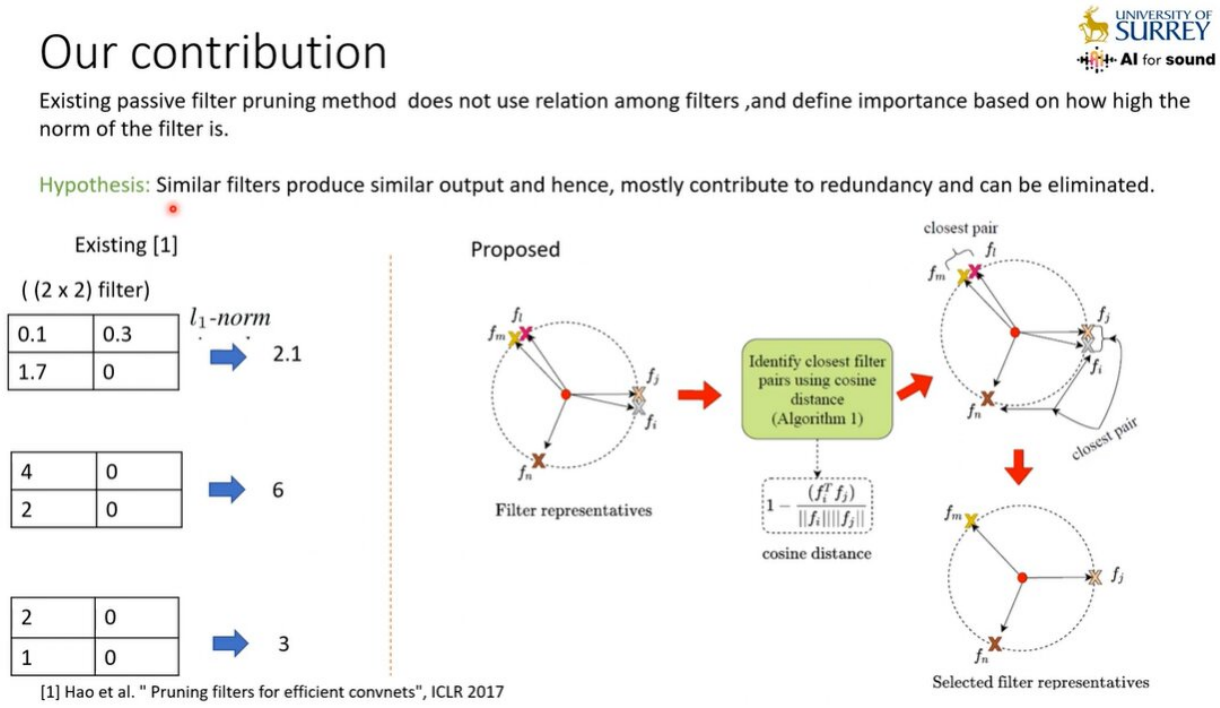
Case study: Efficient audio-based convolutional neural networks via filter pruning

March 20 2023, by Dr. Arshdeep Singh



Credit: University of Surrey

Dr. Arshdeep Singh, a machine learning researcher in sound with Professor Mark D. Plumbley as a part of "AI for sound" (AI4S) project within the Centre for Vision, Speech and Signal Processing (CVSSP), has been focusing on designing efficient and sustainable artificial

intelligence and machine learning (AI-ML) models. Their current study has been accepted to the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, held in Greece, June 4–10.

Recent trends in artificial intelligence (AI) employ convolutional neural networks (CNNs) that provide remarkable performance compared to other existing methods. However, the large size and high computational cost of CNNs is a bottleneck to deploying CNNs on resource-constrained devices such as smartphones.

Moreover, training CNNs for several hours leads to emitting more CO_2. For instance, a computing device (NVIDIA GPU RTX-2080 Ti) used to train CNNs for 48 hours generates the equivalent CO_2 emitted by an average car driven for 13 miles. For estimating CO_2, they researchers used an openly available tool.

Therefore, the researchers aimed to compress CNNs to:

1. Reduce the computational complexity for faster inference.
2. Reduce memory footprints for using underlying resources effectively.
3. Reduce the number of computations during the training stage of CNNs by analyzing how many training examples are sufficient in the fine-tuning process of the compressed CNNs to achieve a similar performance to that obtained using all training examples for uncompressed CNNs.

The solution

One of the directions to compress CNNs is by "pruning," where the unimportant filters are explicitly removed from the original network to build a compact or pruned network. After pruning, the pruned network is fine-tuned to regain the performance loss.

This study proposed a cosine distance-based greedy algorithm to prune similar filters in filter space for openly available CNNs designed for audio scene classification. Further, the researchers improved the efficiency of the proposed algorithm by reducing the computational time in pruning.

They found that the proposed pruning method reduces the number of computations per inference by 27%, with 25% less memory requirements, with less than a 1% drop in accuracy. During fine-tuning of the pruned CNNs, a reduction of training examples by 25% gave a similar performance as that obtained using all examples. They made the proposed algorithm openly available for reproducibility and provided a video presentation explaining the methodology and results from our published work.

In addition, they improved the computational time of the proposed pruning method by three times without degrading performance.

Provided by University of Surrey