Understanding the Relevancy of Modality Information in Multimodal Machine Learning

Jack Geraghty[1], Andrew Hines[1] and Fatemeh Golpayegani[1]

[1]School of Computer Science, University College Dublin, Dublin, D04 V1W8, Ireland

Abstract

Multimodal machine learning aims to improve the performance of a model, including accuracy, precision, and robustness, that is trained using multiple modes of data such as audio, video and text. Recent research has improved these measures by focusing on the intersection of information between modalities. However, there is a lack of research that focuses on how a multimodal model utilizes unique modality information and how a target task changes the distribution of relevant modality information. Understanding the relevancy of each modality can help with the development of future multimodal models, quantifying model robustness when a modality is missing and understanding how the inference process depends on each modality. This paper investigates how multimodal models learn within and across modalities and how the distribution of relevant information per modality changes depending on the target task being carried out. To achieve this, two multimodal classifiers were trained for three different tasks using audio and video. Using these models, we demonstrate that depending on the target task, the amount of relevant modality information will change and that an individual modality may contain more relevant information than the intersection of all modalities for predicting the ground truth.

Keywords

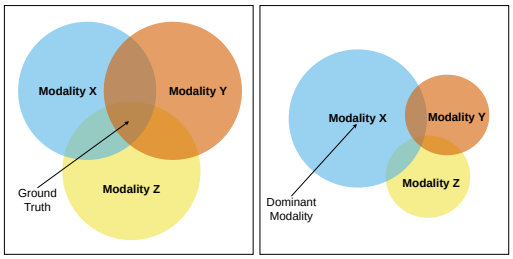
multimodality, multimodal machine learning, modality information, information relevancy

1. Introduction

Understanding the individual modalities and the information they contain is crucial for multimodal machine learning. Each modality will contain information relating to the target task, and how modality information is fused will influence how a multimodal model performs. In recent years, the literature has approached the challenge of understanding modalities and their information in various ways, each with its own merits. Each approach takes a different viewpoint on how the information should be used within and across modalities and are not necessarily mutually exclusive to each other. These approaches include focusing on the intersection of information between modalities, i.e. mutual information and attempting to balance the contribution of dominant modalities towards a learning objective, i.e. imbalanced information. The Venn diagrams in Figure 1 illustrate these two approaches and the relationships between the information of each modality for a given task. Each circle represents all the information (relevant to the task) contained within a modality. The intersection of a group of modalities represents the information shared between them, with the centre containing the information they all share. The non-intersecting parts of each circle represent the unique information contained per modality. Then, the area outside the circles represents all the other information irrelevant to the task.

A significant portion of multimodal research deals with problems that cause poor performance when multiple modalities are involved. These problems include how can you work with modalities that are represented in a fundamentally different way [1, 2, 3] and how can information be translated from one modality to another to achieve effective learning [4, 5, 6]?

These challenges are often amplified in real-world applications, as reasonable assumptions made in multimodal research often become limiting when applied in a real-world system. For example, assuming that the underlying conditional distributions of the modalities are the same or at least similar[7, 8, 9], or that each modality/view contains enough information to accurately carry



(a) Mutual Information (b) Imbalanced Information

Figure 1: Different views of information in multimodal machine learning.

MRC 2023 – The Fourteenth International Workshop Modelling and Representing Context. Held at ECAI 2023, 30.09.-5.10.2023, Kraków, Poland.

jack.geraghty@ucdconnect.ie (J. Geraghty); andrew.hines@ucd.ie (A. Hines); fatemeh.golpayegani@ucd.ie (F. Golpayegani)

out a task[10, 11, 12]. Others take the view that the intersection of the information/representations of the modalities contains the ground truth/optimal intersection of information for a task[13, 14]. All of these assumptions fail to acknowledge that the unique information contained within a modality might be more relevant than others when performing a task.

Other works describe the potential causes of poor performance in multimodal classification networks, such as multimodal models being prone to overfitting due to their increased capacity, which is compounded by the different generalisation rates of the modalities[14], or that there is insufficient structure in some modalities that leads to poor performance[15]. This paper takes the view that for a given observation of modalities, one modality might contain very little relevant information, whereas another contains the majority of relevant information. However, this is counteracted by the way in which multimodal models learn, i.e. there is no guarantee that a model will utilize information in a way that we would expect.

Two multimodal classifiers are implemented to understand better the impact of modality relevancy and imbalanced modality information, one binary and the other multi-class. The binary classifier was trained for two different classification tasks (urban vs. rural and speech vs. non-speech) using the same data from the AudioSet data set[16]. Using the same data for both tasks, with the only difference being the target label, we can investigate how the underlying, relevant information changes depending on the target task. These labels were selected as they had the potential to represent tasks that were vision-based (urban vs. rural) and audio-based (speech vs. non-speech). To gain further insight into modality relevancy and imbalanced modality information, from a more challenging task, the binary classification task was extended to a multi-class classification task. The multi-class classifier was trained on the Kinetics-Sounds[17] data set, which contained the video and corresponding audio of 26 human-action classes, such as "playing guitar", "laughing" and "singing". Once the models were trained, a leave-one-out approach was adopted in order to determine how much the multimodal models relied on the information coming from each modality. Our results show that:

1. Depending on the task, the amount of relevant information present within a modality will change.
2. There is no guarantee that using a good video model and a good audio model will create a good audio-visual model.
3. We provide experimental results that show that multimodal models can ignore almost entire modalities during training and inference, **despite** the individual modalities being capable of carrying out the target task.
4. We provide early results for the impact of representing missing data for audio/visual tasks.

