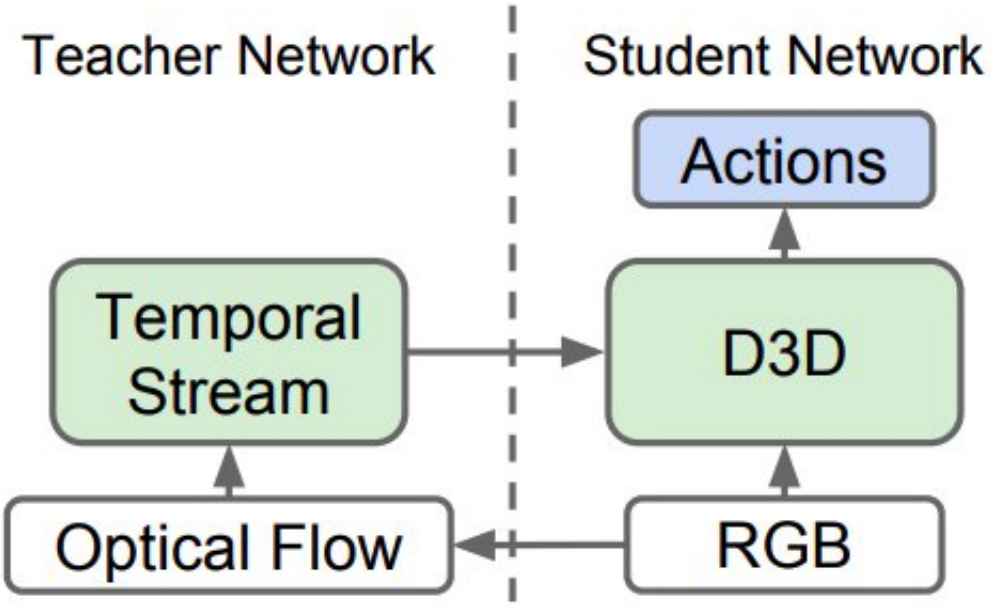
Distilled 3-D (D3D) networks for video action recognition

January 9 2019, by Ingrid Fadelli



Distilled 3D Networks (D3D). The researchers trained a 3D CNN to recognize actions from RGB video while distilling knowledge from a network that recognizes actions from optical flow sequences. During inference, only D3D is used. Credit: Stroud et al.

A team of researchers at Google, the University of Michigan and Princeton University have recently developed a new method for video action recognition. Video action recognition entails identifying particular actions performed in video footage, such as opening a door, closing a door, etc.

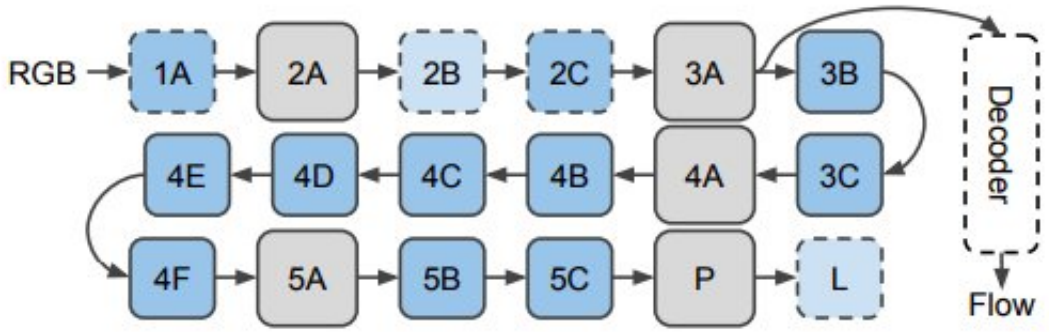
Researchers have been trying to teach computers to recognize human and non-human actions on video for years. Most state-of-the-art video action recognition tools employ an ensemble of two neural networks: the spatial stream and the temporal stream.

In these approaches, one neural network is trained to recognize actions in a stream of regular images based on appearance (i.e. the 'spatial stream') and the second network is trained to recognize actions in a stream of motion data (i.e. the 'temporal stream'). The results attained by these two networks are then combined to achieve video action recognition.

