Multimodal Joint Emotion and Game Context Recognition in League of Legends Livestreams

Charles Ringer, James Alfred Walker, and Mihalis A. Nicolaou

c.ringer@gold.ac.uk james.walker@york.ac.uk m.nicolaou@cyi.ac.cy

OVERVIEW

Livestreaming provides viewers with rich audio-visual data about the game & streamer's emotional state. Analysing this data and discovering correlations between affect and game context is crucial for modelling and understanding the streamer's experience. However, there are a significant set of challenges such as how to fuse multimodal data captured in uncontrolled (`in-the-wild') conditions. This work explores utilising recent advances in deep learning to model both streamer emotion and game context in a single model. We presents a dataset of annotated livestreams as well as three models for joint emotion and game context recognition.

MODELS

Three models are presented, two using existing fusion techniques (Early and Late Fusion) and our proposed model (TensorTrain Fusion). All models follow a similar structure: Feature Extraction → Fusion and Temporal Modelling → Dense Classification Layers (Fig. 1).

Our TensorTrain Model replaces concatenation based fusion with a higher order interaction tensor (Fig. 2) to model interactions between data views and a TensorTrain layer to connect this to the dense layers in a weight efficient manner. We also find that modelling each view temporally before fusion (Late and TensorTrain fusion) results is far fewer weights, these networks are around 33% smaller than Early Fusion.

DATA

We present, to our knowledge, the first annotated dataset of audio-visual stream data. Data from 5 male & 5 female streamers playing League of Legends was collected. For each streamer, 20 minutes was taken from 3 games for 1 hour total footage, segmented into 7200 5-second long clips.

Game Content was annotated across 9 classes: In Lane, Shopping, Returning, Roaming, Fighting, Pushing, Defending, Dead & Miscellaneous. Miscellaneous data were discarded because they did not relate to specific game context.

Affect was annotated across two axes, Valence (Negative, Neutral & Positive) and Arousal (Neutral & Positive). In-the-wild affect data is hugely imbalanced, 'Neutral' makes up 86% of Valence labels and 94% of Arousal labels.

Oversampling was applied to the dataset during pre-processing to partially rebalance the classes (Fig. 3). Additionally, Class Weights were used during training to further address the class imbalance.

RESULTS

Results show that Tensor Train fusion generally outperforms Early Fusion and is comparable to Late Fusion across tasks.

Affect classification is particularly difficult in this environment, all models performed considerably better at game context classification.

For Late and TensorTrain fusion models jointly modelling game context and emotion outperforms learning these separately however Early fusion performs worse in this training paradigm. (Fig. 4).

CONCLUSION

Jointly modelling game context and streamer affect is a challenging, interdisciplinary problem, with challenges arising in areas such as computer vision, machine learning, and affective computing. However, these models go some way to answering key questions and provide baseline results.

Ultimately we invite more researchers to work in this area, as improved results and models are crucial for facilitating audio-visual player experience.