The remainder of the paper is organised as follows: Section 2 details the related works, section 3 describes the models and data sets used for the experiments, section 4 describes the experiment methodology, section 5 then presents the results and a discussion around them and finally, section 6 concludes the work and details several future research directions.

2. Related Works

This section details related works in the areas of multimodal information and multimodal machine learning.

2.1. Multimodal Information

Understanding the individual modalities and the information they contain is crucial for multimodal machine learning. Each modality will contain information relating to the target task and how modality information is fused will influence how a multimodal model performs. An assumption often seen in multimodal learning is that each modality is conditionally independent with respect to some common ground truth label. While each modality is independent of the others, they all agree on one thing: a ground truth label. Research has used the idea of common ground between modalities to propose solutions that optimise the learning process based on the intersection of modality information.

Total Correlation Gain Maximisation(TCGM) is a novel information-theoretic approach to multimodal machine learning[13]. Using the previous assumption, they propose that the information intersection of modalities will contain the best approximation of the ground truth as it is the information they agree on the most. By maximising the total correlation gain across each modality classifier, they all cooperatively discover the information intersection. Another novel classification architecture is Cross Partial Multi-View Networks (CPM-Nets)[18]. The architecture utilises the assumption that the information per view/modality originates from a fused latent representation of the views/modalities. To handle complex correlations between views/modalities, the architecture focuses on learning a complete and versatile fused representation of the data. Another approach, aimed at determining when and why multimodal models outperform their unimodal counterparts, takes the intersection of information assumptions and puts it into the context of latent spaces[14]. They propose that the latent space of each modality is contained within the latent space of the fused modalities. They then propose a novel metric, latent representation quality to measure the distance be-

tween a learned latent representation $\hat{g}$ and a true latent representation g^*.

2.2. Multimodal Machine Learning

Multimodal machine learning aims to build models that can process and understand information from multiple modalities, similar to how humans learn and reason. Each modality contains information that is relevant to the task being carried out. The information contained per modality can be divided into two groups: information unique to the modality and information that is present in the other modalities. This information divide raises challenges to the efficacy of a multimodal machine learning model.

A survey [19] on the current state of multimodal learning identifies the following general challenges in multimodal machine learning: **Representation**; how to work with modalities that are often complementary but are fundamentally different in how they are represented [20, 21], **Translation**; how to map the information from one modality to another [4, 5], **Alignment**; how to identify the direct (and indirect) relations between elements of different modalities [20, 22], **Fusion**; how to join information from multiple modalities [23, 24] and **Co-learning**; how to transfer information from one modality to another, their representations and models that use them [25, 26].

Many multimodal machine-learning solutions have been proposed in the literature in recent years. Unsupervised machine learning techniques such as variational autoencoders have been used for a wide range of tasks, such as detecting fake news [27], anomaly detection [28] and annotating mass spectra in chemistry [29]. Other works have explored combining techniques/models such as convolutional LSTMs with an autoencoder architecture [30], or where two unsupervised models (one for each modality) are trained for image-to-image translation in autonomous vehicles [31].

As the amount of available multimodal information increases, labelling each modality becomes time-consuming and costly. Semi-supervised learning is often used to solve this large-scale labelling problem in unimodal and multimodal learning. To alleviate the problem of not having enough data, semi-supervised learning often involves training a small, supervised classifier to generate labels for unlabeled data. This is a challenge in unimodal learning, which is amplified in multimodal learning. For example, [8] utilizes a supervised classifier per modality to generate labels for unlabeled data. However, they do not account for the possibility that the modalities do not share the same conditional distributions for the data point labels. Other approaches focus on learning a joint representation, which projects the modalities into a multimodal space [14]. Then a classifier is trained using labelled data to predict the joint representation. Another learning approach subdivides an unlabeled input data set and then clusters these divisions based on their fuzziness[32, 33]. Those with a low fuzziness are considered "better" samples and are then labelled using a classifier which has been trained on a smaller, labelled data set.

Supervised learning methods such as Canonical and Deep Canonical Correlation Analysis (CCA/D-CCA) [34, 35] maximise the canonical correlations between modalities, with D-CCA also accounting for the non-linear relationships between modalities. Subspace learning [36] utilises matrix factorisation to factorise modalities into a modality-invariant and a modality-specific part. The invariant part is then used by the predictive model. Knowledge distillation is a method of transferring knowledge from one or more models to another and has traditionally been used for model compression. However, in recent years, it has been explored as a method of multimodal learning. When used for multimodal learning, each expert model produces a soft label which is given to the student model during training [37]. The soft label is used as part of the loss function of the model and is meant to provide a metric for how close a student's prediction is to an expert prediction.

To prevent overfitting while also accounting for the difference in generalization rates a gradient blending approach, which computes an optimal blending based on their overfitting characteristics, is proposed [38]. While the approach offers state-of-the-art performance for various multimodal tasks such as human action recognition and acoustic event detection, there is room for further investigation with regard to how the approach performs with missing data and identifying a root cause for the overfitting.

Current state-of-the-art approaches in literature have made good progress in attempting to understand and use the information contained within modalities. However, the research gap still exists in relation to understanding how multimodal models use the information available to them and how they can be trained in such a way as to make effective use of the varying amounts of information available from all modalities with respect to a target task.

3. Data sets & Models

This section details the data sets used and the models implemented as part of the experiments conducted in this paper. Each convolutional and linear, excluding the output layers, is followed by a ReLU activation layer. Full details on the model parameters are provided in the Appendix.

