Integrating Privileged Information in Different Affect Modelling Paradigms

Progress Report

Nicholas Azzopardi
University of Malta
Msida, Malta
nicholas.azzopardi.23@um.edu.mt

Abstract

Affective computing aims to estimate human emotional states from observable signals such as speech, facial expressions and physiological activity. Although deep learning has significantly advanced this field, affect models that are trained under a controlled laboratory environment (vitro) often fail to generalise to the nosier real world (vivo) in which these applications are designed for. In practice, the richness and quality of data is considerably lower than what's available during model development.

This project addresses the problem of how to take advantage of rich multimodal information during training while still producing affect models that can operate efficiently when only a limited subset of modalities is available at test time. Moreover, the RECOLA dataset is the cornerstone of this implementation, where arousal and valence are considered as the continuous regression and binary classification targets.

The paradigm which is adopted to solve this task is the Learning Using Privileged Information (LUPI) framework which employs a teacher-student architecture. With such a design, knowledge may be transferred from teacher to student during training, with the goal of improving the performance and robustness of single modality affect models.

1 Introduction

Affective computing is constantly being applied in settings where emotional signals must be interpreted from incomplete, noisy or unstable sensory input [1]. Although research experiments rely on rich multimodal data, the reality for many of the real-world applications is a complete different setting. Environmental noise, unreliable sensors and hardware limitations frequently result in situations where only a subset of modalities is available and the richness of those signals is often much lower than in controlled laboratory settings. As a result, affect models that perform well during the development stage tend to deteriorate once deployed [8].

This inconsistency between the data conditions used for training and those experienced during application account for a core challenge in affect modelling. Improving robustness under modality constraints is essential, particularly in applications such as mobile affect sensing, wearable devices, and real-time emotion monitoring, where access to multiple high-quality modalities cannot be guaranteed.

The motivation behind this project is to explore how having additional modalities available during training can alleviate the current discrepancy. Rather than expecting real-world systems to acquire the same multimodal richness as lab data, an alternative approach is to make use of this knowledge only during training in a way that can benefit models that may be operating under restricted conditions [2], [3], [7]. The LUPI framework offers a mechanism to achieve this goal by enabling models to learn from information that is not accessible at test time [3].

2 Aims and Objectives

The main aim of this project is to investigate how integrating the LUPI framework can benefit affect modelling through privileged multimodal information. Towards this endeavour, the tri-modal RECOLA dataset is used. The study focuses on arousal and valence as the targets in both regression and binary classification settings, having the constraint of only a single modality being available at test time. To achieve this aim, the following objectives have been set:

- Develop single-modality baseline models for each modality, for both regression and binary classification, to establish reference performance levels without privileged information.
- Construct tri-modal teacher models for regression and classification that fuse audio, video and physiological features and act as privileged-information experts during training.
- Design and implement student models that operate on a single modality (initially audio only), trained within the LUPI framework to leverage guidance from the tri-modal teacher while remaining deployable under single-modality constraints.
- Define an evaluation protocol and appropriate metrics to compare baselines, teacher models and student models.
- Analyse the impact of privileged information on model performance and robustness, and reflect on the implications for deploying affect recognition systems in realistic, modality-limited environments.

3 Background

3.1 Affective Computing and Dimensional Emotion Models

Affective computing is the strategy of developing systems that can detect, interpret and respond to human emotions [1]. Rather than treating emotions as a small set of discrete categories, modern techniques represent affect through utilising continuous dimensions. Two of the most commonly used dimensions are arousal, reflecting the intensity level of an emotional state, and valence, which reflects how positive or negative that state is. In addition, to develop models for such target variables, two techniques can be considered. These are regression and classification, where in regression the goal is to predict continuous arousal and valence values, whilst in classification, the goal is to mitigate subjectivity bias in affect annotations by discretizing these labels into classes such as high and low. The relevance of implementing two different machine learning techniques furthers the level of appreciation for regression's ability to preserve the small changes of the original annotations, while classification can simplify such tasks to a decision-oriented system.

3.2 Multimodal Affect Modelling

Human affect can be expressed in various ways including speech, facial expressions, body language, and physiological activity. Multimodal affect modelling aims to make use of this by combining information from different modalities. Each modality captures different pieces of information of the underlying emotional state [2]. For example, data derived from audio could capture pitch and Mel-Frequency Cepstral Coefficients (MFCC's), whilst video data could capture body language.

