Deep Learning Multimodal Fusion for Sentiment Analysis - Coursework 4

Group 25 (s1738075, s1427590, s1211898)

Abstract

We present our work on deep neural network (DNN) multimodal fusion for binary sentiment classification using the standardized CMU-MultimodalDataSDK MOSI dataset of text, audio, and visual features extracted from YouTube video movie reviews. We developed and compared three approaches to DNN multimodal fusion: (1) input-level feature fusion, (2) intermediate-level feature fusion, and (3) decision-level fusion. We also experimented with principle component analysis (PCA) for dimensionality reduction to clarify which features were most significant for each modality and find that it increases unimodal performance. Our results were measured on a held-out test set using accuracy. For input-level feature fusion, we showed that our best performance was obtained using bimodal video+text data without PCA and using a 2-layer bidirectional Long Short-Term Memory (BLSTM) DNN architecture with 72.4% accuracy. Our experiments also indicated that bimodal audio+visual consistently under-performs comparative to the other combinations of modalities. Our overall best system was trimodal intermediate-level feature fusion, where weights are merged from each modality during training with subsequent additional training, which achieved an overall accuracy of 73.4%. We show that multimodal fusion outperforms the individual unimodal classifiers.

1. Introduction

Sentiment analysis provides important tools in attempting to uncover the underlying attitude that one holds towards a certain entity. For a long time, text-based sentiment analysis has been the staple in this area and only recently are other modalities being considered for sentiment analysis such as vision and speech (Pérez-Rosas et al., 2013; Wöllmer et al., 2013; Poria et al., 2015c). For text channels, the features usually include information about word sequences and meaning (Mikolov et al., 2013). In visual data, features involve salient points of the face or body (Zadeh et al., 2016a). In audio data, low-level descriptors are collected from the speech signal such as pitch and volume (Zeng et al., 2009). The combination of features which have originated from text, speech and audio is what forms the basis of our multimodal classification work. Features from each modality are modeled, learned, and eventually *fused* together at various levels in a classification Deep Neural Network (DNN) system. When the modalities are fused together, this is called *multimodal fusion*.

DNN multimodal fusion for binary sentiment classification is an active area of research that continues to gain momentum and spark interest due to the challenging nature of the problem (Cambria et al., 2017; Gunes & Piccardi, 2005; Zadeh et al., 2016b; 2017; Poria et al., 2018). In fact, currently there is a new 2018 ACL Workshop multimodal emotion and sentiment analysis shared-task and they are conducting a competition for system performance on a standardized dataset which is very similar to the one that we have used in this work [1].

To expand from our previous coursework on unimodal binary sentiment classification (G25, 2018), in this coursework we have explored the interplay between modalities with our work on DNN multimodal fusion. We focused our attention on three fusion techniques applied to three modalities: text, video, and audio. We developed these techniques inspired from previous work on multimodal fusion including Poria et al. (2018) and Zadeh et al. (2016b).

Mainly, our motivation for fusion techniques is that the audio and text modalities can bring additional information to ambiguous cases. For example, a smile extracted from facial features could help disambiguate cases such us "This movie is sick" - text alone would have trouble interpreting the meaning of the word "sick" in this context.

Our approaches to multimodal fusion are:

1. Input-level features fusion
2. Intermediate features fusion
3. Decision-level fusion (late fusion)

Our first method refers to fusing information at the level of input features, similar to an unweighted concatenation of feature vectors, and it is the most widely used. The second method evokes the notion that each modality can be learned using a unimodal DNN. The weights learned through training each unimodal DNN are concatenated together and training continues before the decision level. The third method, also known as *ensemble fusion* or *late fusion*, fuses multiple modalities at the decision level.

We present our multimodal DNN fusion approaches in detail in our methodology description in Section 3. where we

[1] http://multicomp.cs.cmu.edu/acl2018multimodalchallenge/

further analyze the interactions between modalities. In the current work we have experimented with combinations of modalities as well as system architectures that attempt to capture the interplay between modalities. We advance our previous work by building a multimodal fusion framework for analyzing sentiment.

This paper is organized as follows: in Section 2, we present an overview of related work including performance on this task for existing systems from other researchers. In Section 3 we discuss our methodology including information about our task and data, details about our training hyperparameters, and a description of our machine learning architectures. In Section 4, we present our experiment results. In Section 5, we discuss and analyze the results and prediction decisions. Finally, Section 6 concludes and offers some ideas for future work.

2. Related Work

Sentiment analysis has traditionally been a task for natural language processing and based explicitly on text data, such as online blog posts (Feng et al., 2011). Beyond the scope of text-based sentiment analysis there is the work of Chen et al. (1998) that provides us with an early work on audio-visual emotion recognition. They also showed that bimodal classifiers can perform better than unimodal ones alone.

Schuller (2011) and Wöllmer et al. (2013) approached the sentiment analysis paradigm through fusing audio and visual information at both feature level and decision level. Through their experiments they demonstrated that the audio modality performs better than the visual one. However, their task differed slightly from ours. They predicted emotion as an overlapping multi-class problem whereas our work predicts binary sentiment. While it may be true that audio features have an important role in their task, we know that the state of the art for our MOSI dataset has demonstrated audio features to be the worst predictor (Zadeh et al., 2017), which our work also explores.

