Review on “Temporal Relational Reasoning in Video”

Authors: B. Zhou, A. Andonian, A. Torralba

# Short Summary

The goal of this paper is to reason about temporal dependencies in video frames at multiple time scales. The authors propose Temporal Relation Net (TRN) as a means to improve historical baselines that exceed at action detection but cannot reason about transformations between frames. The module is designed to be extensible enough to be inserted into any CNN architecture as a plug-and-play module.

The architecture involves using different frame relation modules that analyze n-frame relations by sampling from the video. For simplicity, Inception pre-trained on ImageNet is used as the base model and tests are performed with and without the TRN module. The model can capture the essence of activity given only a few frames from an action. Using a larger number of frames generally improves accuracy and ordered frames perform significantly better than shuffled ones suggesting that the temporal order plays an important role.

The model achieves state-of-the-art results on the Something-Something dataset for human-interaction recognition (33.60 top1 % accuracy) and the Jester dataset for hand gesture recognition (94.78). The model also achieves 25.2 mAP on the Charades dataset for daily activity recognition which exceeds previous state of the art. A t-SNE visualization shows that the model is learning to separate actions into distinct regions in space; furthermore, similar actions are closer together suggesting that the model is learning the meaning behind these actions to some degree.

# Main Contributions

1. Propose a Temporal Relation Network (TRN) to learn temporal dependencies between video frames at multiple time scales.
2. Evaluated on Something Something, Jester, and Charades where previous works do not perform well due to lacking reasoning
3. Visualized what the model is learning to show that they are in fact learning relations

# High-Level Evaluation of Paper

This paper placed little focus on theory and spent most of its length discussing interesting experimental results. As such, the way the model is leveraging temporal structure isn’t quite clear, but the extensive results demonstrate that it is in fact succeeding at the proposed task. Given that the model achieves state of the art on many distinct datasets, an interesting argument can be made about the importance of temporal structure for a reasoning task. That being said, I do not agree with the author’s claim that the model is learning “common sense”; rather, it is learning the relationships between frames that correspond to a particular action.

# Discussion on Evaluation Methodology

With an emphasis on results, the paper provides five tables that compare results on the various tested datasets, graphs and accompanying visualizations. The tables are easy to read, the metrics used seem to be the standard for the particular datasets, and thus the results are easy to interpret. I appreciate how most results are kept on a single page, including validation and test results, allowing one to quickly get a sense that the proposed model achieves state-of-the-art and by what margin it exceeds baselines. The charts in particular are also helpful for understanding the impact of the number of frames and their ordering on recognition accuracy.