3.1. Data Sets

AudioSet is a sound and vocabulary data set consisting of an expanded ontology of 632 audio event classes and a collection of over 2M+ human-labelled 10-second sound clips drawn from YouTube videos. Each entry in the data set can have several audio event labels. There are two major points to note about AudioSet. The first is that it is possible for videos, from which the audio clips are taken, to no longer be available for a variety of reasons such as the video being made private or being removed from YouTube. The second is that there is likely modality bias in how the labels were assigned for each audio clip. The way in which human labellers assigned labels involved them both listening and watching the video clip. This means an audio event label was potentially assigned using both the audio and visual modality.

A subset of entries from AudioSet was taken that contained the labels "urban" or "rural". After filtering unavailable videos and corrupted downloads, the final urban-rural subset contained 7276 samples, with a 50/50 label split. Then for the purpose of further evaluation, the urban-rural subset had the labels changed to "speech" and "non-speech". Since each data set entry has multiple labels, these labels were retrieved from the other labels of a given entry. Non-speech is the absence of a speech label. This label change results in a class imbalance of 61/39 in favour of speech.

All of the downloaded audio samples were sampled at 16000Hz with a single channel and a duration of 10s. All of the downloaded videos had a resolution of 420p and a duration of 10s. Prior to training the video model, 8 frames of the 10s clips were taken using a uniform temporal sample, normalized, scaled to a common size and a centre crop operation performed. The resulting images were then converted to a feature representation using the pre-trained ResNet50 model and saved locally to reduce training time. The ResNet features were then used as input to the video mode

Kinetics-Sounds is a subset of the Kinetics data set [39] containing 19k (15k train, 1.9k validation and 1.9k test) 10-second clips formed by filtering on a set of human action classes that potentially manifest both visually and aurally, for example, "playing guitar", "laughing" and "typing". The same pre-processing steps were used for the kinetics-sounds audio and video samples as for the AudioSet samples. The full list of classes can be found in the Appendix.

3.2. Multimodal Knowledge Distillation Model

The first model implemented is a multimodal knowledge distillation model. The model design is based on the knowledge distillation model proposed in [37]. The model used in this experiment consists of an expert video model, an expert audio model and then a multimodal student model which learns from the experts. Knowledge distillation was selected as the learning method as it has shown promising results in the literature and also reflects the human sensory system in terms of having expert senses that are then fused together to provide multimodal perception.

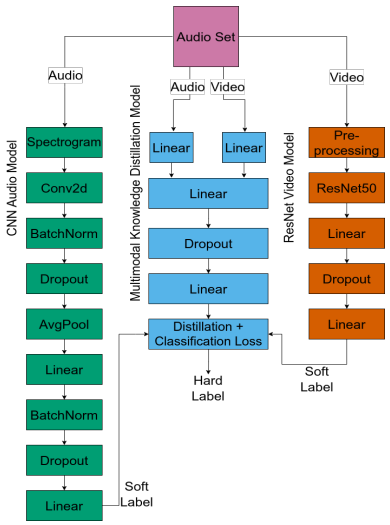


Figure 2: Left: CNN Audio Model. Centre: Multimodal Knowledge Distillation Model. Right: ResNet Video Model. The ResNet Video Model and CNN Audio Model are the expert teachers with the Multimodal model being the student learner.

The **student multimodal model** (centre in Figure 2) consists of two separate fully connected layers, one for the audio input and one for video input. The outputs of these layers are then fused (late feature concatenation) to produce the final prediction.

The **expert audio model** (left in Figure 2) is composed of a single 2D convolutional layer followed by two fully connected layers. When more complex architectures, similar to those seen in audio classification literature [40, 41, 42], were used, the model failed to learn a representation of the classification problem, indicated by a stagnant loss each epoch. Resulting from this, a simplified architecture was adopted. The CNN audio model uses log-mel spectrograms as input features. Log-mel spectrograms are used instead of raw audio signals as they compress the input audio into a single image, which contains a concise representation of the original audio.

The **expert video model** (right in Figure 2) is based on a pre-trained ResNet50 model [43]. ResNet was chosen as a pre-trained version was readily available with PyTorch and it has also been successfully used for video/image

classification tasks in other works [44, 45, 46]. The tunable parameters of the ResNet model were frozen during training and used to convert frames of the input video into a feature vector. The ResNet feature vector is then passed to two fully connected layers to produce a prediction.

3.3. Late Fusion Model

The second model implemented is based on the multimodal model used in [17], but with some adjustments to suit the target task being performed. Figure 3 illustrates the model implemented. It consists of two sub-networks, a video sub-network that uses ResNet50, and a CNN audio sub-network that processes log-mel spectrograms. The sub-networks outputs are naively fused before passing them through the three fully connected layers for classification.

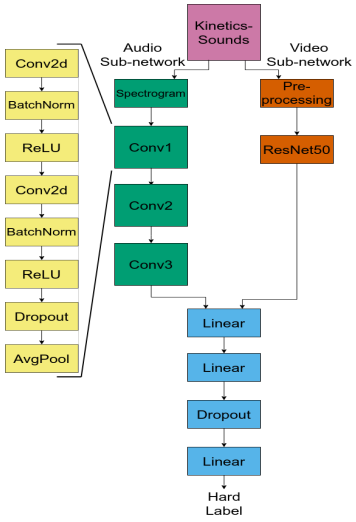


Figure 3: Late Fusion Model consisting of a ResNet50 video network and a CNN audio network. Left: CNN audio sub-network. Centre: Late fusion and output. Right: ResNet video sub-network

Kinetics-Sounds - Training, Inference and Missing Modality Procedure This section describes how the Kinetics-Sounds model was trained, tested and then tested with missing modalities.

