Deep Learning Multimodal Fusion for Sentiment Analysis

Group 25
s1738075, s1427590, s1211898

Abstract

We explore architectures for deep multimodal fusion for the binary classification task of sentiment polarity prediction. In this work, we focus on building separate stand-alone unimodal systems to gain a better understanding of how text, acoustic and visual features each perform on this task. We used the open source MOSI dataset to run multiple experiments with several different deep learning architectures. We found that overall, the text data is very predictive of sentiment and it achieves the highest unimodal accuracy of 70% whereas acoustic data alone consistently performs only slightly better than chance. We then examined two baseline techniques for modality fusion. We compared both of our preliminary fusion techniques with the current state of the art technique, tensor fusion. We show that our method of multimodal ensemble fusion outperforms the state of the art on this task. Finally, for future work we propose a transfer learning approach for taking advantage of previously learned weights. We also propose expanding our work on polarity sentiment classification to include emotion categories such as: *happy*,*sad*,*anger*, *fear*, *disgust*, and *surprise* using a newly released MOSEI dataset from the same open source providers.

1. Motivation

In light of recent successes with deep learning approaches to multimodal classification problems, problems such as sentiment analysis remain truly challenging. Both emotion and sentiment analysis have become increasingly important in recent years, helping companies automatically extract and understand their users' opinions, and helping governments understand political climate within particular geographical regions. However, it remains a difficult task due to the ambiguity of language and the use of slang and sarcasm. Most sentiment analysis tasks are performed on text data, exploiting the sequential structure of texts and using Long Short Term Memory Networks on text. However, rich complementary information can be extracted from audio and visual data. A sentence like "This movie is sick" together with a smile as a visual cue can make the classification task much easier, since it can bring different semantic information. (Peng et al., 2016) Continuous word representations, while popular in recent years, fail to capture ambiguities in language. We aim to show that adding acoustic and visual data can help models make better sentiment predictions.

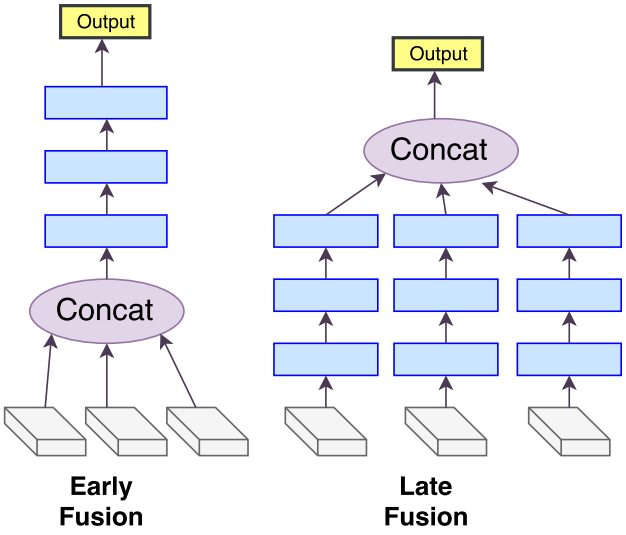

Figure 1. Multimodal fusion techniques

Multimodal machine learning has been recently attracting interest, with the abundance of multimedia data available on the internet, making it easy for researchers to integrate data of multiple modalities. It is a vibrant research field which aims to integrate and model multiple sources of input, usually acoustic, visual and text. In order for Artificial intelligence to make progress, it must be able to integrate multimodal signals together. While it it applicable across many fields and tasks, we look at sentimental analysis, where approaches with interesting results have been proposed. (Pérez-Rosas et al. (2013), Wöllmer et al. (2013), Poria et al. (2015))

A survey by Baltrusaitis et al. (2017) motivates some of the uses of multimodal data, together with five main components:

Representation Representing and summarizing multimodal data

Translation Mapping data from one modality to another

Alignment Identifying relationships between modalities: for example, transcribed text of a video

Fusion Joining information for different modalities in order to perform a prediction

Co-learning Transfering knowledge between modalities

This report will touch on representation and alignment issues, but it will mostly focus on fusion and co-learning techniques. There are two major approaches to multimodal fusion: **early fusion** and **late fusion**. The first refers to fusing information at the feature level, and it is the most widely used. The second is also known as *ensemble fusion*, which fuses multiple modalities at the decision level. In the machine learning community, deep learning approaches have been proposed in an effort to exploit new architectures.