Morency et al. (2011) was one of the first to investigate sentiment analysis on video movie reviews. They analyzed a collection of 47 videos depicting monologues in addition to the corresponding text that they manually transcribed from each 30-seconds excerpt. They evaluated sentiment for each review as a 3-way classification problem: positive, negative or neutral and achieved an F1 measure of 55.3%, which is much better than chance. Furthermore, Wollmer et al. (2013) attempts the same type of multimodal sentiment task for movie reviews through the usage of a linear Support Vector Machines (SVM) for the linguistic features and a Bidirectional Long Short-Term Memory (BLSTM) for the audiovisual ones. Our work continues in this direction of combining data from different modalities and we also used video movie reviews. However, these related studies used very small collections of videos, whereas our work uses more than 2,000 videos.

Even though there is a significant amount of research done on audio-visual emotion recognition, only a few previous works have systematically explored trimodal fusion by combining text data with audio and visual modalities. Rozgic et al. (2012) fuses the visual, textual and audio data at the input-feature level. Poria et al. (2015b) presents a novelty in how it uses a deep Convolutional Neural Network (CNN) to extract features from the text modality and then adopts multiple kernel learning (MKL) for classifying the multimodal fused feature vectors. All of these approaches arrive at the conclusion that multimodal classifiers perform better than unimodal ones, which motivates our interest in exploring multimodal fusion in the current work.

More recently, (Hu & Flaxman, 2018) tried to predict emotion in a holistic manner by inferring the latent emotional state of a Tumblr [2] user rather than predicting if a particular sentence expresses negative or positive sentiment. Also, Poria et al. (2018) presented three fusion techniques for achieving high accuracy when dealing with multimodal data: concatenation-based fusion, context-aware fusion and context-aware fusion with attention. Gunes & Piccardi (2005) had also compared feature fusion and decision fusion. One of the main problems of early fusion is that input-level feature level concatenation will increase the feature space, which can be potentially problematic for very large datasets. To account for this in our work, we employed a dimensionality reduction technique.

There has been recent work on multimodal fusion techniques for binary sentiment classification of YouTube movie reviews using the CMU-MultimodalDataSDK MOSI (Zadeh et al., 2018) dataset, the same dataset that we used in our work. Comparative performance of the top-performing existing systems is shown in Table 1, measured by classification accuracy. The state-of-the-art is Zadeh et al. (2017) which used a tenor-based approach to multimodal fusion as we described in our previous coursework (G25, 2018). The C-MKL system from Poria et al. (2015b), as mentioned above, used a novel approach with CNNs but it does not perform as well as the tensor fusion approach . We also mention a non-DNN system from Zadeh et al. (2016b), which we included in this comparison because it presents performance results based on input-level feature fusion.

System	Authors	Accuracy
TFN	Zadeh et al. (2017)	77.1
C-MKL	Poria et al. (2015a)	73.1
SVM-MD	Zadeh et al. (2016b)	71.6

Table 1. System accuracy results on MOSI dataset for binary sentiment classification using trimodal data (audio, video, and text). Note each of these authors has used slightly different feature extraction and pre-processing.

However, in each of these related works that had used the MOSI dataset researchers extracted features directly from the raw YouTube videos. On the other hand, our current work used features that were already extracted, pre-

[2] https://www.tumblr.com/

processed and made standard for the MOSI dataset that is available from Zadeh et al. (2018). The features that we used in the standardized MOSI dataset are very similar to the features used by others. However, it is possible that one or more of the pre-processing techniques has introduced artifacts and other inconsistencies such as different train/test split, therefore we cannot make a direct system-to-system comparison between our methods and previous work.

3. Methodology

In this section, we provide a brief overview of our data and task, as this was described in more detail in our previous coursework (G25, 2018). Additionally, we provide the technical specifications of the DNN architectures and parameters that we used in this work, followed by details about our three fusion techniques. Finally, we discuss PCA dimensionality reduction, which we employed for our early fusion technique experiments.

3.1. Data and Task Description

In our previous coursework, we used the Multimodal Opinion level Sentiment Intensity (MOSI) dataset from CMU-MultimodalDataSDK (Zadeh et al., 2018). We examined whether or not single modality data features were predictive of binary sentiment classification for YouTube video movie reviews (G25, 2018).

The benefits of using the MOSI dataset were three-fold. First, the developers had already extracted features from the raw video for each of the three different modalities in order to standardize the dataset. Second, there is now a definitive train, validation, and test split of the data. Finally, the developers also provided a way to align text, acoustic and visual data. For these reasons the MOSI dataset allows for a meaningful comparison across systems, something that is important for the current work. A more detailed description of text, audio, and visual features, as well as details about the sentiment class labels can be found in our earlier coursework (G25, 2018) and Zadeh et al. (2018).

The MOSI dataset is a collection of 2199 opinion video clips. Each video is annotated with sentiment data in the range [-3,3]: strongly positive (labeled as +3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3). The multimodal observations consist in transcribed speech and features extracted from the visual and audio data.