The multimodal late-fusion model is first trained using video and audio input. The video sub-network produces a tensor of size 400x1 and the audio sub-network produces a tensor of size 512x1. These tensors are fused to produce a tensor of size 912x1 which is passed to the first fully connected layer. The multimodal late-fusion model is now trained.

For the purpose of evaluation, the individual sub-networks are trained. It is important to note that the trained sub-networks are identical to those described in Figure 3. The **only difference** between the full multimodal models and the individual sub-networks that are trained is the **size of the first fully connected layer**. For training the video sub-network the size of the first fully connected layer is 400x1 and for the audio sub-network, it is 512x1. This is the only difference. All three models, multimodal, video-only and audio-only, are now trained.

The following steps happen at inference time. All three models are evaluated with no missing modalities to provide baselines. Then to evaluate how the multimodal model uses the modalities a leave-one-out approach is used. To evaluate the model when there is only the video modality available, the video feature gets fused with either Gaussian noise or silence at the fusion stage. Then to evaluate the model when there is only the audio modality available, the audio feature is fused with either Gaussian noise or a black video. Full details on the parameters of all the models trained as part of this experiment can be found in the Appendix.

4. Experiments

This section describes the objectives of each experiment and the methodology followed when carrying them out.

Experiment 1: Binary Classification The main objectives of this experiment are to understand to what extent a multimodal model utilises each input modality to perform a prediction and to investigate how the underlying distribution of relevant information changes depending on the target task when only the target label changes.

A binary classifier was trained as a starting point for understanding modality relevance. The aim of this experiment was to investigate which modality was the most relevant to the target task. The multimodal knowledge distillation model described in 3.2 was trained using a 70/20/10 split. The individual modalities are passed to their respective teacher model for each audio and video pair used for training the multimodal model. The teacher models produce a soft label for each modality, which is then used in the loss function of the multimodal model. The multimodal loss function is the summation of the classification loss of the multimodal model for the given sample and the distillation loss for each modality. The distillation loss measures how close the multimodal prediction was compared to that of a teacher model. The Kullback-Leibler divergence loss was used as the distillation loss function.

This process is repeated for both the urban/rural and speech/non-speech subsets. It is important to note that these subsets are identical except for the label associated

with each sample. By using the same data set entries but with different labels, we were also able to investigate how the underlying distribution of relevant information changes depending on the target task.

The trained models were then evaluated under the metrics of accuracy, balanced accuracy, precision, recall and F1-score when both modalities are present. To then determine which/if modality was more dominant, a leave-one-out approach was adopted. The multimodal model was evaluated when only the video was available and then when only the audio was available.

Experiment 2: Multi-class Classification This experiment aims to investigate further how a multimodal model uses the information within and across modalities to perform a multi-class classification task and to gain insights into how the multimodal model used modality information during the training process.

In this experiment, the model described in section 3.3 was trained jointly using both modalities. Both modalities were passed to their respective sub-networks before being fused to produce a 912-D (400-D ResNet, 512-D Audio CNN) feature and passed to the fully connected layers for prediction. The multimodal classifier used a negative log-likelihood loss function.

Alongside the joint training of the multimodal network, the individual networks were trained for the target task to provide an unimodal comparison. The unimodal models were trained by connecting the sub-networks to the fully connected layers of the model, but the first fully connected model had a size equal to that of the sub-network output. In the case of the video model, the first fully connected layer had a size of 400-D and for the audio model, a size of 512-D.

The trained multimodal model was then evaluated under the metrics of accuracy, balanced accuracy, precision, recall and F1-score when both modalities were present and when only one of them was present.

Handling Missing Modalities Two options for representing missing data were used to evaluate the performance of the multimodal models when a modality is missing: Gaussian noise and "appropriate defaults". In the case of the video modality, a black image was passed to ResNet50 to produce a blank feature, with the intuition being that a monochromatic image contains little to no information, i.e. as close to it really being missing. Then, for the audio model, a tensor of all zero values, i.e. silence, was used, again the intuition being that silence is as close to it being missing.

Baselines The baselines used for the experiments are the unimodal models used to train the multimodal models. The unimodal models were selected baselines for two reasons: 1) To investigate the efficacy of multimodal learning over unimodal learning for the same task. 2) To investigate how multimodal models utilise and weigh the features and representations created by the modality sub-networks. The multimodal models trained in the paper contain the unimodal models in their architecture.

5. Results

The following section presents the results obtained from running the experiments outlined in section 4. Each set of results provides an insight into how the models are using the modalities given to them and how, even for the same input data for the same model, modalities will be utilized in different ways and to different extents. This section is divided into results for experiment one and experiment two. Each set of results is followed by a discussion of those results.

Binary Classification of Urban vs. Rural - All Modalities, Audio-only Model and Video-only Model Table 1 presents the results of the urban vs. rural binary classifiers. Firstly, for the unimodal, expert models, it is observed that both models learn a moderate representation of each modality. While the metrics for the audio model are significantly less than the video model, they are still better than a random guess. Secondly, for the multimodal model, it is observed that the performance lies between the two expert models. Knowledge distillation aims to distil "useful" information to the student learners. However, this information is not always beneficial to the student models' performance, as indicated by the results presented. The multimodal model achieves a reasonable performance that is an improvement over using just audio.

Table 1
Results for the unimodal and knowledge distillation (KD) multimodal models for classifying Urban vs. Rural.

