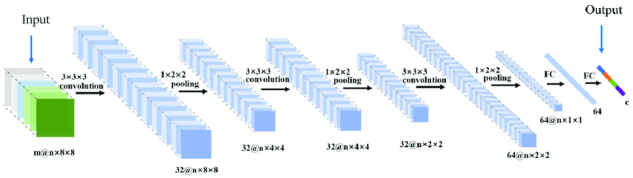
Explaining Temporal Information in Activity Recognition for Situational Understanding

Liam Hiley (Cardiff), **Alun Preece** (Cardiff), **Yulia Hicks** (Cardiff), **David Marshall** (Cardiff) **Supriyo Chakraborty** (IBM US), **Prudhvi Gurram** (ARL)

Objectives

- Better Understand representation of motion in spatio-temporal models
- Expose contribution to a model's decision by salient motion in a video as proposed in [1].
- Improve interpretability of spatio-temporal model explanations

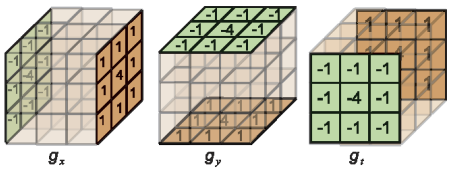


Technical Challenges

- Explanation methods for spatio-temporal models are designed with image recognition in mind.
- Spatio-temporal models such as 3D CNNs view space and time as one fused input.
- Getting such an explanation method to explain contribution in one dimension to the decision of such a model is outside of the intended use of both systems.

Approaches

- In order to measure relevant motion, we first define such motion as that which causes the moving object to gain or lose relevance over time as that motion begins or ends.
- We exploit the image-processing edge detector known as the Sobel operator, to measure the discrete derivative of the explanation in the third dimension.
- We then detect this gain or loss by measuring the change in relevance over time and use sharp change to select relevant motion out of the explanation.

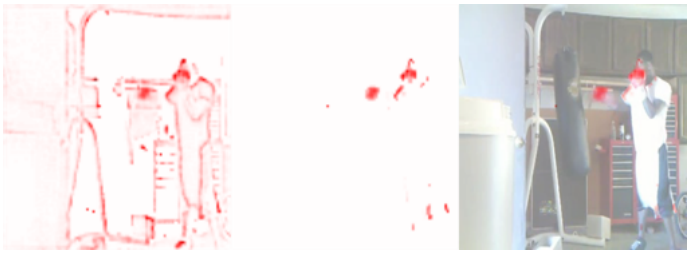


Military & Coalition Relevance

- Provides an improved method for generating transparent explanations from black box classifiers, e.g., where a user from one coalition partner is using a classifier provided by a different partner.
- Focusing on the temporal aspects of an explanation is particularly important in situational awareness in rapidly changing environments.

Results

- When generating explanations using our method on UCF-101 activities, we found that it greatly reduced the regions of relevance down to salient motion.
- The regions selected were ones that would also make sense from a perspective of human intuition.



- In this example for boxing, our method (centre, right) selects the boxer's fists. But filters out much of the background noise that the explanation method (left) picks up as relevant such as the shelving and bag frame as well as spatially relevant features such as the bag itself.
- We use the knowledge that these regions change in relevance over time, to infer that the model recognises the motion of the fist during a punch, as relevant to boxing.
- Note to properly represent the selection of relevance, we have visualised the relevance itself (left, centre). For a more interpretable explanation, we have overlaid this on the input frame for reference (right).

Summary & Future Work

- We are able to select from an explanation, relevant regions within that explanation, selected for their motion. Previously obscured information.
- This allows us to better understand the model's representation of motion, and how much it plays a part in its decision.
- We hope to extend this work to additional functionalities, such as frequency vs. time in the spectral audio space.

Publication(s) & Impact

- [1] Hiley, L.; Preece, A.; Hicks, Y.; Marshall, D.; and Taylor,H. 2019. Discriminating spatial and temporal relevance in deep Taylor decompositions for explainable activity recognition. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence (XAI)* https://arxiv.org/abs/1908.01536v2.