Multi-modal Sequence Fusion via Recursive Attention for Emotion Recognition

Rory Beard[1,*] **Ritwik Das**[2,*] **Raymond W. M. Ng**[1,*] **P. G. Keerthana Gopalakrishnan**[2]
Luka Eerens[2] **Pawel Swietojanski**[1] **Ondrej Miksik**[1]
[1]Emotech Labs [2]Carnegie Mellon University

Abstract

Natural human communication is nuanced and inherently multi-modal. Humans possess specialised sensoria for processing vocal, visual, and linguistic, and para-linguistic information, but form an intricately fused percept of the multi-modal data stream to provide a holistic representation. Analysis of emotional content in face-to-face communication is a cognitive task to which humans are particularly attuned, given its sociological importance, and poses a difficult challenge for machine emulation due to the subtlety and expressive variability of cross-modal cues.

Inspired by the empirical success of recent so-called *End-To-End Memory Networks* (Sukhbaatar et al., 2015), we propose an approach based on recursive multi-attention with a shared external memory updated over multiple gated iterations of analysis. We evaluate our model across several large multi-modal datasets and show that global contextualised memory with gated memory update can effectively achieve emotion recognition.

1 Introduction

Multi-modal sequential data pose interesting challenges for learning machines that seek to derive representations. This constitutes an increasingly relevant sub-field of multi-view learning (Ngiam et al., 2011; Baltrusaitis et al., 2017). Examples of such modalities include visual, audio and textual data. Uni-modal observations are typically complementary to each other and hence they can reveal a fuller and more context-rich picture with better generalisation ability when used together. Through its complementary perspective, each view can unburden sub-modules specific to another modality of some of its modelling onus, which might otherwise learn implicit hidden causes that are over-fitted to training data idiosyncrasies in order to explain the training labels.

On the other hand, multi-modal data introduces many difficulties to model designing and training due to the distinct inherent dynamics of each modality. For instance, combining modalities with different temporal resolution is an open problem. Other challenges include deciding where and how modalities are combined, leveraging the weak discriminative power of training label and the presence of variability and noise or dealing with complex situations such as modelling the emotion of sarcasm, where cues among modalities contradict.

In this paper, we address multi-modal sequence fusion for automatic emotion recognition. We believe, that a strong model should enable:
(i) *Specialisation of modality-specific sub-modules* exploiting the inherent properties of its data stream, tapping into the mode-specific dynamics and characteristic patterns.
(ii) *Weak (soft) data alignment* dividing heterogeneous sequences into segments with co-occuring events across modalities without alignment to a common time axis. This overcomes limitations of *hard alignments* which often introduce spurious modelling assumptions and data inefficiencies (*e.g.* re-sampling) which must be performed again from scratch if views are added or removed.
(iii) *Information exchange* for both view-specific information and statistical strength for learning shared representations.
(iv) *Scalability* of the approach to many modalities using (a) parallelisable computation over modalities, and (b) a parameter set size growing (at most) linearly with the number of modalities.

In the present work, we detail a *recursively attentive* modelling approach. Our model fulfills the desiderata above and performs multiple sweeps of globally-contextualised analysis so that one modality-specific representation cues the at-

[*] Equal contribution.

Proceedings of the 22nd Conference on Computational Natural Language Learning (CoNLL 2018), pages 251–259
Brussels, Belgium, October 31 - November 1, 2018.

tention of the next and vice-versa. We evaluate our approach on three large-scale multi-modal datasets to verify its suitability.

2 Related work

2.1 Multi-modal analysis

Most approaches to multi-modal analysis (Ngiam et al., 2011) focus on designing feature representations, co-learning mechanisms to transfer information between modalities, and fusion techniques to perform a prediction or classification. These models typically perform either "early" (input data are concatenated and pushed through a common model) or "late" (outputs of the last layer are combined together through linear or non-linear weighting) fusion. In contrast, our model does not fall into any of these categories directly as it is "iterative" in the sense that there are multiple fusions per decision, with an evolving belief state – the memory. In addition to that, our model is also "active" since feature extraction from one modality can influence the nature of the feature extraction from another modality in the next time step via the shared memory.

For instance, Kim et al. (2013) used low-level hand crafted features such as pitch, energy and mel-frequency filter banks (MFBs) capturing prosodic and spectral acoustic information and Facial Animation Parameters (FAP) describing the movement of face using distances between facial landmarks. In contrast, our model allows for an end-to-end training of feature representation.

Zhang et al. (2017) learnt motion cues in videos using 3D-CNNs from both spatial and temporal dimensions. They performed deep multi-modal fusion using a deep belief network that learnt non-linear relations across modalities and then used a linear SVM to classify emotions. Similarly, Vielzeuf et al. (2017) explored VGG-LSTM and 3DCNN-LSTM architectures and introduced a weighted score to prioritise the most relevant windows during learning. In our approach, exchange of information between different modalities is not limited to the last layer of the model, but due to memory component, each modality can influence every other in the following time steps.