Urban & Rural	Acc.	Balanced Acc.	Precision	Recall	F1 Score
Audio Model	61.62	61.79	63.38	61.69	60.51
Video Model	77.72	77.71	77.71	77.71	77.72
AV Model	71.66	71.65	71.71	71.69	71.66
Gaussian Audio	**70.84**	**70.91**	**71.28**	**70.91**	**70.71**
Gaussian Video	61.76	61.81	61.93	61.81	61.68
Silent Audio	**71.66**	**71.66**	**71.67**	**71.67**	**71.66**
Black Video	61.76	61.76	62.45	61.87	61.35

Binary Classification of Urban vs. Rural - Missing Modalities A different understanding emerges when a modality is withheld from the multimodal model at inference time. A minor decrease, highlighted in bold, in the performance metrics is observed when audio data is missing, whereas when either Gaussian noise or a black image is used to represent missing video data, a more significant decrease ($\approx$ 10%) in performance metrics is

observed. The different decreases in the performance metrics provide insight into the shared representation of audio and video created by the model. From the audio model, we know a moderate representation can be learned. However, this does not appear to have transferred to the multimodal model. The minor decrease in the metrics caused by missing audio data indicates that the multimodal model weighted the video input more than the audio, to the point where it appears to almost entirely **ignores the contributions** of the audio modality. This occurs even though the audio model itself is capable of performing the classification. These results, combined with a significant drop in metrics for missing video input, provide evidence for the impacts of imbalanced modality information on the training of multimodal models.

Binary Classification of Speech vs. Non-Speech - All Modalities, Audio-only Model and Video-only Model Table 2 presents the results of the speech vs. non-speech classifiers. Each of the unimodal experts achieves a very similar level of performance across all metrics. Due to the label imbalance in the speech vs. non-speech data set the balanced accuracy should be used instead of accuracy. When compared under balanced accuracy, the audio model outperforms the video model by less than 2%. Both experts appear to be equal in their ability to classify samples as speech or non-speech. This carries forward to the multimodal model, which, like the urban vs. rural classifiers, sees performance metrics in between that of the experts. It is worth noting again that the models and data used for the speech vs. non-speech task **are identical to that of the urban-rural task**.

Table 2
Results for the unimodal and knowledge distillation (KD) multimodal models for classifying Speech vs. Non-Speech.

Speech & Non-Speech	Acc.	Balanced Acc.	Precision	Recall	F1 Score
Audio Model	68.36	65.32	66.73	65.32	65.64
Video Model	67.54	63.67	65.95	63.37	66.31
AV Model	**68.09**	**67.28**	**66.89**	**67.28**	**67.02**
Gaussian Audio	64.92	61.63	62.79	61.63	61.80
Gaussian Video	63.55	63.46	62.92	63.46	63.92
Silent Audio	66.85	65.53	65.40	65.53	65.46
Black Video	63.41	62.87	62.44	62.87	62.47

Binary Classification of Speech vs. Non-Speech - Missing Modalities The results of the multimodal model and those of the multimodal model with missing modalities indicate that the shared representation created for the task of speech vs. non-speech uses information from the different modalities more equally compared to the urban vs. rural classifier. When a modality is missing, regardless of how it is missing, no major drop in the performance metrics is observed. It is observed, however, that missing the audio modality results in a larger decrease in the performance metrics. This indicates that while the model created a more balanced shared representation, it still favours using the audio data. Contrasting to the trained model for the urban vs. rural task, the speech vs. non-speech model **has not ignored** the contributions of the input modalities. This happens when using the exact same input data and model architecture, with the only difference being the target label. This shows that depending on the target task of the model, **the underlying distribution of relevant information changes** and the model will make use of the modalities differently. The results also provide early empirical evidence that the choice of how a missing modality is represented can impact performance. For the speech vs. non-speech task, using Gaussian noise to represent missing video and silence to represent audio resulted in less of a decrease in the performance metrics. Based on the results of the binary classifiers the **choice of how "nothing" is represented is also task dependent**.

Multi-class Classification - All Modalities, Audio-only and Video-only Models Table 3 presents the results of the multi-class Kinetics-Sounds classifiers. It is worth noting that the chance of correctly guessing a class at random for the Kinetics-Sounds data set is 3.8%. The results provide another view on modality relevance, shared representation and the challenge of imbalanced modality information. The unimodal audio classifier achieved a final accuracy of 16.2%, which is better than guessing at random and shows that the model implemented is capable of learning a representation of the audio data. However, in stark contrast, the unimodal video classifier achieved consistent metrics of $\approx$ 77%.

Table 3
Results for the unimodal models and late fusion model for the Kinetics-Sounds data set.

Kinetics Sounds	Acc.	Balanced Acc.	Precision	Recall	F1 Score
Audio Model	**16.22**	**16.04**	**42.18**	**16.04**	**11.73**
Video Model	77.2	76.73	78.27	76.73	76.92
AV Model	77.35	76.89	78.36	76.89	77.04
Gaussian Audio	76.65	76.05	77.33	76.05	76.20
Gaussian Video	**5.94**	**6.07**	**20.27**	**6.07**	**4.10**
Silent Audio	76.65	76.05	77.33	76.05	76.20
Black Video	**6.29**	**6.88**	**83.61**	**6.88**	**2.82**

Multi-class Classification - Missing Modalities Building from these unimodal results, the multimodal model, trained by means of late fusion, marginally outperforms the unimodal video model. However, a similar phenomenon to that of the video-only results of the urban vs. rural classifier is observed, i.e. there is **no meaningful decrease in the performance metrics when the audio modality is missing**.