This paper is organized as follows: in Section 2, we describe our research questions and the objectives we hope this work will achieve. In Section 3 we provide an overview of the data we used in our experiments, as well as the experiment task. In Section 4, we describe our methodological approach. In Section 5, we provide the details of our unimodal baseline experiments. In Section 6, we provide preliminary results for our modal fusion experiments. Section 7 discusses our interim conclusions based on our experiments. Finally, Section 8 describes our plans for future work.

2. Research Questions and Objectives

We aim to research how the 3 modalities present in our dataset interact. We start by looking at architectures for unimodal features on their own. Previous work [CITATION] demonstrates the advantages of using multiple modalities. However, fusing these modalities is non trivial.

- What are the significant and least significant features (and modalities) for predicting sentiment?
- Which multimodal fusion architectures perform the best?
- Can we achieve an improvement from modality fusion over using only our best performing modality?

Our final goal is to understand how to take advantage of the richness of the data and combine the 3 modalities for better sentiment classification.

3. Data Set and Task Description

In this section we describe our data in more details. To describe the data, we consider 3 different modalities separately, as each modality consists of different types of features. We also provide a focused explanation of the task.

3.1. Data

We use the Multimodal Opinionlevel Sentiment Intensity (MOSI) dataset, 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, visual gestures and acoustic features. The dataset can be downloaded using the CMU-MultimodalSDK (Zadeh et al., 2018), which also provides preprocessed features and a way to align text, acoustic and visual data.

Text features: word vectors from Global Vectors for Word Representation(GloVe) (Pennington et al., 2014) as well as 1-hot word representations. The duration of each word utterance is also provided by the P2FA forced alignment

Speech features: From the audio modality, the software COVAREP was used to extract acoustic features including 12 Mel-frequency cepstral coefficients, pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients are extracted. The sampling rate of these features is 100hz from the original audio (Degottex et al., 2014)

Video features: Facial action units which are the indicators for facial muscle movements as well as 6 basic emotions regressed from just facial features. The position of facial landmarks is also present in the vision features. These features are sampled at 30hz .

3.2. Task

We look at the task of sentiment binary classification. Therefore, we take a label> 0 as being in the positive class, while labels < 0 are in the negative class. We measure performance with overall accuracy on this binary classification problem.

4. Methodology

Existing work

Cambria et al. (2017) published sentiment analysis benchmarks for many multimodal sentiment datasets, including MOSI. Their state of the art is 76.6%. However, they preprocess their features differently, by training their own word embeddings and using different softwares for speech and visual feature extraction. We chose to use the available features provied by the CMU-MultimodalSDK for the reason that it provides a unified format for many datasets that we potentially want to try in the future. This means that we cannot directly compare our results with their state of the art. We followed the standard of running unimodal experiments first and comparing them with multimodal fusion techniques.

We split our data into training (1283 items), validation (229 items) and test (686 items), which is the split provided by the dataset. The positive class makes up 52% of the training set. Since sequences have different lengths, we need to pad or crop our features to a predefined length. We do so by examining the length frequencies. For text, we use a length of 15, while for video and speech features we take length 20. This choice is based on validation accuracy. To reduce overfitting, we perform early stopping monitoring validation loss on the past 10 epochs and saving the best weights. Each architecture is explained in more detail as it is introduced.

The activation function we use across experiments is ReLu

(Nair & Hinton, 2010). The learning rule is Adam (Kingma & Ba, 2014) with standard parameters. For 1D convolution layers, the kernel size is 3. For max pooling layers, the window size is 2. We vary the number of convolutional layers in $[1, 2, 3]$. For BidirectionalLSTMs, we vary the number of units in $[64, 128, 256]$ and the number of layers in $[1, 2, 3]$. For fully connected layers, we vary the number of units in [100, 200] and the number of layers in $[1, 2, 3, 4]$. We add dropout(Srivastava et al., 2014) between fully connected layers with dropout rate in $[0.1, 0.2, 0.35]$ 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 split our experiments into unimodal and multimodal. In Section 5, we start by analyzing the performance of each of the modalities: we call these unimodal classifiers, operating on either text, visual or acoustic features. In Section 6, we experiment with early and late fusion fusion techniques for exploiting all 3 types of data, in order to model interactions between modalities (known as intermodal dynamics)

4.1. Feature alignment

For our early fusion experiments, we align the modalities, because different features in multimodal datasets are in different temporal frequencies. The CMU-MultimodalSDK alignes data using weighted averaging. The overlap of each modality with a reference one is the weight of each modality. An average is taken with these weights to align them to the reference.