Co-training and co-regularisation approaches of multi-view learning (Xu et al., 2013; Sindhwani and Niyogi, 2005) seek to leverage unlabelled data via a semi-supervised loss that encodes a consensus and complementarity principles. The former encodes the assertion that predictions made be each view-specific learner should largely agree, and the latter encodes the assumption that each view contains useful information that is hidden from others, until exchange of information is allowed to occur.

2.2 Memory Networks

End-To-End Memory Networks (Sukhbaatar et al., 2015) represent a fully differentiable alternative to the strong supervision-dependent *Memory Networks* (Weston, 2017). To bolster attention-based recurrent approaches to language modelling and question answering, they introduced a mechanism performing multiple hops of updates to a "memory" representation to provide context for next sweep of attention computation.

Dynamic Memory Networks (DMN) (Xiong et al., 2016) integrate an attention mechanism with a memory module and multi-modal bilinear pooling to combine features across views and predict attention over images for visual question answering task. Nam et al. (2017) iterated on this design to allow the memory update mechanism to reason over previous dual-attention outputs, instead of forgetting this information, in the subsequent sweep. The present work extends the multi-attention framework to leverage neural-based information flow control by dynamically routing it with neural gating mechanisms.

The very recent work (Zadeh et al., 2018a) also approaches multi-view learning with recourse to a system of recurrent encoders and attention mediated by global memory fusion. However, fusion takes place at the encoder cell level, requires hard alignment, and is performed online in one sweep so it cannot be informed by upstream context. The analysis window of the global memory is limited to the current and previous cell memories of each LSTM encoder, whereas our approach abstracts the shared memory update dynamics away from the ties of the encoding dynamics. Therefore our approach enables post-fusion and retrospective re-analysis of the entire cell memory history of all encoders at each analysis iteration.

3 Recursive Recurrent Neural Networks

Our approach is tailored to videos of single speakers, each divided into segments that roughly span one uttered sentence. We treat each segment as an independent datum constituting an individual

multi-modal event with its own annotation, such that there is no temporal dependence across any two segments. In the following exposition, each of the various mechanisms we describe (encoding, attention, fusion, and memory update) act on each segment in isolation of all others. We will use the terms "view" and "modality" interchangeably.

We refer to our recursively attentive analysis model as a *Recursive Recurrent Neural Network* (RRNN) since it resembles an RNN, but the hidden state and the next cell input are coupled in a recursion. At each step of the cell update there is no new incoming information; rather the *same* original inputs are re-weighted by a new attention query to form the new cell inputs (see discussion in Section 3.5 for more details).

3.1 Independent recurrent encoding

The major modelling assumption herein, is that a single, independent recurrent encoding of each segment of each modality is sufficient to capture a range of semantic representations that can be tapped by several shared external memory queries. Each memory query is formed in a separate stage of an iterated analysis over the recurrent codes. Concretely, modality-specific attention-weighted summaries $(\mathbf{a}^{(\tau)}, \mathbf{v}^{(\tau)}, \mathbf{t}^{(\tau)})$ at analysis iteration τ contribute to the update of a shared dense memory/context vector $\mathbf{m}^{(\tau)}$, which in turn serves as a differentiable attention query at iteration $\tau + 1$ (*cf.* Fig. 1). This provides a recursive mechanism for sharing information within and across sequences, so the recurrent representations of one view can be revisited in light of cross-modal cues gleaned from previous sweeps of other views. This is an efficient alternative to re-encoding each view on every sweep, and is more modular and generalisable than routing information across views at the recurrent cell level.

For each multi-modal sequence segment $\mathbf{x}^n = \{\mathbf{x}_a^n, \mathbf{x}_v^n, \mathbf{x}_t^n\}$, a view-specific encoding is realised via a set of independent bi-directional LSTMs (Hochreiter and Schmidhuber, 1997), run over segments $n \in [1, N]$:

$$\mathbf{h}_s^{fwd}[n, k_s] = LSTM(\mathbf{x}_s^n[k_s], \mathbf{h}_s^{fwd}[n, k_s - 1]) \quad (1)$$

$$\mathbf{h}_s^{bwd}[n, k_s] = LSTM(\mathbf{x}_s^n[k_s], \mathbf{h}_s^{bwd}[n, k_s + 1]) \quad (2)$$

$$\mathbf{h}_s[n, k_s] = [\mathbf{h}_s^{fwd}[n, k_s]; \mathbf{h}_s^{bwd}[n, k_s]] \quad (3)$$

Here, $s \in \{a, v, t\}$ denotes respectively audio, visual and textual modalities, and $k_s \in \{1, ..., K_s\}$ are view-specific state indices.

