Initial Plan

Explainable Artificial Intelligence for Multi-Sensor Information Fusion

CM3203 One Semester Individual Project
40 credits

Author:
Jack Furby

Supervisor:
Prof Alun Preece

Contents

1 Project Description

In recent years artificial intelligence (AI) has seen large advances mainly regarding machine learning (ML) using deep neural networks (DNN) in prediction and classification tasks. The issue with these tasks, however, is they do not represent the real world which is multimodal. This is to say that an environment includes multiple modalities. A human, for instance, has five senses; touch, sight, hearing, smell, and taste. In order for an AI agent to understand the world as we do, it will also need multimodal capabilities (Baltrušaitis, Ahuja, and Morency 2019). Fusion is one aspect of adding multimodal capabilities to an AI. It is the joining together of two or more modalities which are then used for prediction or classification.

DNN is not perfect and has a few limitations. One of these is the fact that they are opaque, meaning that by default it is hard to understand the reason why a particular prediction or classification occurred. This means a user may have a hard time trusting an output produced. Explainable artificial intelligence (XAI) techniques have provided ways of examining these models with visuals, text, examples, and local explanations (Arrieta et al. 2019). One or more of these can be used for explanations, but the problem of which modality to use has largely been missed in current research (Braines, Preece, and Harborne 2018). This problem will only become more difficult when multiple modalities of input data are considered as each one may have a different best fit for the explainability technique.

A couple of researchers in the Crime and Security Research Institute within Cardiff University who have developed an explainability technique for mid-level fusion networks (fig.1) called Video / Audio Discriminative Relevance (VADR). This applies discriminative relevancy to audio and video which is then displayed as a four tuple saliency map (Taylor et al. 2019). Their implementation uses off-the-shelf models for audio and video classification.

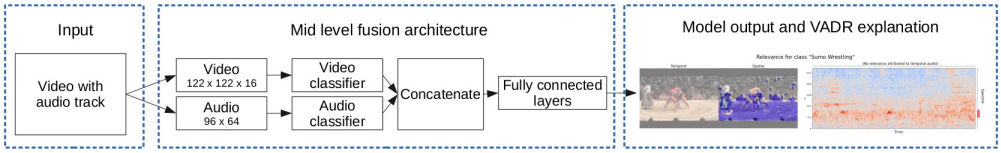

Figure 1: The model used by Taylor et al. 2019 to demonstrate VADR is a mid level fusion of two off-the-shelf classifiers. It was trained on a subset of UCF-101 with 51 categories.

In this project, I aim to (1) evaluate the current state of research regarding XAI for multimodal data fusion and (2) demonstrate the concept of multimodal XAI for activity recognition. From this, I aim to produce two models with version 1 reproducing VADR with state-of-the-art classifiers for audio and video, and version 2 applying additional ideas based on my findings, with the overall aim of making an improved model.

2 Project Aims and Objectives

This project aims to research how explainability can be added to a model that is making a classification or prediction on multimodal data. I will start by reproducing results produced by researchers in the Crime and Security Research Institute who developed the VADR technique, before attempting to make improvements. This will come in the form of two models, both of which will be trained on a subset of the UCF-101 dataset[1] using 51 actions instead of the full 101.

To use existing work such as pre-trained classifiers for audio and video, my implementation will be using a Python ML framework called PyTorch[2] which is already heavily within the Crime and Security Research Institute.

2.1 Aims

2.1.1 Initial Goal

The initial goal, which is considered the minimal viable product, will be to reproduce the results achieved by Taylor et al. 2019. I will swap the off-the-shelf classifiers for state-of-the-art classifiers, but for the most part, the rest of the model will remain the same.

2.1.2 Target Goal

The desirable goal will be to produce an improved model that will look to apply additional research and my ideas to explainability for multimodal data fusion. This model will be evaluated against model 1 to conclude the final report.

2.2 Objectives

1. Research current XAI techniques used for multimodal data fusion. This will be used for the background section of the final report.

2. Reproduce VADR using state of the art models for video and audio classification. This model will be called version 1.

3. Evaluate the version 1 model against the model produced by Taylor et al. 2019. This will be recorded for use in section 5 of the final report.

4. Produce an improved model based on version 1 and the research I conducted for objective 1. This will be the version 2 model and will conclude section 4 of the final report.

5. Evaluate model 2 in comparison to model version 1. This will deliver section 5 of the final report.

[1] UCF-101 is a dataset of 101 different activity categories taken from YouTube. See `https://www.crcv.ucf.edu/data/UCF101.php`.

[2] Pytorch is an ML framework for rapid development in Python. See `https://pytorch.org/`.

3 Work Plan

I have outlined a plan on a two-week basis. This is subject to change as the project develops. I will have a scheduled meeting with my supervisor once every two weeks which will go over the previous two weeks' deliverable(s).

Week 1 and week 2: Write the initial plan and meet with my supervisor to ensure the future direction is clear. At this point, I will also start learning Pytorch which will be used throughout the project. At the end of week 2, I will have at least one example model (not related to this project) implemented in PyTorch.

Week 3 and week 4: Research current XAI techniques for multimodal data fusion. From these two weeks, I will have an understanding of what has and has not worked, which I can then use in the following weeks. This will deliver objective 1.

Week 5 and week 6: Reproduce VADR using state-of-the-art audio and video classifiers. This will deliver objective 2 and 3.

Week 7 and week 8: Design and set up my experiment. During these two weeks, I will be able to start planning and implementing my proposed model version 2. This will deliver part of objective 4.

Week 9 and week 10: Complete and Train my model. From this, I will be able to record metrics that will be used to compare it against model version 1. At the end of week 10, I will have delivered objective 4 and 5.

Week 11 and week 12: Write up and submit the final report, bring together all of the work I have completed in previous weeks.

References

Arrieta, Alejandro Barredo et al. (2019). *Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI*. arXiv: 1910.10045 [cs.AI].

Baltrušaitis, T., C. Ahuja, and L. Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2, pp. 423–443. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2798607.

Braines, Dave, Alun Preece, and Dan Harborne (2018). *Multimodal explanations for AI-based multisensor fusion*. In Proc NATO SET-262 RSM on artificial intelligence for military multisensor fusion engines. URL: http://orca.cf.ac.uk/id/eprint/116675.

Taylor, Harrison et al. (2019). *Multimodal explanations for AI-based multisensor fusion*. In Proc British Machine Vision Association Symposium on Video Understanding, London, September 2019.