The standardized MOSI dataset can be downloaded using the CMU-MultimodalDataSDK [3], which also provides pre-processed features and a way to align text, acoustic and visual data. Being able to align modalities will be important for our fusion techniques. As described in our previous coursework, we align the features using text modality as a reference and we normalize the feature values on a per-modality basis. Due to the different number of timesteps in each utterance, we need to restrict each sentence to a fixed size length, either by padding our cropping sentences. This length becomes a hyper-parameter that we will explore.

Our prediction task is binary classification for sentiment: positive versus negative. An exemplar with score $s > 0$ belongs to the positive class, while scores of $s < 0$ belong to the negative class. We transformed all scores to True/False values, where True corresponds to the positive class. For performance metrics, we used overall accuracy on the held-out test set. Our experiments were also speaker independent. The train (1283 items) validation (229 items) and test (686 items) set split has been made in such way that no speaker is present in more than one of the sets. This ensures tat we can generalize to unseen utterances.

3.2. Training Hyper-parameters

The activation function we used across all of our experiments was ReLu (Nair & Hinton, 2010). The learning rule was Adam (Kingma & Ba, 2014) with standard parameters. For 1D convolution layers, the kernel size was 3. For max pooling layers, the window size was 2. We varied the number of convolutional layers in $[1, 2, 3]$. For Bi-directional LSTMs, we set the number of units to [64] and the number of layers in $[1, 2, 3]$. For fully connected layers, we varied the number of units in [100, 200] and the number of layers in $[1, 2, 3]$. We added dropout (Srivastava et al., 2014) between fully connected layers with dropout rate in $[0.1, 0.2]$. In all experiments, we used early stopping with the stopping criteria set to identify maximum validation accuracy and patience was set to 10. We varied the maximum length setting for the video segments in our dataset, known as *maxlen*, in $[15, 20, 25, 30]$. The experiments employed batch normalization with batch size set to $b = 64$ (Ioffe & Szegedy, 2015). Since it is a binary classification task, we use a single output unit with sigmoid activation. The loss function we use is binary cross-entropy.

We choose the best model based on validation performance and present test set results measuring overall accuracy.

3.3. Unimodal classifiers

As described in our previous work, various networks can be employed for sentiment prediction on text, audio and video.

CNN: Convolutional Neural Networks (CNNs) are prominent in various sentiment and emotion detection tasks in natural language processing of text (Kim, 2014). In addition, CNNs constitute the core of OpenFace (Baltrušaitis et al., 2016) an open-source face recognition tool, that is also employed by MOSI and is relevant to our work. While there is not much previous work on using CNN architectures for predicting sentiment from speech, we note that others have tried this deep learning approach specifically by working directly on the spectrogram (Niu et al., 2017).

LSTM: We used RNNs in this work, including Long Short-Term Memory (LSTM) as Yin et al. (2017) showed comparable results for both CNNs and RNNs. We also know that

[3] https://github.com/A2Zadeh/CMU-MultimodalDataSDK

LSTMs have been studied with moderate success for video emotion detection Chen et al. (2017). LSTMs are popular with sequence prediction tasks, as they can capture context from previous steps.

Bidirectional LSTM: The employment of Bidirectional LSTMs (BLSTMs) for emotion detection from visual and audio features is becoming more prevalent (Ullah et al., 2018; Ghosh et al., 2016; Lee & Tashev, 2015; **?**; Chernykh et al., 2017). BLSTMs increase the amount of available contextual information. The principle is to use both a forward pass and a backward pass through a sequence.

3.4. Input-Level Feature Fusion

Input-level feature fusion, also known as very early fusion, refers to simply concatenating features from all the modalities, after they have been aligned and transformed to fixed size length. The concatenation is performed on the time step dimension. After input concatenation, the process follows a standard deep learning pipeline. On top of the concatenated features, we can experiment with different deep learning structures.

In our experiments, we varied CNN, LSTM and BLSTM. Finally, and we used one fully connected hidden layer and one output layer for the final sigmoid prediction. In each case, the DNN was trained with the hyper-parameters that we described earlier.

We experimented with dimensionality reduction of each modality, prior to concatenation. This is motivated by the fact that we noted by inspection that many of the visual and audio features were zero valued and we want to be able to extract the most important features.

We show the system architecture with and without dimensionality reduction from principle component analysis (PCA). For a more illustrative understanding, the concept of input-level fusion is displayed in Figure 1.

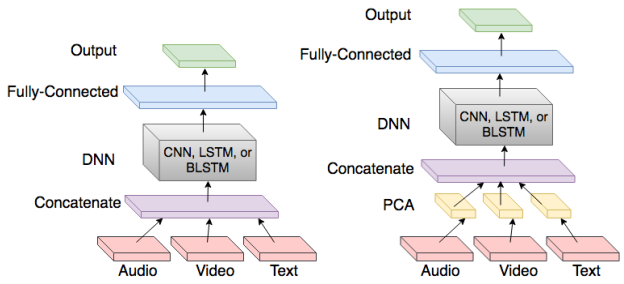

Figure 1. Input-level feature fusion architecture with and without Principle Components Analysis (PCA).

3.5. Intermediate-Level Feature Fusion

In the case of intermediate-level feature fusion, the concatenation is done after some intermediate steps. Thus, data from each modality is given as the input to independent networks which will learn and extract intermediate features. For this step, we chose the best performing unimodal network: thus, for video and audio we use CNN, while for text we have employed BLSTMs.