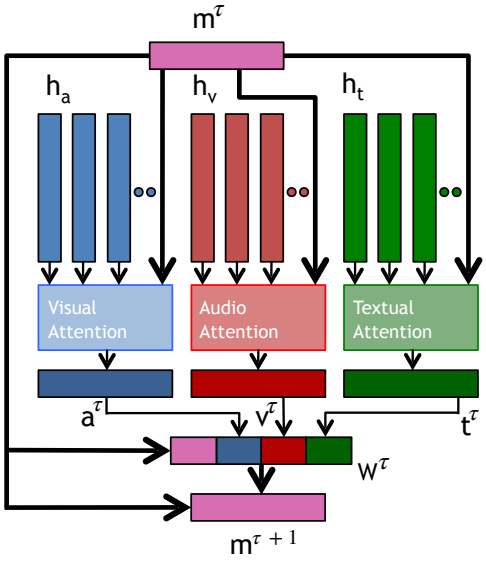

Figure 1: Schematic overview of the proposed neural architecture. Shared memory m^τ is updated with with the contextualised embeddings from a^τ, v^τ and t^τ.

The number of recurrent steps is view-specific (*i.e.* $K_a \neq K_v \neq K_t$) and is governed by the feature representation and sampling rate for the given view, *e.g.* number of word (embeddings) in a the text contained within a time-stamped segment. This is in contrast to Zadeh et al. (2018a), where the information in different views was grounded to a common time axis or the number of steps in an early stage, either via up-sampling or down-sampling. Thus the extracted representations in our approach preserve the inherent time-scales of each modality and avoid the need for hard alignment, satisfying desiderata (i) and (ii) outlined in Section 1. Note that the input sequences $\mathbf{x}_s^{(n)}$ may refer to either raw or pre-processed data (see Section 4 for details). In the remainder, we drop the segment id n to reduce notational clutter.

3.2 Globally-contextualised attention

We used a view-specific attention-based weighting mechanism to compute a contextualised embedding $\mathbf{c}_s$ for a view s. Encoder output $\mathbf{h}_s$ is stacked along time to form matrices $\mathbf{H}_s \in \mathbb{R}^{(D \times K_s)}$. A shared dense memory $\mathbf{m}^{(\tau=0)}$ is initialised by summing the time-average of the $\mathbf{H}_s$ across three modalities. $\mathbf{M}^{(\tau)}$ is then constructed by repeating the shared memory, $\mathbf{m}^{(\tau)}$, K_s times such that it has the same size as the corresponding context $\mathbf{H}_s$,

i.e. $\mathbf{H}_s, \mathbf{M} \in \mathbb{R}^{(D \times K_s)}$. An alignment function then scores how well $\mathbf{H}_s$ and $\mathbf{M}^{(\tau)}$ are matched

$$\tilde{\boldsymbol{\alpha}}_s^{(\tau)} = \text{align}(\mathbf{H}_s, \mathbf{M}^{(\tau)}). \quad (4)$$

The alignment mechanism entails a feedforward neural network with $\mathbf{H}_s$ and $\mathbf{M}^{(\tau)}$ as inputs. A softmax is applied on the network output to derive the attention strength $\boldsymbol{\alpha}$. This architecture resembles that in Bahdanau et al. (2014); concretely

$$\mathbf{R}^{(\tau)} = \tanh\big(\mathbf{W}_{s1}^{(\tau)}\mathbf{H}_s\big) \odot \tanh\big(\mathbf{W}_{s2}^{(\tau)}\mathbf{M}^{(\tau)}\big), \quad (5)$$

$$\tilde{\boldsymbol{\alpha}}_s^{(\tau)} = \mathbf{w}_{s3}^{(\tau)\mathrm{T}}\mathbf{R}^{(\tau)}, \quad (6)$$

$$\boldsymbol{\alpha}_s^{(\tau)}[k_s] = \frac{\tilde{\boldsymbol{\alpha}}_s^{(\tau)}[k_s]}{\sum_l \tilde{\boldsymbol{\alpha}}_s^{(\tau)}[l]}. \quad (7)$$

In Eq. (5), $\mathbf{W}_s^{(\tau)}$ (where $s \in \{s1, s2\}$) are square or fat matrices in the first layer of the alignment network, containing parameters governing the *self-influence* within view s and influence from the shared memory $\mathbf{M}$. For the majority of our experiments, we used the multiplicative method of Nam et al. (2017) to combine the two activation terms, but similar results were also obtained with the concatenative approach of Bahdanau et al. (2014). In eq. (6), $\mathbf{w}_{s3}^{(\tau)\mathrm{T}}$ is a vector projecting an un-normalised attention weight $\mathbf{R}$ onto an alignment vector $\tilde{\boldsymbol{\alpha}}$, which has the same dimensions as K_s. Finally, eq. (7) applies the softmax operation along the time step k_s.