When modelling human effect having more than one modality, individual encoders are often used for each. This allows the model to create an embedding that acts as a representation for each type of information. The goal is then to process diverse data into a common format that allows a model to perform multimodal reasoning across different types of information. Moreover, following this step, a fusion technique would be implemented to combine their representations. Fusion can take place at different levels, where early-fusion concatenates features before the network, mid-level fusion to combine intermediate representations, and late-fusion, which groups modality-specific predictions at the output level. The specific choice of fusion strategy comes down to the design choices based on how much the model can exploit cross-modal relationships and how robust it is to missing modalities [2].

3.3 The RECOLA Dataset

The RECOLA (Remote Collaborative and Affective Interactions) dataset is a multimodal corpus designed for continuous emotion recognition [11]. It consists of roughly 20 French-speaking participant recordings of spontaneous interactions in a remote collaborative task, having synchronised audio, video, and physiological signals with continuous annotations for both arousal and valence.

For this project, RECOLA serves as the primary dataset and provides a number of advantages. Firstly, the ability to have multiple modalities captured in a controlled environment aligns with the concept of privileged information. Secondly, the continuous arousal and valence annotations work well for both regression and classification techniques. Third, the dataset design involves real interactions rather than it being some sort of setup or act, making the data extracted represent the participant's true emotions.

3.4 Learning Using Privileged Information and Teacher-Student Learning

The LUPI framework allows a learning algorithm to have access to additional information during training that may not be present during testing. The key idea is that the model can use this extra information to organise its internal representation of the problem more effectively, even though it will need to operate without it [3], [4].

A common way to implement LUPI in deep learning is through teacher-student architectures. In this setup, a teacher model is trained with access to the full set of inputs, including the privileged information, and learns a knowledgable representation of the task. A student model is then trained to operate with less inputs, such as a single modality, while at the same time having the guidance of the teacher. This guidance can be applied at different levels of the architecture, being typically referred to as knowledge transfer [5], [6].

4 Literature Review

4.1 Learning Using Privileged Information

The Learning Using Privileged Information (LUPI) paradigm was introduced by Vapnik as an extension to classical supervised learning. In the original writing, the learner only has access during training to an additional "privileged" representation of each example. For instance, expert annotations or a richer feature set which is not available during testing. The idea is that this extra information can help formulate the decision boundary and speed up the learning process, even though the final model operates only on the standard features [3].

Future work refined the theory of LUPI and proposed mechanisms to exploit privileged information effectively. Vapnik and colleagues introduced methods such as similarity control and knowledge transfer to accelerate learning and improve generalisation using SVM-based models [4]. Around the same time, Lopez-Paz et al. unified LUPI with knowledge distillation in the context of generalised distillation [5]. In this study, both Hinton's and Vapnik's work are

interpreted as techniques that allow machines to teach other machines using richer representations or additional predictive information [5], [6]. Moreover, the work performed at Stanford University further clarified how LUPI can be implemented with neural networks, by for instance incorporating privileged features into auxiliary losses [7].

4.2 Privileged Information and Distillation in Affective Computing

Makantasis et al. represent affect modelling as a vitro versus vivo problem, where models trained in a controlled laboratory must operate in the challenging settings of the real-world. They propose using privileged information to bridge this gap by allowing affect models to be trained on all modalities available within the lab, but at deployment, the models deal with a reduced set of inputs. Their experiments on two multimodal affect databases show that models trained with privileged information can achieve performance close to fully multimodal systems, even when only a subset of modalities is available during testing [8], [12]. This work clearly presents this FYP's motivation on using multimodal signals as privileged information for training models that are deployed having access to less information.

A similar study shows privileged information through the lens of knowledge distillation for dimensional emotion recognition. Aslam et al. propose Privileged Knowledge Distillation (PKD) for emotion recognition in the wild, where a multimodal audio-visual teacher is trained using both speech and facial expressions, where its knowledge is distilled into a visual-only student model. Vocal expressions were defined as privileged information and show that distilling from the stronger teacher significantly improves the performance of the unimodal student on continuous arousal and valence estimation [9]. Recent follow-up work extends this concept to multi-teacher privileged knowledge distillation, where multiple teachers having different modality combinations are used to guide a student using structural similarity [10].

4.3 Multimodal Affect Modelling and RECOLA

The RECOLA corpus is a widely used benchmark for continuous affect recognition. Ringeval et al. present RECOLA as a multimodal dataset of remote collaborative interactions, having synchronised audio, video, ECG and EDA signals, along with continuous arousal and valence annotations [11]. For this FYP, RECOLA is well suited as it includes both behavioural and physiological modalities, allowing for the study of multimodal affect models.