The intermediate weights from these separate networks are concatenated and from that point we added fully connected layers to train the concatenated features. The goal is to capture interactions between modalities. We refer the reader to Figure 2, for a visual illustration. We experimented with and without PCA on the input-level features.

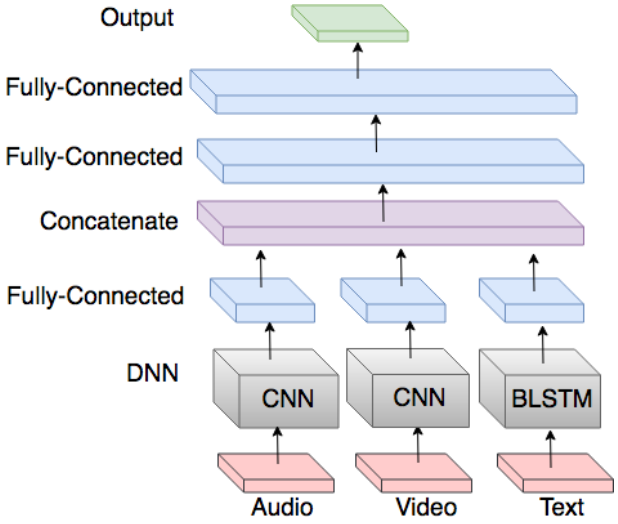

Figure 2. Intermediate-level feature fusion architecture. Not shown in this diagram: PCA for dimensionality reduction.

3.6. Decision-Level Feature Fusion (Late Fusion)

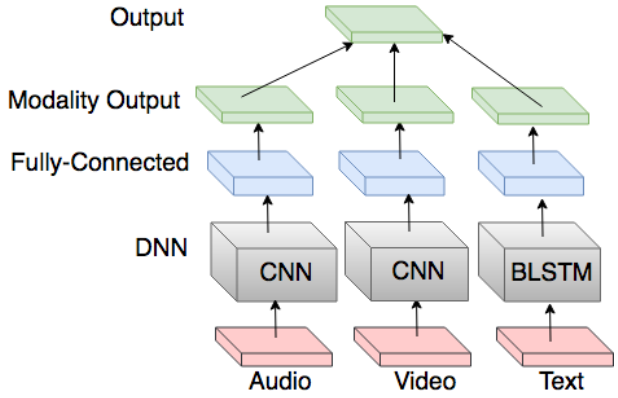

Figure 3. Late decision-level fusion architecture. Not shown in this diagram: PCA for dimensionality reduction.

Decision-level feature fusion is the easiest to understand, as it simply involves using a separate classifier to weight the decisions of DNNs that had been trained independently on each modality, in an ensemble like fashion. The most straightforward way is to train separate classifiers and weight their outputs with a tuple consisting of one weight, λ for each modality as in: $w = (\lambda_1, \lambda_2, \lambda_3)$. These weights can either be learned by another classifier, or set experimentally.

For this fusion technique, there is no concatenation performed. Compared to intermediate level fusion, where the sub-networks where simply extracting intermediate fea-

tures, here we output a decision from each modality. The goal is to improve robustness by combining the results.

Commonly, an SVM or another classifier is used on to of the decisions of each unimodal classifier.Our approach is different from existing literature in that we train the 3 networks together, by pre-training each sub-network and subsequently designing a system that contains the 3 component networks. For an illustration, refer to Figure 3. The top layer of this network is simply an output layer that receives the output of each modality sub-network (so the input is a one dimensional vector of size 3) and assigns a weight for each. This architecture acts as an ensemble of the 3 separate modalities classifiers.

The advantage of training separate classifiers is that the modalities do not need to be aligned because there is no concatenation. It also allows us to choose a best performing architecture for each modality independent of other modalities in the whole network. Although it is not the case for our experiments, in general it is possible to pre-train each modality on a different dataset, if there is more data available (Wu et al., 1999)

3.7. Principal Component Analysis

Principal Component Analysis (PCA) is an important linear transformation technique that is used to perform dimensionality reduction. PCA yields the ordered feature vectors, commonly referred to as *principal components*, which maximize the variance of the data by removing redundant features (Abdi & Williams, 2010). As a data reduction technique, PCA is commonly used for handling high-dimensional visual information. From medical images (K et al., 2017; Ma & Li, 2007) to vehicle and face detection (Sun et al., 2004; Kwak, 2008), PCA proves to be a method that gives very good results in the case of feature selection and extraction.

As PCA extracts low dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. We decided to use this algorithm to reduce the dimensionality of our data and thus attempt to increase the accuracy of our experiments. We chose to apply the PCA algorithm in order to find the best, least redundant components to the unimodal representation of our data since features are semantically different even when they have been min/max normalized between 0 and 1. After using PCA on each of the 3 modalities, individually we can then continue with the classifier training according to the fusion architectures described earlier.

Initially, each of our modalities can be described with a tuple for shape of the form $s = (examples, segments, features)$. More precisely, the shape of the training set for text is $s_{text} = (1283, maxlen, 300)$, for video is $s_visual = (1283, maxlen, 46)$ and for audio is $s_audio = (1283, maxlen, 74)$. After flattening the last two dimensions, we proceed in applying the Python Sklearn PCA decomposition function (Pedregosa et al., 2011) to each training set. By using a scree plot[4] to compute the proportion of variance explained by the number of principal components utilized, we can infer how many components are responsible for a high enough cummulative variance, as seen in Figure 4.