4.2. MC Dropout

The data set is relatively small which is problematic when fitting Deep Neural Networks to it. Deep Neural Networks are known to overfit easily and with small data sets it easier for models with such low bias to have poor generalisation. It is well known that Bayesian methods perform well with small data sets and offer robustness to overfitting (MacKay, 1992). When applied to deep learning Bayesian Inference poses tractability challenges and has had very few success stories when approximated with variational methods.

We decided to use a cheap and empirically successfully approximation to Bayesian deep learning known as MC Dropout (Gal & Ghahramani, 2016).

The dropout objective in DNNs can be shown to minimise the KL divergence between a selected candidate distribution over weight matrices $Q(\omega)$ and the posterior of a Deep Gaussian Process. This results highlight that a multilayer neural network $\hat{y}(\omega, x)$ with dropout applied before each weight layer is mathematically equivalent to an approximate Deep GP posterior. Having tractably approximated the posterior we can write down the approximate posterior distribution as:

$$Q(y^*|x^*) = \int p(y^*|x^*, \omega) Q(\omega) d\omega \tag{1}$$

$$Q(y^*|x^*) = \int \mathcal{N}\left(y^* \middle| \hat{y}(\omega, x^*), \tau^{-1}\mathbb{I}\right) Q(\omega) d\omega \tag{2}$$

The mean of the posterior predictive distribution can then cheaply be approximated by drawing Montecarlo samples from the dropout distribution over weight matrices $Q(\omega)$:

$$\mathbb{E}_{Q(y^*|x^*)}(y^*) \approx \frac{1}{T} \sum_{s=1}^{T} \hat{y}(\omega_s, x^*), \quad \omega_s \sim Q(\omega) \tag{3}$$

This is known as MC dropout and just equates to doing T stochastic forward passes through the dropout network and averaging them. This can also be interpreted as using the MC dropout samples to achieve an mean ensemble effect of different neural network topologies. Applying this to our best performing network over text features achieves an increase in accuracy from 69.5% to 70%.

5. Unimodal Approaches

In this section we describe the approach to each modality on its own. We also explain the features in more depth. It is important to understand which approaches work best for each modality. We used this information later when deciding how to approach multimodal fusion.

5.1. Text

To extract textual features, we represent the transcribed words as bag of words with a (filtered) vocabulary size of 2713.

GloVe Embeddings GloVe allows obtaining vector space representation for words. We use pretrained vectors on Common Crawl data with 300 dimensions. The advantages of using continuous vector representation as opposed to simple bag of words models have been outlined many times in literature, as words end up nearby to semantically similar words, in terms of vector distance. GloVe uses a count based approach, computing co-occurrence statistics in a large corpus, as opposed to predictive based models such as word2vec (Mikolov et al., 2013).

Experiments

Common network architectures for NLP tasks use either CNN or RNN. A an extensive published by Yin et al. (2017) notices comparable results for both. We therefore experiment with these two approaches on our unimodal textual features.

CNN Convolutional Neural Networks are often used in NLP in various prediction tasks, including sentiment analysis. (Kim, 2014) The interpretation is not as straightforward as for images, but we can still argue that semantically related vectors will be close to each other within a context

window. As outlined in the methodology, we use 1D Convolutional layers.

LSTM Recurrent Neural Networks(RNNS) are variants have been proven very successful for many tasks including sentiment analysis on text and are known for their ability to model invariances across time. Recent advancements propose variants of RNNs that do not suffer from the problem of vanishing gradients: Long Short Term Memory(LSTM). The goal of LSTMs is to capture long term dependencies, such as context words.

Bidirectional LSTM Bidirectional LSTMs increase the amount of available contextual information. The principle is to use both a forward pass and a backward pass through the sequence of words.

We chose our best performing model on the validation set, which was Bidirectional LSTM. CNN achieves 73.8%, LSTM 74.24% and BiLSTM achieves 75.1% accuracy.