4.4 Bridging the Gap

Existing LUPI applications in affective computing, such as Makantasis et al. and Aslam et al. demonstrate multimodal teachers distilling to unimodal students, achieving arousal/valence gains through a single or multi-teacher setup [8],[9],[10]. However, only a few studies systematically compare regression to classification within distillation pipelines, overlooking how continuous annotations manage to preserve subtle emotional hints that may play a critical role during real interactions.

This FYP addresses the gaps by leveraging RECOLA's audio, physiological, and video data as privileged information within the teacher-student framework, focusing on audio-only deployment. It evaluates both regression and classification through quantifying trade-offs in knowledge transfer efficiency, and investigates the impact of different knowledge transfer strategies.

5 Proposed Solution and Methodology

The proposed solution consists of a teacher-student architecture under the LUPI framework on the RECOLA dataset [3]–[5], [8], [9], [11]. The main idea is to use teacher models that have access to privileged information during training, and transfer that knowledge to the single-modality student models that operate under realistic conditions at both training and testing. The methodology may be structured as a pipeline compromising of data preparation, baseline model construction, teacher model development, and LUPI-oriented student models, followed by an evaluation strategy.

At the time of writing this progress report, the preprocessing steps, six single-modality baseline models (three for regression and three for classification for audio, video, and physio data), two teacher models (one for regression and one for classification), and an audio-only regression student model have been implemented. The classification student models form the next stage of implementation.

5.1 Data Preparation

The initial stage of this FYP involves preparing the RECOLA dataset, which can be split into two steps. First, all non-target features are standardised using scikit-learn's StandardScalar [13]. This is fitted on the full feature matrix, excluding the Participant ID and two continuous targets. The original median arousal and median valence attributes are preserved as target variables for regression. Second, the binary classification targets are obtained by computing global median thresholds for arousal and valence across all instances and assigning each sample to the high or low class accordingly. Samples with values above the median are labelled as "high", while those below are labelled as "low", resulting in the arousal and valence classification targets.

The processed data once saved, is then further processed by splitting the data into three modality-specific datasets:

- An audio file containing ComParE-style acoustic features and a voice similarity indicator
- A video file containing numeric visual features and a face detection probability
- A physiological file containing ECG and EDA features

5.2 Data Splitting and Validation Strategy

The data splitting strategy applied is Leave-one-participant-out (LOPO) cross-validation, implemented with GroupKFold [13]. This technique allows for subject-independent grouping, where each participant is treated as a group. Within each fold, one participant is left out to be used for the test set, and all instances forming part of that participant are excluded from the train set. This ensures that no participant appears in both the train and test data, improving the purity of the predications made.

The validation strategy within each fold that was adopted during neural network training is a validation split that is taken from the training portion to monitor overfitting and support early stopping. The same grouping strategy is applied across all models to ensure that each model is evaluated under the same conditions.

5.3 Single-Modality Baseline Models

The baseline models are implemented as multi-layer perceptrons having access to a single modality. For each of the three modalities, two baseline models were developed. The first being a classification baseline that predicts high/low arousal and valence from the given modality, and the second is a regression baseline model that predicts continuous median arousal and valence.

In both cases, the architecture consists of fully connected layers having ReLU as their activation function and dropout for regularisation, followed by an output layer [14].For classification, the sigmoid activation function is used at the output layer with binary cross-entropy loss. On the other hand, regression consists of a linear output having mean squared error as the loss function. Training for both techniques is performed using the Adam optimiser [15],[16], and early stopping based on the validation loss.

For each fold, the model is trained from scratch and performance is recorded on the participant that had been left out. Classification performance is summarised using accuracy and F1-score computed between the ground-truth binary labels and the predicted class labels, while Pearson's correlation is computed between the binary labels and the predicted probabilities. For the regression baselines, performance is measured using mean squared error and Pearson's correlation between the continuous predictions and the continuous target values. These baselines shall then be used as reference performance for when comparisons take place with the student models.

5.4 Tri-Modal Teacher Models with Privileged Information

The teacher models are designed to utilise all three modalities available in RECOLA. Both regression and classification teachers share a common architectural design.

Each modality is processed by its own encoder branch, implemented as a small MLP with two hidden layers, ReLu activations, dropout, and L2 weight regularisation. The outputs of the three encoders are then concatenated into a single fused representation, which is then passed through a fusion head consisting of additional dense layers with ReLU and dropout, which learns cross-modality interactions and produces a compact tri-modal embedding [2],[8].

As with the baseline models, the teacher models are trained using LOPO cross-validation through GroupKFold, having early stopping based on the validation loss. For each task, the mean and standard deviation of the overall performance is computed.