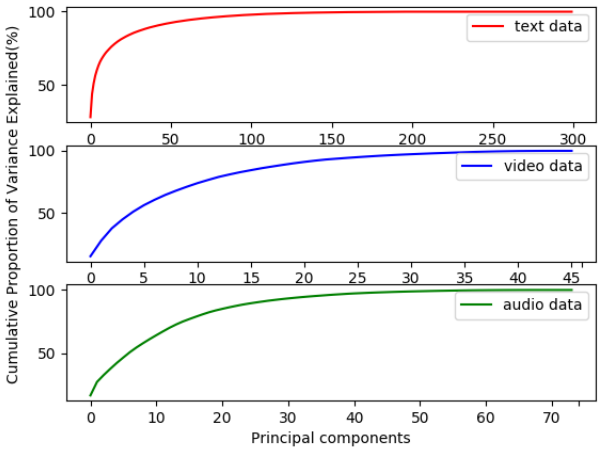

Figure 4. Principal components for every modality and their respective contributions to the cumulative variance

Thus, we choose values for k (where k is the number of principal components) close to a variance of 98%, but we soon discovered that our best accuracy could not achieved when using these particular k values. Only on text modality the best performing DNN is obtained when the variance is 98% as it can be seen in Table 2. We attribute this irregularity to the fact that the audio and visual data are likely noisy with the audio information being the noisiest.

After noticing the noisiness of our data, we proceeded to sweep the following values for k in PCA for unimodal performance, using the same DNN architectures and parameter sweeps as described earlier in the methodology section regarding training hyper-parameters(Section 3.2):

1. For video: $k = [10, 15, 20, 25, 30]$
2. For audio: $k = [10, 15, 20, 25, 30]$
3. For text: $k = [100, 110, 120, 130, 140]$

Finally, we applied the PCA fit that we learned from training data and used it as the PCA transform on our validation and test data. We then examined the unimodal test accuracy on each DNN architecture and the corresponding best value for k in PCA in Table 2. Many of these unimodal results that are using PCA outperformed our best-performing unimodal results from our earlier work (G25, 2018).

Bailey (2012) were among the first to begin looking at the problem of noisy and/or missing data through a generalization of the traditional PCA algorithm that results in faster

[4]commonly employed when there is a need to access components or factors which explain the most of variability in the data

Modality	DNN	Accuracy	k-PCA	Var
Audio	LSTM	55.2	10	0.6104
	BLSTM	55.1	10	0.6104
	CNN	**57.2**	**20**	**0.8267**
Visual	LSTM	56.7	25	0.9436
	BLSTM	56.5	20	0.8995
	CNN	**57.1**	**25**	**0.9436**
Text	**LSTM**	**71.7**	**110**	**0.9836**
	BLSTM	70.8	110	0.9836
	CNN	68.5	130	0.9907

Table 2. Unimodal accuracy on test set for best-performing k value for PCA and showing the corresponding variance threshold. Top unimodal system highlighted in bold. Parameters refer to DNN layers, number of nodes in fully-connected layer, dropout rate, segment maxlen.

DNN	Mode	Test Accuracy		Best Params
		-PCA	+PCA	
LSTM	A,V,T	**70.5**	70.1	1,100,0.2,25
	A,T	69.2	**70.8**	2,100,0.2,30
	A,V	55.1	**55.8**	3,100,0.1,20
	V,T	**72.3**	69.5	2,100,0.2,30
BLSTM	A,V,T	71.4	**71.8**	3,100,0.2,25
	A,T	**71.2**	**71.2**	1,100,0.2,25
	A,V	55.1	**56.7**	3,100,0.1,30
	V,T	**72.4**	69.3	2,128,0.2,30
CNN	A,V,T	**69.2**	68.5	1,100,0.2,20
	A,T	**68.3**	**68.3**	1,100,0.1,30
	A,V	55.6	**57.4**	2,100,0.1,30
	V,T	**69.3**	68.8	3,100,0.2,30

Table 3. Test accuracy results with and without PCA for early fusion experiments using combinations of 2 or 3 modalities (A=audio, V=video, T=text). Top multimodal system for +/- PCA is highlighted in bold. Parameters refer to DNN layers, number of nodes in fully-connected layer, dropout rate, segment maxlen.

run times on data that contains a large amount of observations. They developed the Expectation-Maximization-PCA (EMPCA). Following their initial work, Delchambre (2014) improved on the EMPCA algorithm with the idea of focusing on the maximization of the weighted variance explained by each principal component through the diagonalization of the associated weighted covariance matrix.

We tried an implementation of the EMPCA algorithm[5] for the audio modality, but the need for fine-tuning proved to be essential, and is out of scope for our current work. Our initial result with EMPCA revealed that a variance of approximately 65% is more suitable for a competitive result in the case of the audio data. Improvements are necessary and needed to this implementation as a variance of 82% yields a set that is easier to predict, as shown in Table 2.

