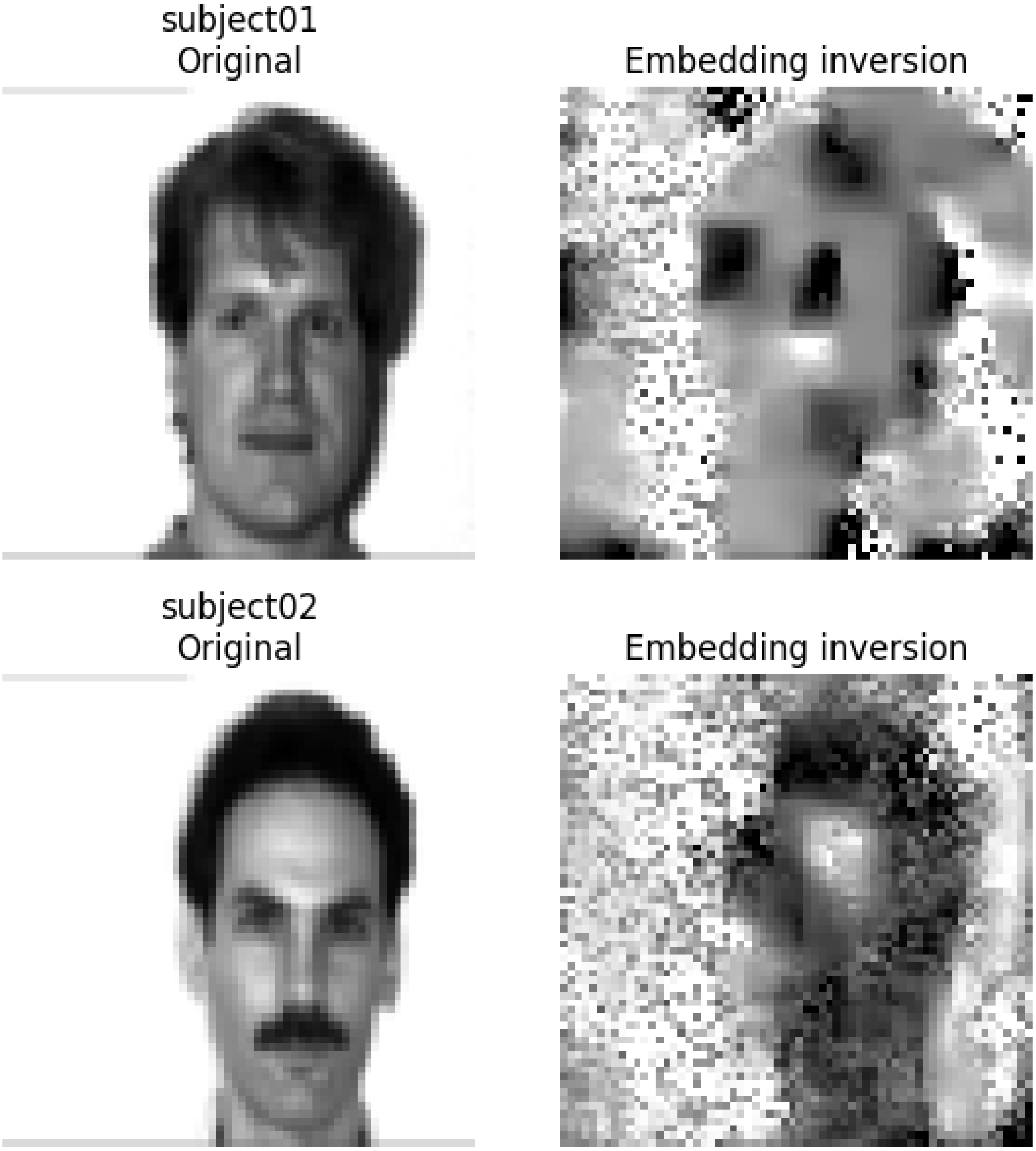


Figure 10: Neurons in layers: 128x256x512