5.5 Student Models under the LUPI Framework

The student models are intended to have an identical architecture design to those of the baseline models, while now incorporating additional loss terms that help reflect and interpret the influence the teacher model is having on the student. The key factor is that the student models only have knowledge of a single modality during both training and testing. However, as training commences, the student can utilise the teacher's knowledge indirectly through the loss function [4],[5],[6].

For regression, an audio-only student model has been implemented. Its architecture follows the same structure as the audio regression baseline model, being an MLP with the same number of hidden layers and dropout rate, followed by a final linear output to predict continuous arousal or valence. Moreover, the only addition to the student model is the student representation layer, which acts as a bottleneck placed before the output, allowing the student and teacher to align.

The student is optimised using a combined loss function made from the following two components:

- **Data Loss (Mean Squared Error):** Measures the discrepancy between the student's predictions and the ground-truth continuous labels.
- **Representation Loss (Cosine Distance):** Encourages the student's internal representation to align with the corresponding hidden representation from the teacher model. This is computed as:

$$\text{cosine\_distance}(T, S) = 1 - \textit{cosine\_similarity}(T, S),$$

  where T denotes the teacher's fusion-layer representation and S denotes the student's bottleneck representation [8],[9].

These two components are combined into a single objective using a weighting parameter α [5],[9], which controls the teacher's influence during training:

$$\text{total\_loss} = (1 - \alpha) * \text{MSE} \; + \; \alpha * \text{cosine\_similarity}(T, S).$$

When $\alpha = 0$, the student ignores the teacher and focuses only on the ground truth labels. As α increases, the influence

of the teacher on the student rises, driving the student to follow the ground truth less and the teacher's representation more. Furthermore, in terms of the experimentation approach taken for the FYP, within a single student model, a total of eight experiments take place. This is constructed by four values for alpha (0.25, 0.5, 0.75 and 1.0), creating four trained models for arousal and valence respectively.

For each fold, the teacher is first trained using the training participants, with early stopping to prevent overfitting. Once the teacher has been trained, the fusion-layer representation is extracted and used as privileged information. When the student then begins training, the audio-only model receives the audio features, but its internal representation is encouraged to match the teacher fusion representation according to the combined loss function.

Moreover, for classification, rather than developing a single student model, a total of three classification models shall be developed. These consist of:

- A representation-based student model, where cosine distance will be used to encourage the student's hidden representation to match those of the classification teacher.
- A prediction-based student model, where KL-divergence encourages the student's probabilistic outputs to match the teacher's predicted probabilities
- A combination of the above two, where both representation and prediction based loss terms are added, allowing the student to benefit from both types of privileged information at the same time [5],[6],[9],[10].

For all cases, the student architecture remains exactly the same as the corresponding single modality baseline model, ensuring that when evaluating the results, the true improvement with the help of privileged information is obtained.

6 Evaluation

The models in this project will be evaluated using the RECOLA dataset [11], following a consistent implementation across all baselines, teacher models, and student models. Due to the dataset having multiple recordings for each participant, evaluation will make use of the LOPO cross-validation strategy. This is to ensure that the models are always tested on participants that are not present during training.

For the regression tasks, the evaluation will rely on two well known metrics in affect modelling [17]:

- **Mean Squared Error (MSE):** Measures the average squared difference between the predicted and true values, capturing the overall prediction accuracy.
- **Pearson Correlation Coefficient (PCC):** Measures how strong the predictions follow the trend of the ground truth annotations.

For the classification tasks, performance will be evaluated using PCC once again, along with:

- **Accuracy:** The proportion of correctly classified recordings
- **F1 Score:** The harmonic mean of precision and recall

All metrics will be computed for each fold, and the final results will be summarised using the mean and standard deviation across the participants. Moreover, appropriate visualisations will also be designed.

Finally, the student models will be compared against their corresponding single-modality baseline. This comparison will help identify whether incorporating privileged information during training produces measurable improvements.

7 Conclusion

This progress report displays the work completed so far in developing the teacher-student framework for affect modelling using privileged information. The RECOLA dataset was preprocessed and structured into modality-specific feature sets, allowing the development for each of the single modality baseline models for both regression and classification.

Teacher models were then implemented for both techniques, establishing the means of obtaining and further transferring the knowledge to the student. Moreover, the first of four student models is in the works, specifically the audio regression model.

The next stage of this FYP will focus on taking the logic implemented for the first student model and adjusting it accordingly for the set of classification models. With the remaining components defined, a project schedule outlining the upcoming steps is provided in the Gantt chart below (Figure 1).

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000. doi: https://doi.org/10.7551/mitpress/1140.001.0001.

[2] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, Feb. 2019. doi: https://doi.org/10.1109/TPAMI.2018.2798607.

[3] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5–6, pp. 544–557, Jul. 2009. doi: https://doi.org/10.1016/j.neunet.2009.06.042.