Like the urban vs. rural classifier, the results of the unimodal audio demonstrate that the model can learn from the audio, but when there is no audio modality

present, the multimodal model still performs the target task with the same accuracy as having the audio data present. This again indicates the final multimodal model **did not balance the information from both modalities**. This imbalanced use of modality information is further exemplified by the significant drop in the performance metrics when the video modality is missing. **A drop of** $\approx 70\%$ **is observed** for all metrics excluding precision, which dropped by $\approx 60\%$. There is no difference recorded in using Gaussian noise versus silence for representing missing audio. There is a minor difference recorded across all metrics, excluding precision, between using Gaussian noise or a black image.

To better understand the performance of the models, Figure 4 presents a breakdown of the accuracies per label for the multimodal model, the multimodal model with missing audio or video (silence and black video) and the audio-only and video-only models. The results clearly show the disparity between how the multimodal model treats audio versus video.

In Figure 4 it can be seen for the majority of the classes the multimodal model and the audio-only model (which is a sub-network of the multimodal model) have not represented the audio modality in the same way. The classes in which the audio-only model showed the best performance ("playing bagpipes", "playing bass guitar", and "playing trombone") still had good performance even when the audio was missing in the multimodal model. The majority of the remaining classes were not accurately predicted by the audio model but were predicted with reasonable to high accuracy by the multimodal model, even when audio was missing. This indicates again that for such classifications, the **multimodal model relies on the video**. The opposite is then observed with the video-only model. It accurately predicts the majority of classes and outperforms the multimodal model for a significant number of classes, but then when the video is missing there is a sharp decrease in accuracy. This also indicates that the multimodal model relies on the video modality.

From the numerical results presented in Tables 1, 2 and 3 and the more granular breakdown of the models in Figures 4 is clear that the way in which the models use information from the various modalities **is imbalanced** and that in the Kinetics-Sounds model, **the audio modality was almost entirely ignored**. This imbalance appears to be the result of a small change (the size of the fully connected layer), that you would not expect to have such an impact, but does in fact, cause wildly different results when multimodal joint learning is happening. **The assumption that if we have a "good" video network and a "good" audio network, they can create a good Audio-Visual network is incorrect.**

6. Conclusion

This paper presents experimental results that provide insight into how multimodal machine-learning models use information from different modalities. To achieve this, two different multimodal models were implemented for two audio-visual data sets. The models were then evaluated with each modality missing to determine how the models used information from the modalities.

The results of the experiments demonstrate the **inherent imbalanced use of information in multimodal models** and that there is no guarantee that multimodal models will use the information from the input modalities equally. The results also show that it is possible for such a model to almost **entirely ignore the contributions of a modality**, even if the individual modality is capable of carrying out the target task. This is most noticeably observed in the urban vs. rural and Kinetics-Sounds classification tasks. The combination of the urban vs. rural and speech vs. non-speech classification tasks empirically demonstrates that depending on the target task being performed, **the amount of relevant information present within a modality will change**. We also provide initial results that show the potential impact of **how "nothing" is represented for modalities**.

Using the results and insights gained from carrying out this research it can be extended in several ways. The most pressing future research relates to the significant difference in the performance metrics for the late-fusion model with Audio Only and the unimodal Audio model. It will involve investigating its root cause(s) by further analysing the results and models. The results presented in this paper showed poor performance for the audio-only model but good enough for the purpose of the experiment. A reasonable assumption is that an audio model should be able to classify the various instruments and sound events better. Therefore, a subjective study will be carried out to get a baseline for the expected performance of the audio model for the Kinetics-Sounds classification task. This study will have participants relabel some of the data set based on only listening to the audio and then again but with only the video. A similar subjective study will be carried out on the AudioSet data set as it was noted in Section 3.1 that there is a possibility of modality bias in the data set. The experiments described in this paper will also be extended to train the AudioSet tasks using the late fusion model and the Kinetics-Sounds task using knowledge distillation. Further investigating the distributions of relevant information per modality could provide insight into how to best exploit these distributions during the training stage of multimodal models. The experimental results in this paper highlight the impact of imbalanced modality information and the need for methods to address it.