Parameters $\mathbf{W}_{s1}, \mathbf{W}_{s2}, \mathbf{w}_{s3}$ for deriving attention strength $\boldsymbol{\alpha}_s$ are in general distinct parameters for each memory update step, τ. However, they could also be tied across steps. In the standard attention schemes, attention weight $\boldsymbol{\alpha}_s$ is a vector spanning across K_s. Note, that $\mathbf{w}_{s3}^{(\tau)}$ in eq. (6) could be replaced by a matrix-form $\mathbf{W}_{s3}^{(\tau)}$ to produce a *multi-head attention* weight (Vaswani et al., 2017). Alternatively, the transposition of network inputs can be performed such that attention scales each dimension, D, instead of each time step k. This can be seen as a variant of *key-value attention* (Daniluk et al., 2017), where the values differ from their keys by a linear transformation with weights governed by the alignment scores.

Each globally-contextualised view representation $\mathbf{c}_s$ is defined as the convex combination of the view-specific encoder outputs weighted by attention strength

$$\mathbf{c}_s^{(\tau+1)} = \sum_k \boldsymbol{\alpha}_s^{(\tau)}[k_s]\mathbf{h}_s[k_s]. \quad (8)$$

3.3 Shared memory update

The previous section described how the current shared memory state is used to modulate the attention-based re-analysis of the (encoded) inputs. Here we detail how the outcome of the re-analysis is used to update the shared memory state.

In contrast to the memory update employed in Nam et al. (2017), our approach includes a set of coupled gating mechanisms outlined below, and depicted schematically in Fig. 2:

$$\mathbf{g}_w^{(\tau)} = \sigma\big(\mathbf{W}_{wm}\mathbf{m}^{(\tau-1)} + \mathbf{W}_{ww}\mathbf{w}^{(\tau)} + \mathbf{b}_w\big) \quad (9)$$

$$\mathbf{g}_c^{(\tau)} = \sigma\big(\mathbf{W}_{cm}\mathbf{m}^{(\tau-1)} + \mathbf{W}_{cw}\mathbf{w}^{(\tau)} + \mathbf{b}_c\big) \quad (10)$$

$$\mathbf{g}_s^{(\tau)} = \sigma\big(\mathbf{W}_{sm}\mathbf{m}^{(\tau-1)} + \mathbf{W}_{ss}\mathbf{c}_s^{(\tau)} + \mathbf{b}_s\big) \quad \forall s \in \{a, v, t\} \quad (11)$$

$$\mathbf{u}^{(\tau)} = \tanh\big(\mathbf{W}_{um}\mathbf{m}^{(\tau-1)} + \mathbf{W}_{uw}\mathbf{g}_w^{(\tau)} \odot \mathbf{w}^{(\tau)} + \mathbf{b}_u\big) \quad (12)$$

$$\mathbf{m}^{(\tau)} = \big(1 - \mathbf{g}_c^{(\tau)}\big) \odot \mathbf{m}^{(\tau-1)} + \mathbf{g}_c^{(\tau)} \odot \mathbf{u}^{(\tau)}, \quad (13)$$

where $\mathbf{w}^{(\tau)} = [\mathbf{a}^{(\tau)}; \mathbf{v}^{(\tau)}; \mathbf{t}^{(\tau)}]$, $\mathbf{m}^{(0)} = \mathbf{0}$ and $\sigma()$ denotes an element-wise sigmoid non-linearity. The function of the view context gate defined in eq. (9) and invoked in eq. (12), is to block *corrupted* or *uninformative* view segments from influencing the proposed shared memory update content, $\mathbf{u}^{(\tau)}$. The attention mechanism, outlined in eq. (5)-(7), cannot fulfill this task alone since the full attention divided over a view segment must sum to 1 even if no part of that segment is pertinent/salient. The utility of this gating will be empirically demonstrated in noise-injection experiments in Section 5.

The new memory content $\mathbf{u}^{(\tau)}$ is written to the memory state according to eq. (12), subject to the action of the memory update gate defined in eq. (10). This update gate determines how much of the past global information should be passed on to contextualise subsequent stages of re-analysis. If parameters $\mathbf{W}_{s1}$, $\mathbf{W}_{s2}$, $\mathbf{w}_{s3}$ are untied across each re-analysis step, this update gate additionally accommodates short-cut or "highway" routing (Srivastava et al., 2015) of regression error gradients from the end of the multi-hop procedure back through the parameters of the earlier attention sweeps.