[4] V. Vapnik, R. Izmailov, A. Gammerman, and V. Vovk, "Learning Using Privileged Information: Similarity Control and Knowledge Transfer," *Journal of Machine Learning Research*, vol. 16, pp. 2023–2049, 2015. [Online]. Available: https://jmlr.org/papers/volume16/vapnik15b/vapnik15b.pdf

[5] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik, "Unifying distillation and privileged information," arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1511.03643

[6] G. Hinton, O. Vinyals, and J. Dean, "Distilling the Knowledge in a Neural Network," arXiv, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[7] A. Momeni and K. Tatwawadi, "Understanding LUPI (Learning using Privileged Information)," 2018. [Online]. Available: https://web.stanford.edu/~kedart/files/lupi.pdf

[8] K. Makantasis, K. Pinitas, A. Liapis, and G. N. Yannakakis, "From the Lab to the Wild: Affect Modeling Via Privileged Information," *IEEE*

Transactions on Affective Computing, vol. 15, no. 2, pp. 380–392, Apr. 2023. doi: https://doi.org/10.1109/TAFFC.2023.3265072.

[9] M. H. Aslam, M. Zeeshan, M. Pedersoli, A. L. Koerich, S. Bacon, and E. Granger, "Privileged Knowledge Distillation for Dimensional Emotion Recognition in the Wild," in *Proc. IEEE/CVF CVPR Workshops*, 2023, pp. 3338–3347. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2023W/FGAHI/papers/Aslam_Privileged_Knowledge_Distillation_for_Dimensional_Emotion_Recognition_in_the_Wild_CVPRW_2023_paper.pdf

[10] M. H. Aslam, M. Pedersoli, A. L. Koerich, and E. Granger, "Multi Teacher Privileged Knowledge Distillation for Multimodal Expression Recognition," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2408.09035

[11] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013. doi: https://doi.org/10.1109/FG.2013.6553805.

[12] K. Makantasis, D. Melhart, A. Liapis, and G. N. Yannakakis, "Privileged Information for Modeling Affect In The Wild," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2107.10552

[13] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html

[14] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: https://jmlr.org/papers/v15/srivastava14a.html

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[16] B. Tuychiev, "Adam Optimizer Tutorial: Intuition and Implementation in Python," DataCamp, Aug. 29, 2024. [Online]. Available: https://www.datacamp.com/tutorial/adam-optimizer-tutorial

[17] "Emotional Impact," MediaEval, 2016. [Online]. Available: http://www.multimediaeval.org/mediaeval2016/emotionalimpact/

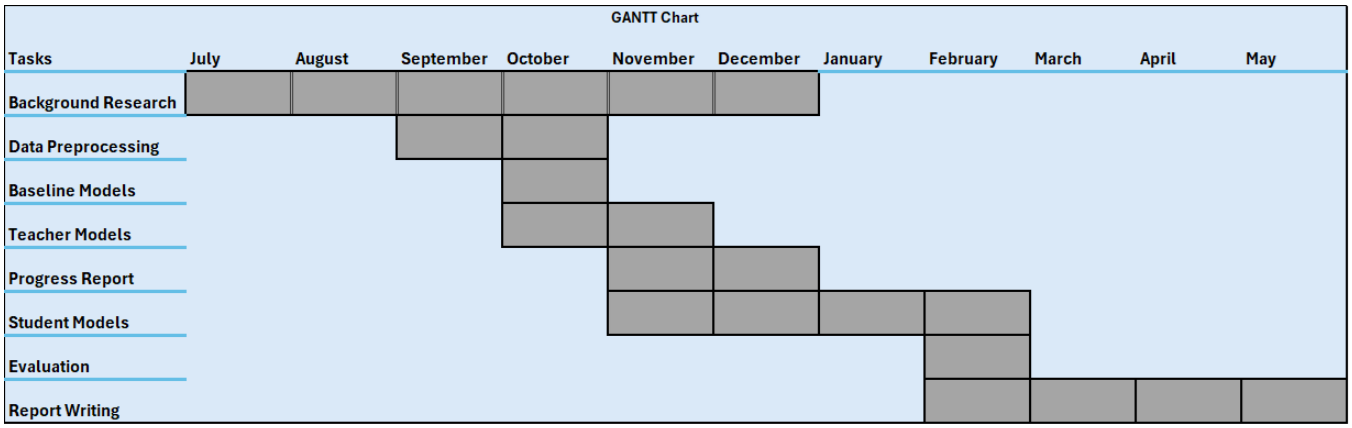

Figure 1. Project Gantt Chart