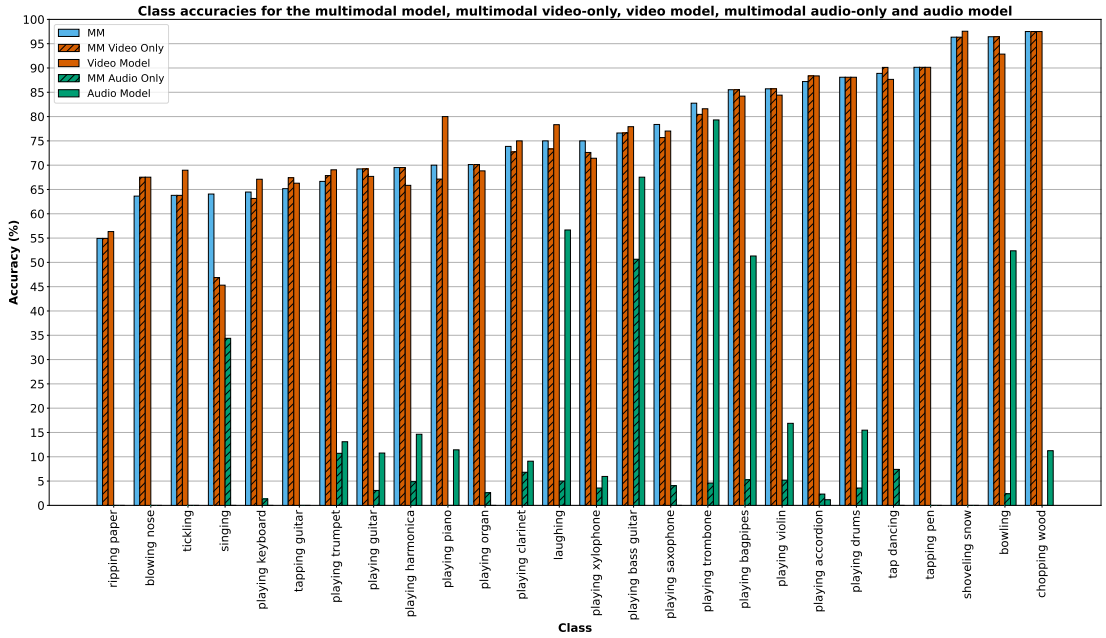


Figure 4: Breakdown of accuracies per class in Kinetics-Sounds for the multimodal model, the multimodal model with video input only, video sub-network model, multimodal with audio input only and audio sub-network model.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (d-real) under Grant No. 18/CRT/6224.

References

[1] R. Kiros, R. Salakhutdinov, R. S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, 2014. URL: https://arxiv.org/abs/1411.2539. doi:10.48550/ARXIV.1411.2539.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 26, Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

[3] W. Wang, R. Arora, K. Livescu, J. Bilmes, On deep multi-view representation learning, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, p. 1083–1092.

[4] A. Dutt, A. Zare, P. Gader, Shared manifold learning using a triplet network for multiple sensor translation and fusion with missing data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 15 (2022) 9439–9456. doi:10.1109/JSTARS.2022.3217485.

[5] B. Wehbe, N. Shah, M. Bande, C. Backe, Sonar-to-rgb image translation for diver monitoring in poor visibility environments, in: OCEANS 2022, Hampton Roads, 2022, pp. 1–9. doi:10.1109/OCEANS47191.2022.9977024.

[6] P. Poklukar, M. Vasco, H. Yin, F. S. Melo, A. Paiva, D. Kragic, Geometric multimodal contrastive representation learning, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 17782–17800. URL: https://proceedings.mlr.press/v162/poklukar22a.html.

[7] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, A. Zisserman, Self-supervised multimodal versatile networks, Advances in Neural Information Processing Systems 33 (2020) 25–37.

[8] Y. Cheng, X. Zhao, K. Huang, T. Tan, Semi-supervised learning and feature evaluation for rgb-d object recognition, Computer Vision and Image

Understanding 139 (2015) 149–160.

[9] C. M. Christoudias, K. Saenko, L.-P. Morency, T. Darrell, Co-adaptation of audio-visual speech and gesture classifiers, in: Proceedings of the 8th international conference on Multimodal interfaces, 2006, pp. 84–91.

[10] M. Federici, A. Dutta, P. Forré, N. Kushman, Z. Akata, Learning robust representations via multi-view information bottleneck, arXiv preprint arXiv:2002.07017 (2020).

[11] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, 2013. URL: https://arxiv.org/abs/1304.5634. doi:10.48550/ARXIV.1304.5634.

[12] M. R. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views - an application to multilingual text categorization, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems, volume 22, Curran Associates, Inc., 2009. URL: https://proceedings.neurips.cc/paper/2009/file/f79921bbae40a577928b76d2fc3edc2a-Paper.pdf.

[13] X. Sun, Y. Xu, P. Cao, Y. Kong, L. Hu, S. Zhang, Y. Wang, Tcgm: An information-theoretic framework for semi-supervised multi-modality learning, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, Springer International Publishing, Cham, 2020, pp. 171–188.

[14] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, L. Huang, What makes multi-modal learning better than single (provably), in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., 2021, pp. 10944–10956. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/5aa3405a3f865c10f420a4a7b55cbff3-Paper.pdf.

[15] Y. Huang, J. Lin, C. Zhou, H. Yang, L. Huang, Modality competition: What makes joint training of multimodal network fail in deep learning? (Provably), in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 9226–9259. URL: https://proceedings.mlr.press/v162/huang22e.html.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in: Proc. IEEE ICASSP 2017, New Orleans, LA, 2017.

[17] R. Arandjelovic, A. Zisserman, Look, listen and learn, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.

[18] C. Zhang, Z. Han, y. cui, H. Fu, J. T. Zhou, Q. Hu, Cpm-nets: Cross partial multi-view networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper/2019/file/11b9842e0a271ff252c1903e7132cd68-Paper.pdf.

[19] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: A survey and taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (2019) 423–443. doi:10.1109/TPAMI.2018.2798607.

[20] V. Rajan, A. Brutti, A. Cavallaro, Robust latent representations via cross-modal translation and alignment, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 4315–4319. doi:10.1109/ICASSP39728.2021.9413456.

[21] Z. Quan, T. Sun, M. Su, J. Wei, X. Zhang, S. Zhong, Multimodal sentiment analysis based on nonverbal representation optimization network and contrastive interaction learning, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2022, pp. 3086–3091. doi:10.1109/SMC53654.2022.9945514.

[22] S. Jose, R. H. Ngouna, K. T. Nguyen, K. Medjaher, Solving time alignment issue of multimodal data for accurate prognostics with cnn-transformer-lstm network, in: 2022 8th International Conference on Control, Decision and Information Technologies (CoDIT), volume 1, 2022, pp. 280–285. doi:10.1109/CoDIT55151.2022.9804090.