Although empirical results achieved using 'two-stream' approaches are great, these methods rely on two distinct networks, rather than a single one. The aim of the study carried out by the researchers at Google, the University of Michigan and Princeton was to investigate ways of improving this, in order to replace the two streams of most existing approaches with a single network that learns directly from the data.

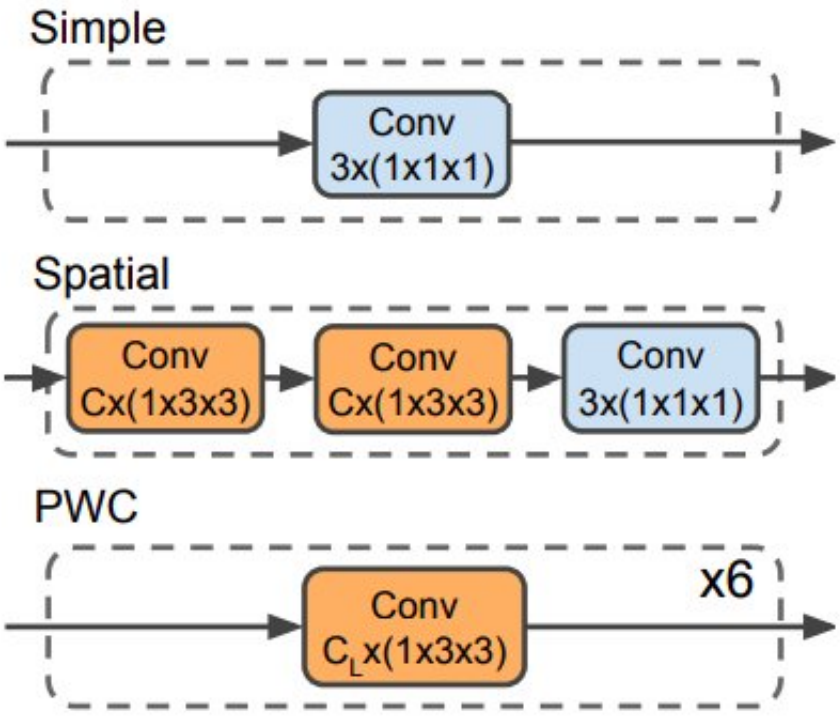
In most recent studies, both spatial and temporal streams consist of 3-D convolutional neural networks (CNNs), which apply spatiotemporal filters to the video clip before attempting classification. Theoretically, these applied temporal filters should allow the spatial stream to learn motion representations, hence the temporal stream should be unnecessary.

In practice, however, the performance of video action recognition tools improves when an entirely separate temporal stream is included. This suggests that the spatial stream alone is unable to detect some of the signals captured by the temporal stream.



The network used to predict optical flow from 3D CNN features. The researchers apply the decoder at hidden layers in the 3D CNN (depicted here at layer 3A). This diagram shows the structure of I3D/S3D-G, where blue boxes represent convolution (dashed lines) or Inception blocks (solid lines), and gray boxes represent pooling blocks. Layer names are the same as those used in Inception. Credit: Stroud et al.

To examine this observation further, the researchers investigated whether the spatial stream of 3-D CNNs for video action recognition is indeed lacking motion representations. Subsequently, they demonstrated that these motion representations can be improved using distillation, a technique for compressing knowledge in an ensemble into a single model.