4. Experiment Results

In this section we talk about the experiments. We experimented with the 3 fusion techniques with and without PCA, for predicting the positive/negative sentiment of the videos. Thus we report accuracy for the binary sentiment classification problem.

4.1. Input-Level Feature Fusion

We explored input feature fusion with and without PCA. When we ran early fusion with PCA, we used the k-PCA components value found for a modality/architecture combination. For example, in our first experiment shown in Table 3 for LSTM and trimodal fusion, we used the PCA component values $k = [10, 25, 110]$ for the corresponding modality $m = [A, V, T]$ as reported earlier.

Our experiment results for early fusion are displayed in Table 3. The table allows us to compare results between systems with and without PCA, as well as the bimodal ablation groups, across DNN systems. We provide the parameters for the best-performing system configuration. Reported best parameters are based on highest accuracy score and ordered as: DNN layers, number of nodes in fully-connected layer, dropout rate, video segment maxlen. In the case of a tie in performance, the parameters are shown for the system without PCA.

The gains from PCA for input-level fusion are somewhat small, especially considering that only two of our input-level fusion techniques begin to approach state-of-the-art performance reported by Zadeh et al. (2016b). We noticed that the best-performing systems that incorporated PCA used a longer maxlen context (e.g. *maxlen* = 30) compared to the corresponding system without PCA (e.g. *maxlen* = 15). That might be due to the fact that dimensionality reduction resulted in less noise which made the context more relevant for sequence learning, though further studies with EMPCA are needed in order to confirm this.

4.2. Intermediate-Level Feature Fusion

The intermediate features fusion model that we propose adds dense layers on top of the intermediate weights extracted from each modality. There are other possible configurations to be explored, but we experimented with the simplest one. Compared to early fusion, the features for each modality are first fed to a different network. We have chosen the best performing network for each modality (CNN for audio and video, BLSTM for text) for the pre-fusion stages, but again there are other configurations to be experimented with. This configuration makes it possible to make a direct comparison with our other approaches. Results are summarized in Table 4. We achieve our highest performance so far. We find that this approach benefits from having all 3 modalities, compared to the early fusion approach, where bimodal models achieved the highest results.

[5] https://github.com/sbailey/empca/

Mode	Test Accuracy		N_layers, Dropout, Maxlen
	-PCA	+PCA	
A,V,T	**73.4**	71.2	1, 0.1, 25
A,T	69.4	69.9	3, 0.2, 30
V,T	69.7	68.1	2, 0.2, 20

Table 4. Test accuracy results with and without PCA for intermediate feature fusion experiments using combinations of 2 or 3 modalities (A=audio, V=video, T=text).

4.3. Decision-Level Fusion (Late Fusion)

For our fate fusion experiments, again we studied both trimodal and bimodal interactions. We experiment with the same parameters as in methodology. We keep the pre-fusion network consistent (CNN for audio and video, BLSTM for text)). Results are in Table 5. Our best result is for the trimodal inputs. We find that the results are not much different from a carefully trained text only predictor. Since the video and audio classifiers are much worse predictors than text, we consider that a decision level classifier is not the best approach for this dataset.

Mode	Test Accuracy		N_layers, Dropout, Maxlen
	-PCA	+PCA	
A,V,T	**71.7**	69.6	3, 0.2, 20
A,T	69.3	**69.8**	3, 0.2, 30
V,T	70.6	68.6	3, 0.1, 30

Table 5. Test accuracy results with and without PCA for decision level fusion experiments using combinations of 2 or 3 modalities (A=audio, V=video, T=text).

5. Analysis of experiments

In our previous work G25 (2018), we have experimented with unimodal classifiers. We have improved our best unimodal performance by adding PCA, as compared to the previous coursework. We found that text is the best predictor for this task.

We proposed fusing all 3 modalities in order to take advantage of additional information and the interaction between features. In Table 6, we compare unimodal models with our new findings. Note that for each model we have chosen the best validation score over the hyper-parameters noted in Methodology. As noted, intermediate level and decision level fusion techniques use the best performing Text, Audio and Video configurations as pre-fusion processing. Our initial unimodal experiments were therefore useful for this phase. These results show that fusion techniques are able to take advantage of the additional information available from fusing the modalities.

We present a sample of both negative and positive sentences and the scores of our best performance classifier. A score above 0.5 classifies the sentences as positive. The examples outline the difficulty of the task and it is clear that some sentences are difficult to label, even for human annotators.

Mode	Accuracy
Text	71.7
Audio	57.2
Visual	57.1
Input-Level Fusion V,T	72.4
Intermediate-Level Fusion A,V,T	**73.4**
Decision-Level Fusion A, V,T	71.7
Most Frequent Class (0,1)	41

Table 6. Summary of best unimodal and fusion results (A=audio, V=video, T=text).

Sentence text	True Label	Classif
1. Other than that it was a good movie	+	0.85
2. It was cute you know the actors did a great job bringing the smurfs to life such as Joe	+	0.95
3. The voice acting was phenomenal	+	0.94
4. It was like this like pouty like grumpy look	-	0.31
5. I didn't really care about it at all	-	0.23
6. Or just really doesn't make any difference to us today at all	-	0.27
7. This looks like it just has a polyurethane coating on it	-	0.56
8. Now the real Steven Russel has like an IQ like 163 which is like wow genius.	+	0.49
9. If you know they're in there this is a cheesy um movie.	-	0.80

Table 7. Example text input and their labels. Wrong classification is distinguished in red.