5.2. Video

FACET features

According to Zadeh et al. (2016), the visual features include 16 Facial Action Units, 68 Facial Landmarks, Head Pose and Orientation, 6 Basic Emotions and Eye Gaze(Wood et al. (2015), Baltrusaitis et al. (2014)). These emotions were extracted using FACET's Emotient. FACET is a state-of-the-art facial expression recognition and analysis software that was preceded by a research version known as the Computer Emotion Recognition Toolbox, CERT(Littlewort et al. (2011)). FACET provides frame-by-frame tracking of facial action units according to the Facial Action Coding Scheme. These action units include such expressions as AU4 Brow Lowerer, AU15 Lip Corner Depressor, and AU23 Lip Tightener (see Figure 2 for illustration). The FACET software provides an Evidence measure for each facial action unit, indicating the chance that the target expression is present.

Figure 2. Main Action Units(AUs) from the Facial Action Coding System as depicted in la Torre et al. (2015)

CNN Convolutional Neural Networks deeply influenced the evolution of the field of face recognition thanks to their feature learning and transformation invariance properties. From the first time employing CNNs for face recognition(Lawrence et al. (1997)) to the present times when sentiment analysis revolves around the usage of CNNs(Tripathi et al. (2017), Xu et al. (2014), Pereira et al. (2016)), one can easily notice the realm of possibilities the CNNs bring with them in performing multimodal sentiment analysis. Moreover, 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. CNNs performance on visual data ranges in literature and we tried to obtain a high accuracy. For the model that we currently have, we managed to obtain a 56.4% accuracy.

LSTM

5.3. Acoustic

Previous work has shown that there are particular elements of the speech signal which are most indicative of emotional state of the speaker (Chang et al. (2011); Zeng et al. (2009)). The features of speech which are very predictive of speaker affect are called low-level descriptors. These descriptors are extracted directly from the speech signal waveform. They include prosody, pitch (fundamental frequency), voice quality, harmonics, frame-level energy, mel-frequency cepstral coefficients (MFCCs), as well as the delta or first derivative for each. Once these low-level descriptors have been extracted from the audio signal with the COVAREP software, they are arranged into feature vectors for our machine learning experiments. In our experiments, the acoustic feature vectors consist of sequences of 36-dimensional COVAREP features for each video segment. In the MOSI dataset as given, some segments are longer or shorter than others. To achieve uniformity of the feature vectors, they are either zero-padded or truncated with a maximum-length value threshold.

SVM Early work on audio emotion detection used Support Vector Machines (SVM) with accuracy ranging between 75-80%, all of which had been measured on older datasets that we did not use (Eyben et al. (2009); Chang et al. (2011); Rahman & Busso (2012)). Recently one group experimented with the MOSI dataset and they reported the best unimodal SVM accuracy on acoustic MOSI data as 58.5%. However, they do not report any particular training parameters for the SVM, such as kernel type (Poria et al., 2017).

In our baseline SVM experiments, we tried 3 different kernels using default parameters, from the sklearn SVM toolkit: polyfit, radial basis, and linear (Pedregosa et al., 2011). Unlike previous work, which used zero-padded features for segments, we calculated the feature mean over all sequences for a given segment. We then range-normalized the mean by dividing each feature mean by the feature maximum. We achieved poor performance for each kernel type and default parameters, 40.3% accuracy.

BLSTM Speech data is often considered sequentially informative. For example, the rise and fall of prosody can form meaningful patterns. Many approaches to detect-

ing emotion in speech use sequential learning methods such as LSTM (Lim et al., 2016). More specifically, there has been previous work using BLSTM, which we have also attempted to replicate in this work for our baseline Ghosh et al. (2016); Lee & Tashev (2015); Han et al. (2014); Chernykh et al. (2017). We specifically attempted to replicate the BLSTM baseline presented in (Lee & Tashev, 2015), which achieved an unweighted accuracy above 60% for the IEMOCAP dataset and 32 acoustic descriptors. Their work describes a bi-directional LSTM with 2 hidden layers, and 64 forward and 64 backward nodes. We attempted to replicate this BLSTM architecture. However, our best results achieved only 52.9% unweighted accuracy on the MOSI dataset.

CNN 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 spefically by working directly on the spectrogram (Niu et al., 2017). The spectrogram contains the same information as a waveform, however it can be treated as an image with multiple color channels. While our work is not directly comparable to the above, due to different datasets and features, we performed our own CNN baseline experiments using our methodology and achieved an accuracy of %.