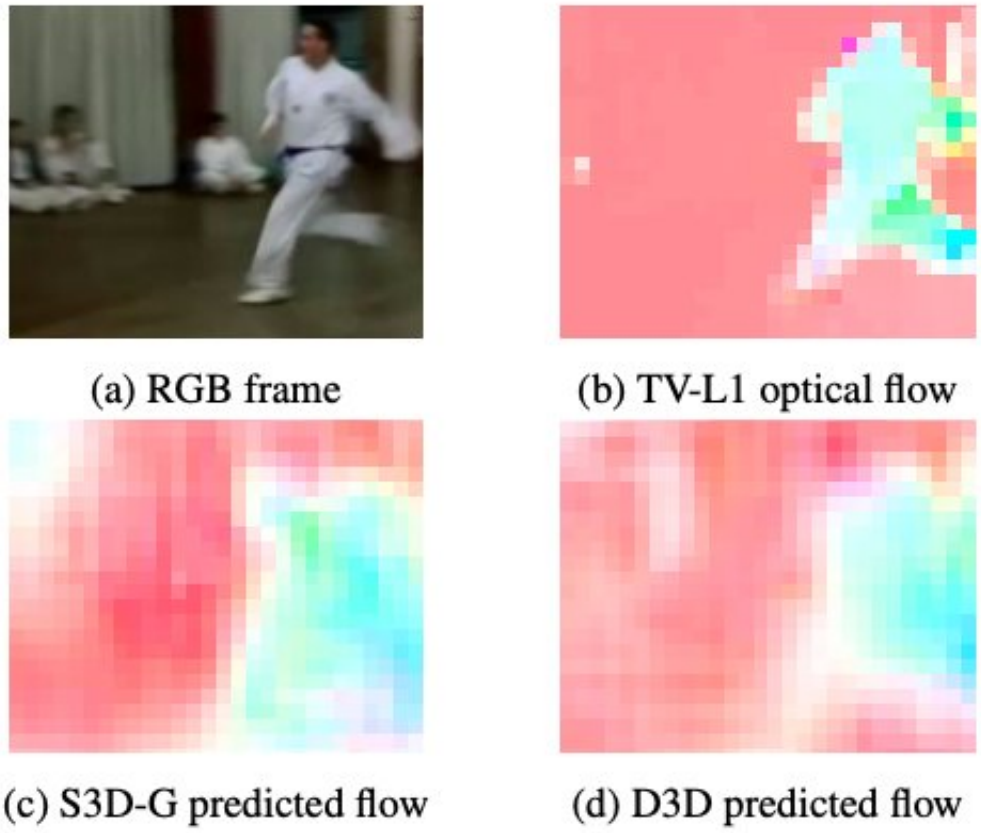


Three decoders used to predict optical flow. The PWC decoder resembles the optical flow prediction network from PWC-net. No decoder makes use of temporal filters. Credit: Stroud et al.

The researchers trained a 'teacher' network to recognize actions given the motion input. Then, they trained a second 'student' network, which is only fed the stream of regular images, with a dual objective: do well at the action recognition task and mimic the output of the teacher network. Essentially, the student network learns how to recognize based on both appearance and motion, better than the teacher and as well as the larger and more cumbersome two-stream models.

Recently, a number of studies also tested an alternative approach for video action recognition, which entails training a single network with two different objectives: performing well at the action recognition task and directly predicting the low-level motion signals (i.e. optical flow) in

the video. The researchers found that their distillation method outperformed this approach. This suggests that it is less important for a network to effectively recognize the low-level optical flow in a video than it is to reproduce the high-level knowledge that the teacher network has learned about recognizing actions from motion.



Examples of optical flow produced by S3DG and D3D (without fine-tuning) using the PWC decoder applied at layer 3A. The color and saturation of each pixel corresponds to the angle and magnitude of motion, respectively. TV-L1 optical flow is displayed at 28 × 28px, the output resolution of the decoder. Credit: Stroud et al.

The researchers proved that it is possible to train a single-stream neural network that performs as well as two-stream approaches. Their findings suggests that the performance of current state-of-the-art methods for video action recognition could be attained using approximately 1/3 the compute. This would make it easier to run these models on compute-constrained devices, such as smartphones, and at larger scales (e.g. to identify actions, such as 'slam dunks', in YouTube videos).

Overall, this recent study highlights some of the shortcomings of existing video action recognition methods, proposing a new approach that involves training a teacher and a student network. Future research, however, could try to attain state-of-the-art performance without the need for a teacher network, by feeding the training data directly to the student network.

More information: D3D: Distilled 3-D networks for video action recognition. arXiv:1812.08249 [cs.CV]. arxiv.org/abs/1812.08249

Distilling the knowledge in a neural network. arXiv:1503.02531 [stat.ML]. arxiv.org/abs/1503.02531