For input-level feature fusion, we showed that our best performance was obtained using bimodal video+text data without PCA and using a 2-layer bidirectional Long Short-Term Memory (BLSTM) DNN architecture with 72.4% accuracy. This is already over the best performance of the text only unimodal classifier. (71.7 %).

The decision level classifier does not performs so well on this dataset, as the audio and video predictors are too weak to be useful in our ensemble. Therefore, its accuracy is similar to the text only approach.

Finally, our overall best performing architecture was intermediate-level fusion of textual and visual data. This was somewhat expected, as it improves on the early concatenation approach by extracting intermediary fusion and it is not affected by the noise in the lower performing classifier as the decision level fusion.

6. Conclusions and Future Work

Discussion

Despite our efforts to reduce feature redundancy during

early fusion, we found an apparent ceiling in terms of the best overall accuracy, as it never reached above 73.4%.

We were able to show that PCA improves test accuracy for most unimodal experiments, and that gave a significant performance improvement from our previous coursework (G25, 2018). We were also able to show that PCA sometimes improved early fusion performance. Interestingly, during our bimodal ablation study, we found that leaving out audio, to focus only on video+text features, sometimes improved performance. This finding correlates with our previous coursework and also the state of the art Zadeh et al. (2017) on the MOSI dataset, that audio is the worst performing of all three modalities for this dataset.

Something to consider when interpreting our results is that we treated the task as a binary sentiment classification problem which implies that we did not have a category for "neutral". It could be the case that some of our data exemplars were a better fit for this third category, or that audio features correspond better to a neutral category. This is an important point to be explored in future work.

We show that both late decision-level fusion and early fusion can achieve comparable results. However, while decision fusion has the advantage of a smaller feature space, because we do not concatenate all the initial features, it is also harder to train, as it involves 3 different classifiers.

As the goal of the project was to explore multimodal fusion techniques, we managed to explore 3 interesting such architectures that all yield better results that unimodal classifiers. We conclude that there are interactions to be learned during the fusion process.

Futher work

The presence of 3 modalities leaves a large number of experiments that could be run, as we could use a different approach for each of the fused modalities. We have not experimented with a much deeper network, due to our desire to keep the architectures consistent across techniques.

The data set is relatively small which is problematic when fitting Deep Neural Networks to it. Deep Neural Networks are known to over-fit easily and with small data sets it easier for models with such low bias to have poor generalization.

In future work it would be interesting and helpful to examine which low-level descriptors, facial features, and words had been identified by our application of PCA. That type of information would be helpful for anyone who operates on the raw data, whereas for this work we had relied on the standardized data that was extracted and pre-processed by the CMU-MultimodalSDK developers. Still, we recognize the value in using a standardized dataset because it allows for detailed and meaningful comparison of systems.

Finally, an interesting extension is proposed by Poria et al. (2017), where contextual sentiment analysis is suggested. The dataset that we have worked with breaks down each movie review into sentences to be classified individually, thus losing context that could be gained by looking at the other neighboring sentences. An interesting extension would be to take advantage of this additional information when predicting the sentiment of a sentence. This would mean that instead of considering each utterance a separate entity, we would add context from the rest of the sentences that are part from the same speech. In this case, the format of the data would hint to an LSTM approach.

References

Abdi, Hervé and Williams, Lynne J. Principal component analysis, 2010.

Bailey, Stephen. Principal component analysis with noisy and/or missing data. *Publications of the Astronomical Society of the Pacific*, 124(919):1015, 2012.

Baltrušaitis, Tadas, Robinson, Peter, and Morency, Louis-Philippe. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

Cambria, Erik, Hazarika, Devamanyu, Poria, Soujanya, Hussain, Amir, and Subramaanyam, R. B. V. Benchmarking multimodal sentiment analysis. *CoRR*, abs/1707.09538, 2017.

Chen, L. S., Huang, T. S., Miyasato, T., and Nakatsu, R. Multimodal human emotion/expression recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 366–371, Apr 1998. doi: 10.1109/AFGR.1998.670976.

Chen, Minghai, Wang, Sen, Liang, Paul Pu, Baltrušaitis, Tadas, Zadeh, Amir, and Morency, Louis-Philippe. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 163–171. ACM, 2017.

Chernykh, Vladimir, Sterling, Grigoriy, and Prihodko, Pavel. Emotion recognition from speech with recurrent neural networks. *arXiv preprint arXiv:1701.08071*, 2017.

Delchambre, Ludovic. Weighted principal component analysis: a weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society*, 446(4):3545–3555, 2014.

Feng, Shi, Wang, Daling, Yu, Ge, Gao, Wei, and Wong, Kam-Fai. Extracting common emotions from blogs based on fine-grained sentiment clustering. *Knowledge and information systems*, 27(2):281–302, 2011.

G25. MLP Coursework 3. 2018. URL /afs/inf.ed.ac.uk/group/teaching/mlp/2017-18/s1427590/cw3/coursework3/interimReport-G25.pdf.

Ghosh, Sayan, Laksana, Eugene, Morency, Louis-Philippe, and Scherer, Stefan. Representation learning for speech emotion recognition. In *INTERSPEECH*, pp. 3603–3607, 2016.