FCN We replicated the fully-connected network (FCN) as described in the work of (Zadeh et al., 2017). They used an FCN on the audio modality, and reported unimodal performance, before fusing audio with the other modalities in their tensor approach. Their paper reports that the acoustic FCN achieved 65.1% accuracy, however upon replication we only achieved 53.06% accuracy. We think this could be due to our methodology of using the maximum context length of 20. It is uncertain what type of parameters the authors used when setting their FCN baseline.

6. Multimodal Approaches

From our unimodal experiments, we conclude that textual features have the most predictive power. Images, text and acoustic can be complementary to each other, because they contain different semantic information. We will explore ways to take advantage of the other modalities, which are not so powerful on their own. The standard fusion approaches are early fusion and late fusion (Figure 1).

6.1. Early Fusion

In the early fusion approach, features from the 3 modalities are simply concatenated and become the input vector. Since sequences have different lengths, all modalities are processed to keep only the first 20 features, in order for the concatenation to be possible. We chose this value by monitoring validation loss. The concatenated features are fed into an LSTM with 256 units.

6.2. Late Fusion

Late fusion techniques fuse modalities at decision level. We compare our approach (trimodal ensemble fusion) with the existing state of the art(tensor fusion).

Trimodal Ensemble Fusion

Our proposed architecture (Figure 3) consists of 3 subnetworks, Acoustic CNN, Visual CNN and Text LSTM. It is a late fusion technique in the sense that the modalities are fused at the decision level. The difference from existing literature is that each subnetwork has a sigmoid output and is pretrained to make its own prediction. Our final layer takes the output from the 3 sigmoid units (one for each modality) and outputs the final prediction, learning to weight the predictions of the already pretrained network in an ensemble like fashion. Most multimodal ensembles set the weights manually [CITATION]

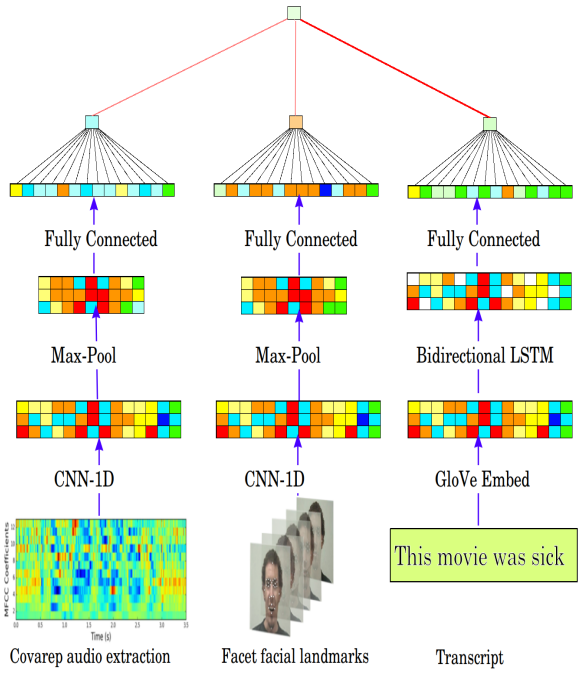

Figure 3. Our proposed Trimodal ensemble fusion network architecture

Figure 4 compares validation accuracy on our proposed Ensemble Fusion Network. The way in which we built the architecture allows us to also visualize the pretrained networks separately (unimodal features)

Tensor fusion

The state of the art multimodal approach is introduced by Zadeh et al. (2017).

They propose a late fusion approach. The text modality subnetwork uses an LSTM. For speech and visual data, they chose to take the mean across timesteps and use fully connected layers. We argue that this approach discards sequential information. Each subnetwork outputs a vector. The proposed approach takes the tensor product of these

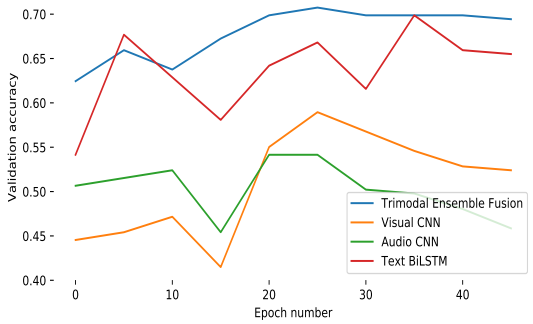

Figure 4. Experimental results on MOSI: Trimodal Ensemble Fusion and its unimodal subnetworks

vectors:

$$\left(T_{fusion}\right)_{ijk} = (T_{acoustic} \otimes T_{visual} \otimes T_{text})_{ijk} = \\ = (T_{acoustic})_i (T_{visual})_j (T_{text})_k$$

The dataset has been updated since the paper was published. In order to have a fair comparison with our methods, we have reproduced the results by modifying the code released by the author together with their paper.[1] This might be because the stratification of the classes in the test set might have changed. Their code sets no random seeds, which makes it harder to reproduce.