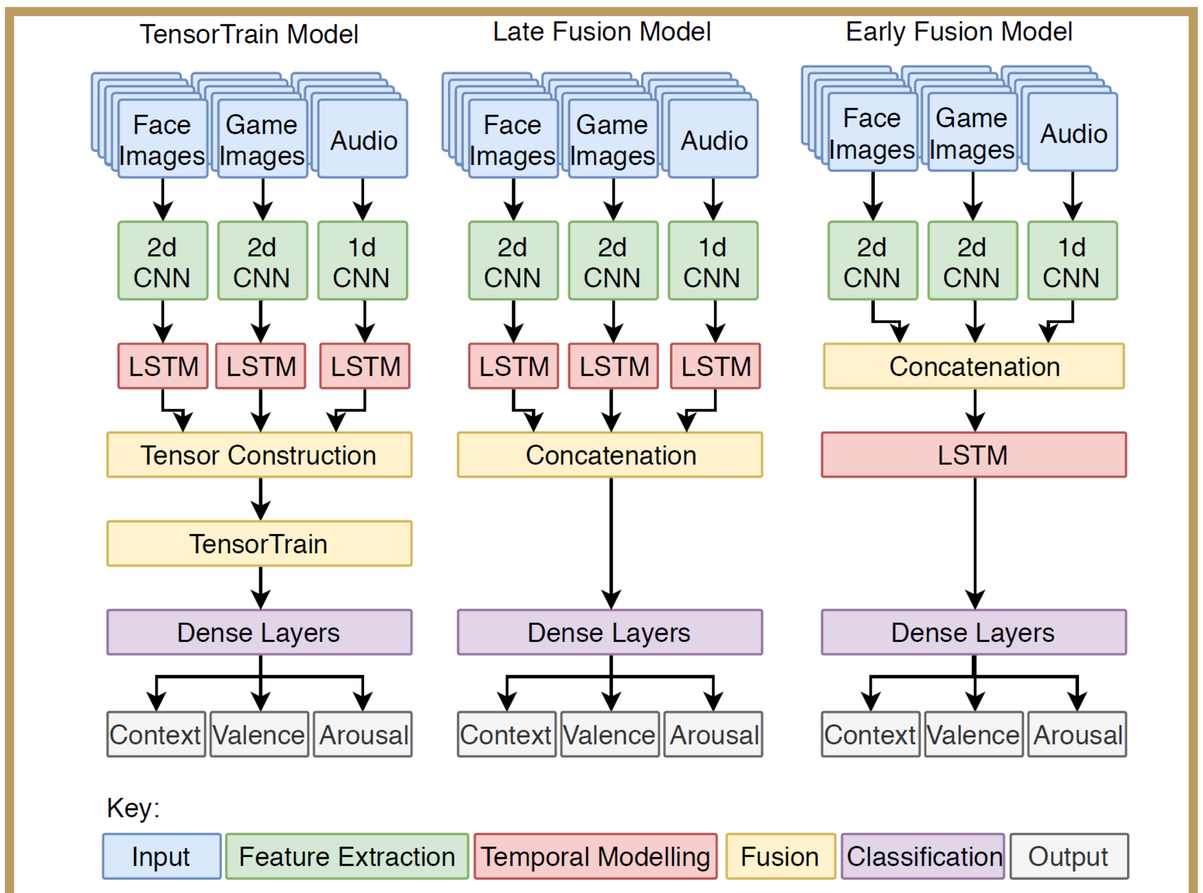
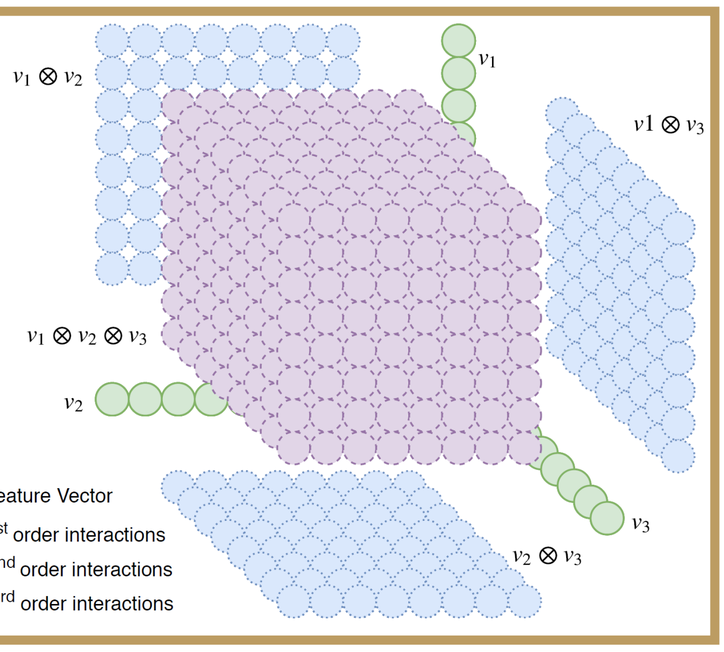


Fig. 1. *Overview of the three neural network architectures presented. All use the same CNN layers for feature extraction. Early fusion fuses each frame via concatenation before temporal modelling. Late fusion extracts a single feature vector per data view across time before concatenation. TensorTrain is similar to late fusion but replaces concatenation with a tensor of higher order features then uses a TensorTrain layer for weight efficient connection. Each network then uses the same dense architectures for classification.*