[23] S. Li, Y. Xie, G. Wang, L. Zhang, W. Zhou, Adaptive multimodal fusion with attention guided deep supervision net for grading hepatocellular carcinoma, IEEE Journal of Biomedical and Health Informatics 26 (2022) 4123–4131. doi:10.1109/JBHI.2022.3161466.

[24] T. Marteau, D. Sodoyer, S. Ambellouis, S. Afanou, Level fusion analysis of recurrent audio and video neural network for violence detection in railway, in: 2022 30th European Signal Processing Conference (EUSIPCO), 2022, pp. 563–567. doi:10.23919/EUSIPCO55093.2022.9909622.

[25] Y. Xie, J. Tian, X. X. Zhu, A co-learning method to utilize optical images and photogrammetric point clouds for building extraction, International Journal of Applied Earth Observation and Geoinformation 116 (2023) 103165. URL: https://www.sciencedirect.com/science/article/pii/S1569843222003533. doi:https://doi.org/10.1016/j.jag.2022.103165.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: ICML, 2011.

[27] D. Khattar, J. S. Goud, M. Gupta, V. Varma, Mvae: Multimodal variational autoencoder for fake news

detection, in: The World Wide Web Conference, WWW '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 2915–2921. URL: https://doi.org/10.1145/3308558.3313552. doi:10.1145/3308558.3313552.

[28] D. Park, Y. Hoshi, C. C. Kemp, A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder, IEEE Robotics and Automation Letters 3 (2018) 1544–1551. doi:10.1109/LRA.2018.2801475.

[29] S. Kutuzova, O. Krause, D. McCloskey, M. Nielsen, C. Igel, Multimodal variational autoencoders for semi-supervised learning: In defense of product-of-experts, 2021. URL: https://arxiv.org/abs/2101.07240. doi:10.48550/ARXIV.2101.07240.

[30] Y. Zhang, Y. Chen, C. Gao, Deep unsupervised multi-modal fusion network for detecting driver distraction, Neurocomputing 421 (2021) 26–38. URL: https://www.sciencedirect.com/science/article/pii/S0925231220314302. doi:https://doi.org/10.1016/j.neucom.2020.09.023.

[31] M. Arar, Y. Ginger, D. Danon, A. H. Bermano, D. Cohen-Or, Unsupervised multi-modal image registration via geometry preserving image-to-image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[32] M. J. Patwary, W. Cao, X.-Z. Wang, M. A. Haque, Fuzziness based semi-supervised multimodal learning for patient's activity recognition using rgbdt videos, Applied Soft Computing 120 (2022) 108655. URL: https://www.sciencedirect.com/science/article/pii/S1568494622001326. doi:https://doi.org/10.1016/j.asoc.2022.108655.

[33] X.-Z. Wang, R. A. R. Ashfaq, A.-M. Fu, Fuzziness based sample categorization for classifier performance improvement, Journal of Intelligent & Fuzzy Systems 29 (2015) 1185–1196. URL: https://doi.org/10.3233/IFS-151729. doi:10.3233/IFS-151729, 3.

[34] H. Hotelling, Relations Between Two Sets of Variates, Springer New York, New York, NY, 1992, pp. 162–190. URL: https://doi.org/10.1007/978-1-4612-4380-9_14. doi:10.1007/978-1-4612-4380-9_14.

[35] G. Andrew, R. Arora, J. Bilmes, K. Livescu, Deep canonical correlation analysis, volume 28, PMLR, 2013, pp. 1247–1255. URL: https://proceedings.mlr.press/v28/andrew13.html.

[36] Q. Tan, G. Yu, C. Domeniconi, J. Wang, Z. Zhang, Incomplete multi-view weak-label learning., in: Ijcai, 2018, pp. 2703–2709.

[37] Q. Wang, L. Zhan, P. Thompson, J. Zhou, Multimodal learning with incomplete modalities by knowledge distillation, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1828–1838. URL: https://doi.org/10.1145/3394486.3403234. doi:10.1145/3394486.3403234.

[38] W. Wang, D. Tran, M. Feiszli, What makes training multi-modal classification networks hard?, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[39] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, 2017. arXiv:1705.06950.

[40] H. Dubey, D. Emmanouilidou, I. J. Tashev, Cure dataset: Ladder networks for audio event classification, in: 2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), 2019, pp. 1–6. doi:10.1109/PACRIM47961.2019.8985061.

[41] Q. Kong, C. Yu, Y. Xu, T. Iqbal, W. Wang, M. D. Plumbley, Weakly labelled audioset tagging with attention neural networks, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 1791–1802. doi:10.1109/TASLP.2019.2930913.

[42] G. Tang, R. Liang, Y. Xie, Y. Bao, S. Wang, Improved convolutional neural networks for acoustic event classification, Multimedia Tools and Applications 78 (2019) 15801–15816. URL: https://doi.org/10.1007/s11042-018-6991-4. doi:10.1007/s11042-018-6991-4.

[43] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. doi:10.48550/ARXIV.1512.03385.

[44] D. Kondratyuk, L. Yuan, Y. Li, L. Zhang, M. Tan, M. Brown, B. Gong, Movinets: Mobile video networks for efficient video recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16020–16030.

[45] D. Tran, H. Wang, L. Torresani, M. Feiszli, Video classification with channel-separated convolutional networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019.

[46] M. E. Kalfaoglu, S. Kalkan, A. A. Alatan, Late temporal modeling in 3d cnn architectures with bert for action recognition, in: A. Bartoli, A. Fusiello (Eds.), Computer Vision – ECCV 2020 Workshops, Springer International Publishing, Cham, 2020, pp. 731–747.