Gunes, H. and Piccardi, M. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pp. 3437–3443 Vol. 4, Oct 2005. doi: 10.1109/ICSMC.2005.1571679.

Hu, Anthony and Flaxman, Seth. Multimodal sentiment analysis to explore the structure of emotions, 2018.

Ioffe, Sergey and Szegedy, Christian. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

K, Padmavathi, Bhat, Mahima, and Karki, Maya V. Feature selection based on pca and pso for multimodal medical image fusion using dtcwt. *CoRR*, abs/1701.08918, 2017.

Kim, Yoon. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

Kingma, Diederik P. and Ba, Jimmy. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Kwak, Nojun. Feature extraction for classification problems and its application to face recognition. *Pattern Recogn.*, 41(5):1701–1717, May 2008. ISSN 0031-3203.

Lee, Jinkyu and Tashev, Ivan. High-level feature representation using recurrent neural network for speech emotion recognition. 2015.

Ma, Lin and Li, Naimin. Texture feature extraction and classification for iris diagnosis. In *Proceedings of the 1st International Conference on Medical Biometrics*, ICMB'08, pp. 168–175, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-77410-6, 978-3-540-77410-5.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

Morency, Louis-Philippe, Mihalcea, Rada, and Doshi, Payal. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pp. 169–176, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0641-6.

Nair, Vinod and Hinton, Geoffrey E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, USA, 2010.

Niu, Yafeng, Zou, Dongsheng, Niu, Yadong, He, Zhongshi, and Tan, Hua. A breakthrough in speech emotion recognition using deep retinal convolution neural networks. *arXiv preprint arXiv:1707.09917*, 2017.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Poria, S., Majumder, N., Hazarika, D., Cambria, E., Hussain, A., and Gelbukh, A. Multimodal Sentiment Analysis: Addressing Key Issues and Setting up Baselines. *ArXiv e-prints*, March 2018.

Poria, Soujanya, Cambria, Erik, and Gelbukh, Alexander. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2539–2544, 2015a.

Poria, Soujanya, Cambria, Erik, and Gelbukh, Alexander F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, 2015b.

Poria, Soujanya, Cambria, Erik, and Gelbukh, Alexander F. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*, 2015c.

Poria, Soujanya, Cambria, Erik, Hazarika, Devamanyu, Mazumder, Navonil, Zadeh, Amir, and Morency, Louis-Philippe. Context-dependent sentiment analysis in user-generated videos. In *Association for Computational Linguistics*, 2017.

Pérez-Rosas, Verónica, Mihalcea, Rada, and Morency, Louis-Philippe. Utterance-level multimodal sentiment analysis, 08 2013.

Rozgic, Viktor, Ananthakrishnan, Sankaranarayanan, Saleem, Shirin, Kumar, Rohit, and Prasad, Rohit. Ensemble of svm trees for multimodal emotion recognition. *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, 2012.

Schuller, B. Recognizing affect from linguistic information in 3d continuous space. *IEEE Transactions on Affective Computing*, 2:192–205, 07 2011. ISSN 1949-3045.

Srivastava, Nitish, Hinton, Geoffrey, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Sun, Zehang, Bebis, George, and Miller, Ronald. Object detection using feature subset selection. *Pattern Recogn.*, 37(11):2165–2176, November 2004. ISSN 0031-3203.

Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., and Baik, S. W. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6: 1155–1166, 2018.

Wöllmer, Martin, Kaiser, Moritz, Eyben, Florian, Schuller, Björn, and Rigoll, Gerhard. Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2):153–163, 2013.

Wollmer, Martin, Weninger, Felix, Knaup, Tobias, Schuller, Bjorn, Sagae, Kenji, and Morency, Louis-Philippe. YouTube Movie Reviews: In, Cross, and Open-domain Sentiment Analysis in an Audiovisual Context. *IEEE Intelligent Systems*, 28(3), March 2013.

Wu, Lizhong, Oviatt, S. L., and Cohen, P. R. Multimodal integration-a statistical view. *IEEE Transactions on Multimedia*, 1(4):334–341, Dec 1999. ISSN 1520-9210. doi: 10.1109/6046.807953.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., and Morency, L. P. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, May 2013. ISSN 1541-1672.

Yin, Wenpeng, Kann, Katharina, Yu, Mo, and Schütze, Hinrich. Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923, 2017.

Zadeh, A, Liang, PP, Poria, S, Vij, P, Cambria, E, and Morency, LP. Multi-attention recurrent network for human communication comprehension. In *AAAI*, 2018.

Zadeh, Amir, Zellers, Rowan, Pincus, Eli, and Morency, Louis-Philippe. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016a.

Zadeh, Amir, Zellers, Rowan, Pincus, Eli, and Morency, Louis-Philippe. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016b.

Zadeh, Amir, Chen, Minghai, Poria, Soujanya, Cambria, Erik, and Morency, Louis-Philippe. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250, 2017. URL http://arxiv.org/abs/1707.07250.

Zeng, Zhihong, Pantic, Maja, Roisman, Glenn I, and Huang, Thomas S. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE transactions on pattern analysis and machine intelligence*, 31 (1):39–58, 2009.