7. Interim Conclusions

We have explored unimodal architectures for text, speech and visual features for the task of sentiment polarity prediction. We find that text is the most significant feature for this task, which can be explain by the possibility to learn when particular words carry positive or negative connotations (*love*, *hate*), whereas it is more difficult to use acoustic features such as pitch or visual cues such as facial landmarks. We find that for unimodal approaches, both CNNs and LSTMs (or Bidirectional LSTMs) perform well, which aligns to results from litereature. We have 3 multimodal architectures in place (early fusion, tensor fusion, ensemble fusion), the later being our proposed architecture. We have explored interesting fusion techniques to take advantage of the richness of the data. Results on the test set are summarized in Table 1.

8. Project Plans

Based on the initial experimentation in this report, we are interested in exploring multimodal fusion and ensembling approaches further. Our goal is to improve our work on fusing the modalities together for our last research question.

[1] https://github.com/A2Zadeh/TensorFusionNetwork

Model	Test acc
Predict one class	41%
Text BiLSTM	69.5%
Text BiLSTM MC Dropout	70%
Acoustic CNN	53.9%
Visual CNN	57.7%
Multimodal Early Fusion	68 %
Multimodal Tensor Fusion	69.3%
MUltimodal Ensemble Fusion	71.7%

Table 1. Test accuracy on MOSI binary sentiment classification

8.1. Improved neural ensembles

Taking inspiration from the ensemble interpretation of MC Dropout (Section 4.2) we can use this idea to combine multiple modalities similar to to the trimodal ensemble fusion, only that here we would have dropout in every layer including the weighting (combining) layer such that when we predict with multiple MC Dropout samples we will switch on and off of modalities and components of modalities in order to achieve an averaging/ensemble effect. Using dropout as a way of achieving ensemble learning has been proposed before. (Hara et al., 2017)

8.2. Transfer learning

Transfer learning refers to storing knowledge gained while solving one problem and applying it to a related problem. A straightforward way to achieve this in our case is to:

- Align all modalities (already possible through the provided timestamps in the data)
- Train a model on our best performing modality (text)
- Train the exact same architecture on another modality but place a prior on the previously learnt weights on text.

A straightforward way to place a prior on the previous learned weights for one modality (text) by adding a modified L_2 regularization penalty centered on the optimal acoustic weights to the loss of the text model:

$$\hat{\omega}_{text} = \underset{\omega}{\text{argmax}}\, \mathcal{L}_{text}(\omega)$$
$$\mathcal{L}_{transfer}(\omega) = \mathcal{L}_{acoustic}(\omega) + \lambda \|(\omega - \hat{\omega}_{text})\|^2$$

where $\mathcal{L}$ is the loss function. This idea is inspired by the work of Kirkpatrick et al. (2016) for elastic weight consolidation, as a way to "slow down learning on certain weights based on how important they are to previously seen tasks".

8.3. Other datasets

Time permitting, we would like to explore these same approaches on a dataset with the same format, but 10 times larger, where no work has been published. [CITATION MOSEI]

References

Baltrusaitis, Tadas, Robinson, Peter, and Morency, Louis-Philippe. Continuous conditional neural fields for structured regression. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 593–608, 2014.

Baltrusaitis, Tadas, Ahuja, Chaitanya, and Morency, Louis-Philippe. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017.

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.

Chang, Keng-hao, Fisher, Drew, and Canny, John. Ammon: A speech analysis library for analyzing affect, stress, and mental health on mobile phones. *Proceedings of PhoneSense*, 2011.

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

Degottex, Gilles, Kane, John, Drugman, Thomas, Raitio, Tuomo, and Scherer, Stefan. Covarep—a collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 960–964. IEEE, 2014.