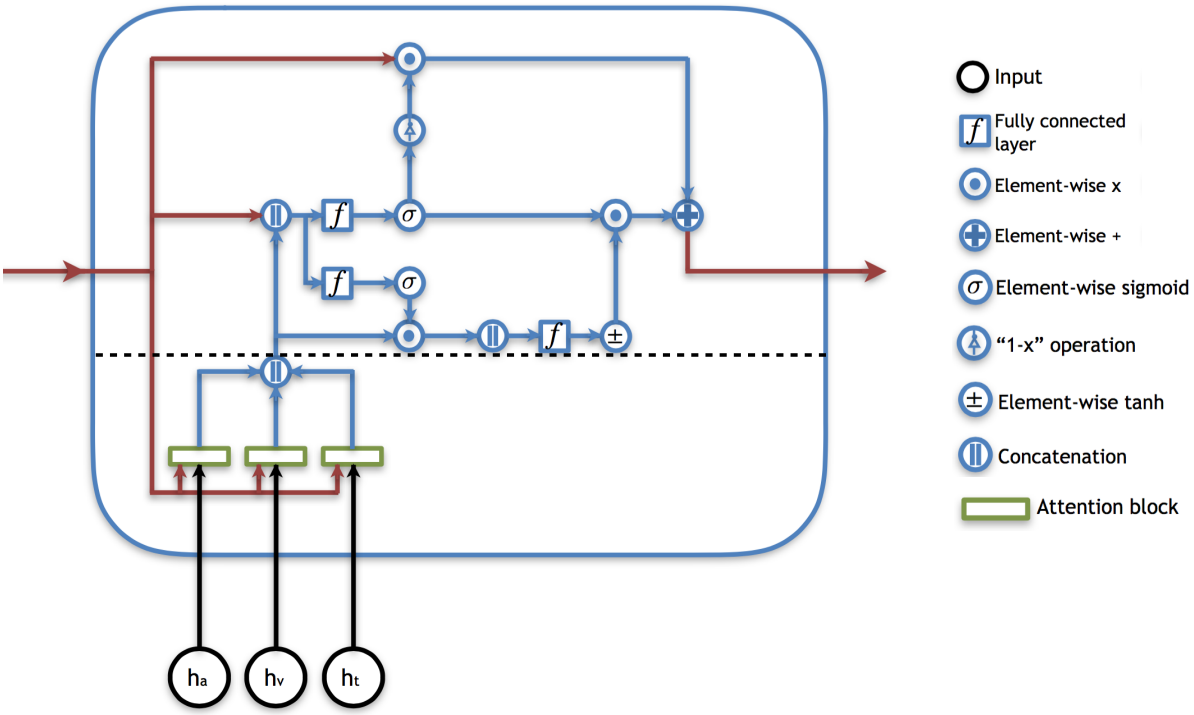

Figure 2: A detailed schematic of the proposed RRNN cell (left) and its legend (right). The routing above the dashed black line resembles that of a (non-recursive) GRU cell, where the concatenated attention output constitutes the cell's input. In this case, the cell's input at time τ is available only once the cell's state at time $\tau - 1$ has been computed. When the static representations $\{\mathbf{h}_a, \mathbf{h}_v, \mathbf{h}_t\}$ are instead viewed as the cell's input, then the cell forms a recursive RNN, which subsumes the attention mechanism as a cell sub-component.

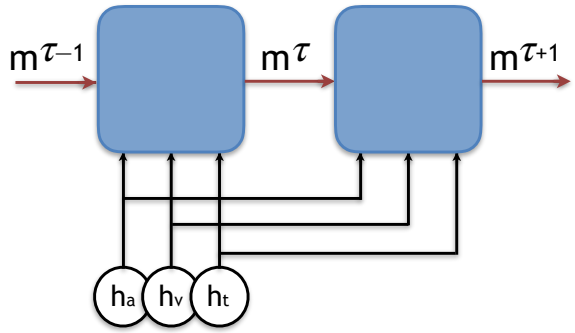

Figure 3: Two consecutive cells of a Recursive Recurrent Neural Network. Note that the cells share a common input, in contrast with a typical RNN which has a separate input to each cell.

3.4 Final Projection

After τ iterations of fusion and re-analysis, the resulting memory state $\mathbf{m}^{(\tau)}$ is passed through a final fully-connected layer to yield the output corresponding to a particular task (regression predictions or logits in case of classification). In our experiments we found that increasing τ yields meaningful performance gains (up to $\tau = 3$).

3.5 Recursive RNN: another perspective

The proposed gated memory update corresponds to maintaining an external recurrent cell memory that is recurrent in the consecutive analysis hops, τ, rather than the actual time-steps of the given modality, k_s. This allows the relevant memories of older hops to persist for use in the subsequent analysis hops.