Fig. 2. *High order interaction tensor. 1st order interaction are latent features from each view, 2nd order interactions contain variables representing the relationship between all variables in two views, 3rd order interactions relate all three views. The tensor therefore consists of all 1st , 2nd and 3rd order interactions between data views.*



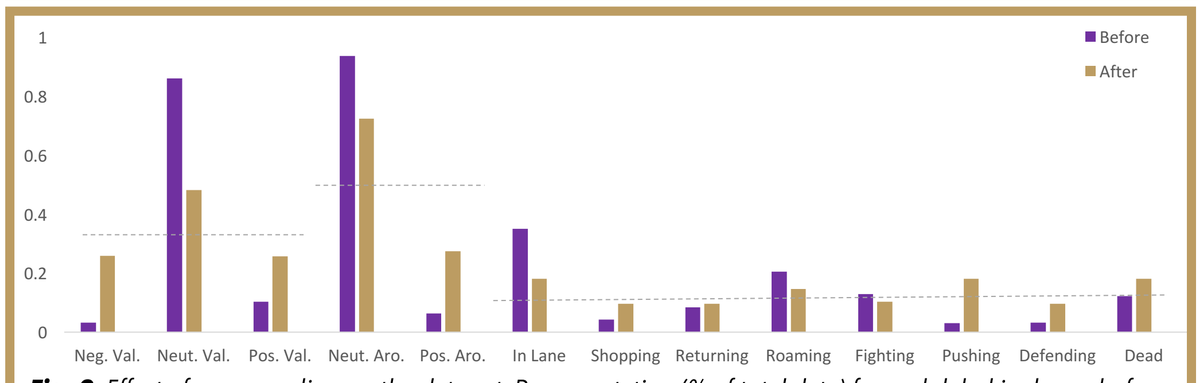
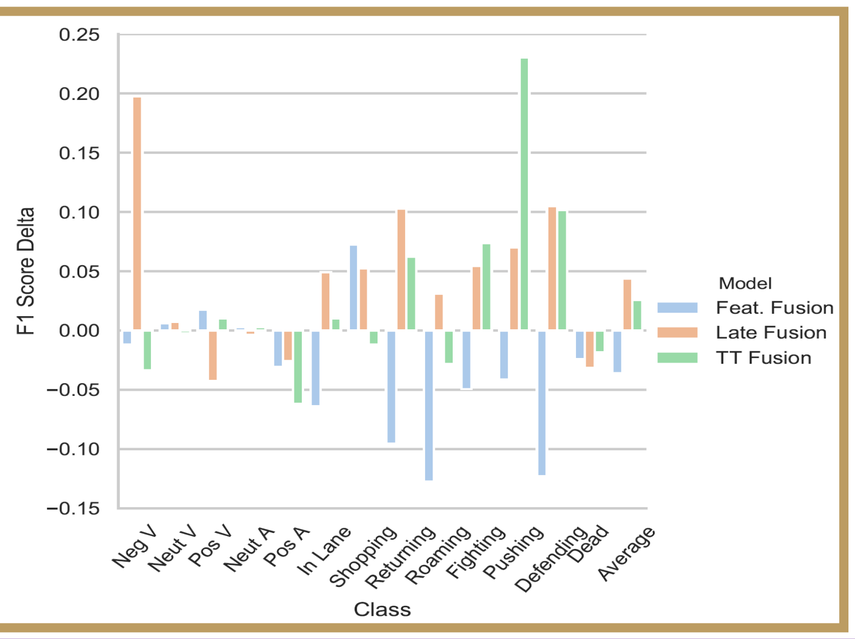


Fig. 3. *Effect of oversampling on the data set. Representation (% of total data) for each label is shown before and after oversampling alongside an even 'target' distribution (dashed lines).*

Fig. 3. *Performance delta between joint and single task learning. Positive values represent examples of joint task learning performing better. The last column shows the average performance change.*



Paper

Data & Code

EPSRC Engineering and Physical Sciences Research Council

iGGi digital creativity labs

Goldsmiths UNIVERSITY OF LONDON

University of York

The Cyprus Institute RESEARCH·TECHNOLOGY·INNOVATION