Eyben, Florian, Wöllmer, Martin, and Schuller, Björn. Opensmile—introducing the munich open-source emotion and affect recognition toolkit. In *Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on*, pp. 1–6. IEEE, 2009.

Gal, Yarin and Ghahramani, Zoubin. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059, 2016.

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

Han, Kun, Yu, Dong, and Tashev, Ivan. Speech emotion recognition using deep neural network and extreme learning machine. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Hara, Kazuyuki, Saitoh, Daisuke, and Shouno, Hayaru. Analysis of dropout learning regarded as ensemble learning. *CoRR*, abs/1706.06859, 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.

Kirkpatrick, James, Pascanu, Razvan, Rabinowitz, Neil C., Veness, Joel, Desjardins, Guillaume, Rusu, Andrei A., Milan, Kieran, Quan, John, Ramalho, Tiago, Grabska-Barwinska, Agnieszka, Hassabis, Demis, Clopath, Claudia, Kumaran, Dharshan, and Hadsell, Raia. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016. URL http://arxiv.org/abs/1612.00796.

la Torre, Fernando De, Chu, Wen-Sheng, Xiong, Xuehan, Vicente, Francisco, Ding, Xiaoyu, and Cohn, Jeffrey F. Intraface. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–8, 2015.

Lawrence, Steve, Giles, C. Lee, Tsoi, Ah Chung, and Back, Andrew D. Face recognition: a convolutional neural-network approach. *IEEE Trans. Neural Networks*, 8(1): 98–113, 1997. URL http://dblp.uni-trier.de/db/journals/tnn/tnn8.html#LawrenceGTB97.

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

Lim, Wootaek, Jang, Daeyoung, and Lee, Taejin. Speech emotion recognition using convolutional and recurrent neural networks. In *Signal and information processing association annual summit and conference (APSIPA), 2016 Asia-Pacific*, pp. 1–4. IEEE, 2016.

Littlewort, Gwen, Whitehill, Jacob, Wu, Tingfan, Fasel, Ian R., Frank, Mark G., Movellan, Javier R., and Bartlett, Marian Stewart. The computer expression recognition toolbox (CERT). In *Ninth IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011), Santa Barbara, CA, USA, 21-25 March 2011*, pp. 298–305, 2011.

MacKay, David J. C. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4(3): 448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

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

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.

Peng, Y., Zhou, X., Wang, D. Z., Patwa, I., Gong, D., and Fang, C. V. Multimodal ensemble fusion for disambiguation and retrieval. *IEEE MultiMedia*, 23(2):42–52, Apr 2016. ISSN 1070-986X.

Pennington, Jeffrey, Socher, Richard, and Manning, Christopher D. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Pereira, Moises H. R., Pádua, Flávio L. C., Pereira, Adriano C. M., Benevenuto, Fabrício, and Dalip, Daniel Hasan. Fusing audio, textual and visual features for sentiment analysis of news videos. *CoRR*, abs/1604.02612, 2016. URL http://arxiv.org/abs/1604.02612.

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*, 2015.

Poria, Soujanya, Cambria, Erik, Hazarika, Devamanyu, Majumder, Navonil, Zadeh, Amir, and Morency, Louis-Philippe. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pp. 873–883, 2017.

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

Rahman, Tauhidur and Busso, Carlos. A personalized emotion recognition system using an unsupervised feature adaptation scheme. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 5117–5120. IEEE, 2012.

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.

Tripathi, Samarth, Acharya, Shrinivas, Sharma, Ranti Dev, Mittal, Sudhanshu, and Bhattacharya, Samit. Using deep and convolutional neural networks for accurate emotion classification on DEAP dataset. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pp. 4746–4752, 2017. URL http://aaai.org/ocs/index.php/IAAI/IAAI17/paper/view/15007.

Wood, Erroll, Baltrusaitis, Tadas, Zhang, Xucong, Sugano, Yusuke, Robinson, Peter, and Bulling, Andreas. Rendering of eyes for eye-shape registration and gaze estimation. In *Proc. of the IEEE International Conference on Computer Vision (ICCV 2015)*, 2015.

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.

Xu, Can, Cetintas, Suleyman, Lee, Kuang-Chih, and Li, Li-Jia. Visual sentiment prediction with deep convolutional neural networks. *CoRR*, abs/1411.5731, 2014. URL http://arxiv.org/abs/1411.5731.

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, 2016.

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

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.