The memory update equations (9)-(13) strongly resemble the GRU cell update (Cho et al., 2014); we treat concatenated view context vectors as the GRUs inputs, one at each analysis hop, τ. When viewed as a recurrent encoding of inputs $\{\mathbf{h}_s\}$, we refer to this architecture as a *recursive recurrent neural net* (RRNN), due to the recursive relationship between the cell's recurrent state and the attention-based re-weighting of the inputs. From this perspective, the attention mechanism forms a sub-component of the RRNN cell.

The key distinction from a typical GRU cell is that the *reset* or *relevance* gate $\mathbf{g}_w$ in a GRU typically gates the recurrent state ($\mathbf{m}^{(\tau)}$ in our case), whereas we use it to gate the input, allowing for

uninformative view contexts to be excluded from the memory update. Gating the recurrent state is essential for avoiding vanishing gradients over long sequences, which is not such a concern for our recursion lengths of ≈ 3. One could of course reinstate the gating of the recurrent state, should recursions grow to more appreciable lengths.
A further distinction is that here the GRU "inputs" (view contexts $\{\mathbf{a}^{(\tau)}, \mathbf{v}^{(\tau)}, \mathbf{t}^{(\tau)}\}$ in our case) are computed online as the memory state recurs, unlike the standard case where they are data or pre-extracted features available before the RNN begins to operate. Figure 3 depicts 2 consecutive RRNN cells, illustrating the recycling of the same cell inputs. Figure 2 shows the details of a single cell, which subsumes the globally-contextualised attention mechanism detailed in Section 3.2.

4 Experimental setup

Datasets. We evaluated our approach on CREMA-D (Cao et al., 2014), RAVDESS (Livingstone and Russo, 2012) and CMU-MOSEI (Zadeh et al., 2018b) datasets for multimodal emotion analysis. The first two datasets provide audio and visual modalities while CMU-MOSEI adds also text transcriptions. The CREMA-D dataset contains $\sim$7400 clips of 91 actors covering 6 emotions. The RAVDESS is a speech and song database comprising of $\sim$7300 files of 24 actors covering 8 emotional classes (including two canonical classes for "neutral" and "calm"). The CMU-MOSEI dataset consists of $\sim$3300 long clips segmented into $\sim$23000 short clips. In addition to audio and visual data, it contains also text transcriptions allowing evaluation of tri-modal models.

These datasets are annotated by a continuous-valued vector corresponding to multi-class emotion labels. The ground-truth labels were generated by multiple human transcribers with score normalisation and agreement analysis. For further details, refer to respective references.

Test conditions and baselines. Since each dataset consists of different emotion classification schema, we trained and evaluated all models separately for each of them. The training was performed in an end-to-end manner with $L2$ loss defined over multi-class emotion labels.

To establish a baseline, we evaluated a naive classifier predicting the test-set empirical mean intensities (with MSE loss function) for each output regression dimension. Similar baselines were obtained for other loss functions by training a model with just one parameter per output dimension on that loss, where the model has an access to the training labels but not the training inputs.

Evaluation. For CREMA-D and RAVDESS, we report the accuracy scores as these datasets contain labels for multiclass classification task.

For CMU-MOSEI, we report the result of the 6-way emotion recognition. Recursive models as described in Sec. 3 predicted the 6-dimensional emotion vectors. Their values represent the emotion intensity of the six emotion classes and are continuous-valued. Following Zadeh et al. (2018b), these predictions were evaluated against the reference emotions using the criteria of mean square error (MSE) and mean absolute error (MAE), summing across 6 classes. In addition, an acceptance threshold 0.1 was set for each dimension/emotion, and weighted accuracy (Tong et al., 2017) was computed.

Complementary views across modality. All experiments in this paper use independent recurrent encoding (Sec. 3.1). The encoding scheme differs for every modality. COVAREP (Degottex et al., 2014) was used for the audio modality. OpenFace (Amos et al., 2016) and FACET (iMotion, 2017) were used for visual one and Glove (Pennington et al., 2014) was used for encoding the text features.

Independent recurrent encoding used bi-directional view-specific encoders with 2×128 dimensional outputs on CREMA-D and RAVDESS and 2×512 on CMU-MOSEI. The complementary effects of multiple views from different modalities would be illustrated by controlling the available input views to different systems.

Attention. Global contextualised attention (GCA) was implemented for the emotion recognition systems. Global and view-specific memory were projected to the alignment space (Eq. (5)). The attention weights were computed (Eq. (6)-Eq. (7)) and the contextual view representation was derived (Eq. (8)). For more details, refer to Sec. 3.2. The encoder-decoder used a 128 dimensional (or 512 for CMU-MOSEI) fully-connected layer. A final linear layer mapped the decoder output to multi-class targets.

GCA was compared to standard "early" and "late" fusion strategies. In early fusion, encoders

Model	Modality	Accuracy
Human performance	Audio	40.9
COVAREP Features + LSTM Decoder	Audio	41.5
OpenFace Features + LSTM Decoder	Vision	52.5
Human performance	Vision	58.2
Human performance	Vision+Audio	63.6
(OpenFace features + LSTM) + (COVAREP Features + LSTM) + Dual Attention	Vision+Audio	65.0

Table 1: Results on the CREMA-D dataset across 8 emotions

Modality	Feature	Encoder	Attention	Accuracy
Audio	COVAREP	LSTM	Nil	41.25
Vision	OpenFace	LSTM	Nil	52.08
Audio + Vision	COVAREP, OpenFace	LSTM	GCA	58.33

Table 2: Results on the RAVDESS dataset across 8 emotions for normal speech mode

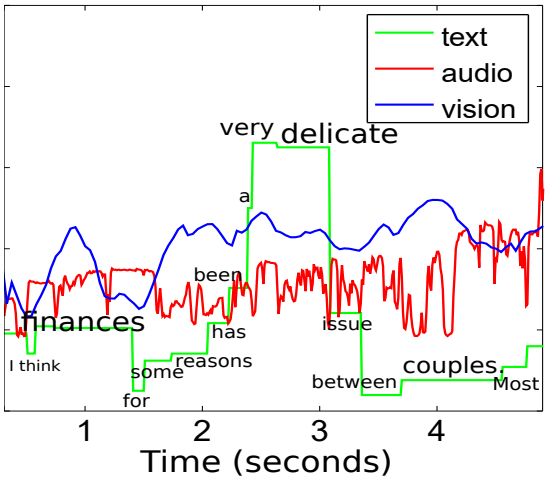

Figure 4: Visualisation of view-specific attention across time. Attention in the text modality focuses on the words "very" and "delicate" as cues for emotion recogntion. Also, the difference in oscillation rates between the audio and visual modalities is noted.

outputs across all views are resampled to their highest temporal resolution (*i.e.* audio, at 100Hz), and resulting (aligned) outputs are concatenated across views. We used similar encoder-decoder structure to one described in Sec 3.2 (Fig. 1), except that the three parallel blocks for modalities were reduced to one. In late fusion, the final-step encoder outputs from all modalities were independently processed by 1-layer feed-forward networks (Sec 3.4) and view-specific multi-class targets were combined using linear weighting.

Memory updates and ablation study. GCA was enhanced with the extra gating functions (*cf.* Eq. (9)-(13), Sec. 3.3) . The extended system was compared with the GCA system on CMU-MOSEI data. To this end, we perform an ablation study using the test data corrupted by additive Gaussian white noise added to the visual modality.

5 Results

Table 1 and 2 show the results of emotion recognition on the CREMA-D and RAVDESS dataset respectively. Audio, visual and the joint use of bi-modal information were compared using identification accuracy. Models trained on the visual modality consistently outperformed models that use solely audio data. Highest accuracy was achieved when the audio and visual modality were jointly modelled, giving 65% and 58.33% on the two datasets. Interestingly, the joint bi-modal system outperformed human performance on CREMA-D (Cao et al., 2014) by 1.4%.

On CMU-MOSEI, the errors between the reference and hypothesis six-dimensional emotion vectors were computed and the results were shown in Table 3.

The use of visual modality resulted in the lowest mean square error (MSE). Meanwhile, when evaluated by mean absolute error (MAE) and weighted accuracy (WA), text modality gave the best performance. Basic techniques in combining information among modalities was not very effective, as indicated by the neglible gain in early and late fusion model.

Globally contextualised attention (GCA) gave an MSE of 0.4696. Gating on global and view-specific memory updates led to further improvements to 0.4691. The improvement in terms of MAE is even more significant (from 0.9412 to 0.8705).

Figure 4 visualises the attention weights in different modalities on a CMU-MOSEI test sentence. The x-axis denotes time t and y-axis is the magnitude of attention $\alpha_s(t)$ in different views $s \in \{a, v, t\}$. The transcribed text was added alongside the attention profile of the textual modality to align the attention weights with the recording. It can be seen that the GCA emotion recognition

Modality	Feature	Encoder	Attention/Fusion	Corruption	MSE	MAE	WA
Text (T)	Word-vec	LSTM	Nil	Nil	0.6326	0.9830	0.5485
Audio (A)	COVAREP	LSTM	Nil	Nil	0.6049	1.0562	0.5249
Vision (V)	FACET	LSTM	Nil	Nil	0.5026	0.9909	0.5476
T+A+V	COVAREP, FACET, Word-vec	LSTM	Early fusion	Nil	0.5319	0.7694	0.5188
T+A+V	COVAREP, FACET, Word-vec	LSTM	Late fusion	Nil	0.5047	0.9825	0.5889
T+A+V	COVAREP, FACET, Word-vec	LSTM	GCA	Nil	0.4696	0.9412	0.6163
T+A+V	COVAREP, FACET, Word-vec	LSTM	GCA	Vision	0.5034	0.9920	0.6068
T+A+V	COVAREP, FACET, Word-vec	LSTM	GCA + Gating	Nil	0.4691	0.8705	0.5765
T+A+V	COVAREP, FACET, Wrod-vec	LSTM	GCA + Gating	Vision	0.4742	0.8857	0.5688

Table 3: Results on CMU-MOSEI dataset

system was trained to attend dynamically to features of varying importance across the time, unlike systems performing early or late fusion. Attention weights of text modality show a clear jump for the words "very" and "delicate". The word "very", combined with an adjective, is often a strong cue to sentiment analysis, resulting in a spike in attention. The subject in this clip was speaking mostly in a neutral tone, with a nod and slight frowning towards the beginning of the sentence. This may correspond to the first peak in the attention trajectory of visual data. The weight of audio modality exhibited a higher oscillation rate compared to the counterpart on visual data. COVAREP features had $4\times$ higher temporal frequency than FACET.

Finally, we verified contribution of the gating system to the GCA using the corrupted visual data. When the GCA system is used without the gating mechanism, corrupted data results in increased MSE (from 0.4696 to 0.5034) and MAE (from 0.9412 to 0.9920). This is in contrast to the full system with gating (GCA + Gating in Table 3). The system cancels the effects of additive visual noise, which is evidenced by the small gap in MSE (0.4691 vs 0.4742) and MAE (0.8705 vs 0.8857) between clean and noisy data.

6 Conclusion

We have presented an approach for combining sequential, heterogeneous data. An external memory state is updated recursively, using globally-contextualised attention over a set of recurrent view-specific state histories. Our model was tested on the challenging tasks of emotion recognition from audio, visual, and textual data on three large-scale datasets. The complementary effect of joint modelling of emotions using multi-modal data was consistently shown across experiments with multiple datasets. Importantly this approach eschews hard alignment of the data streams, allowing view-specific encoders to respect the inherent dynamics of its input sequence. Encoder state histories are fused into cross-modal features via an attention mechanism that is modulated by a shared, external memory. The control of information flow in this fusion is further enhanced by using a GRU-like gating mechanism, which can persist shared memory through multiple iterations while blocking corrupted or uninformative view-specific features. In future study, it would be interesting to investigate more structured fusion operations such as sparse tensor multilinear maps (Benyounes et al., 2017).

References

Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. Openface: A general-purpose face recognition library with mobile applications. Technical report, CMU-CS-16-118, CMU School of Computer Science.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

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

Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: multimodal tucker fusion for visual question answering. *CoRR*, abs/1705.06676.

Houwei Cao, David G. Cooper, and Michael K. Keutmann. 2014. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*.

Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*.

Michal Daniluk, Tim Rocktäschel, Johannes Welbl, and Sebastian Riedel. 2017. Frustratingly short at-

tention spans in neural language modeling. *CoRR*, abs/1702.04521.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP – a collaborative voice analysis repository for speech technologies. In *ICASSP*.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. In *Neural Computation*.

iMotion. 2017. Facial expression analysis.

Yelin Kim, Honglak Lee, and Emily Mower Provost. 2013. Deep learning for robust feature generation in audiovisual emotion recognition. In *ICASSP*.

S. R. Livingstone and F. A. Russo. 2012. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. *CVPR*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *ICML*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *EMNLP*.

Vikas Sindhwani and Partha Niyogi. 2005. A co-regularized approach to semi-supervised learning with multiple views. In *Proceedings of the ICML Workshop on Learning with Multiple Views*.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. 2015. Training very deep networks. In *NIPS*.

Sainbayar Sukhbaatar, arthur szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. In *NIPS*.

Edmund Tong, Amir Zadeh, Cara Jones, and Louis-Philippe Morency. 2017. Combining human trafficking with multimodal deep models. *ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, and Lukasz Kaiser. 2017. Attention is all you need. *NIPS*.

Valentin Vielzeuf, Stéphane Pateux, and Frédéric Jurie. 2017. Temporal multimodal fusion for video emotion classification in the wild. *CoRR*, abs/1709.07200.

Jason Weston. 2017. Memory networks for recommendation. In *RecSys*.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *CoRR*, abs/1304.5634.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. *CoRR*, abs/1802.00927.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Emdund Tong, Erik Cambria, Minghai Chen, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *ACL*.

Shiqing Zhang, Shiliang Zhang, Tiejun Huang, Wen Gao, and Qi Tian. 2017. Learning Affective Features with a Hybrid Deep Model for Audio-Visual Emotion Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*.