Dynamic Fusion for Multimodal Data

Gaurav Sahu
University of Waterloo
gaurav.sahu@uwaterloo.ca

Olga Vechtomova
University of Waterloo
ovechtom@uwaterloo.ca

Abstract

Effective fusion of data from multiple modalities, such as video, speech, and text, is challenging pertaining to the heterogeneous nature of multimodal data. In this paper, we propose dynamic fusion techniques that model context from different modalities efficiently. Instead of defining a deterministic fusion operation, such as concatenation, for the network, we let the network decide "how" to combine given multimodal features in the most optimal way. We propose two networks: 1) transfusion network, which learns to compress information from different modalities while preserving the context, and 2) a GAN-based network, which regularizes the learned latent space given context from complimenting modalities. A quantitative evaluation on the tasks of machine translation, and emotion recognition suggest that such adaptive networks are able to model context better than all existing methods.

1 Introduction

Multimodal deep learning is an active field of research where, for a single event, one is presented with information across multiple modalities, such as video, speech, and text, so that they may be combined to gain a better contextual understanding. Combining, or more precisely, fusing information from multiple modalities is, thus, a vital step for any multimodal task. However, multimodal data is highly heterogeneous in nature making fusion a challenging task. Moreover, the extent to which signals from complimenting modalities are helpful for a downstream task is not always clear.

The most common fusion technique used in the literature is concatenation, where representations from all the modalities are simply concatenated. However, this results in a shallow network (Ngiam et al., 2011), where the model, instead of learning inter-modal features, focuses more on learning intra-modal features. Later, Zadeh et al. (2017) proposed tensor fusion network (TFN), in which, the unimodal, bimodal, and trimodal interactions are modelled using a 3-fold Cartesian product. However, it imposes high computational requirements as information from all the modalities is used as-is, without any prior information extraction. Liu et al. (2018) proposed a low rank multimodal fusion technique (LMF) to address the previous problem. Such fusion techniques are effective but often result in a complex architecture with a lot of computation.

In this paper, we propose dynamic fusion techniques which allow the model to decide "how" to combine multimodal data for an event in the best possible manner. The first technique, transfusion, learns to compress multimodal information while preserving as much meaning as possible. Our second technique employs an adversarial network which regularizes the learned latent space for a target modality (text, in our case) according to information presented by the remaining complimentary modalities. Since our models are generic in nature, the need to specify a pre-determined fusion operation such as concatenation or Cartesian product is alleviated, and the network is incentivized to model intermodal interactions by itself. Moreover, our models are based on lightweight components such as linear transformation layers, thereby, checking unnecessary computational load.

We evaluate our models on three benchmark datasets, namely, How2 (Sanabria et al., 2018), Multi30K (Specia et al., 2016), and IEMOCAP (Busso et al., 2008). Quantitative evaluation shows that our models outperform the existing state-of-the-art methods. The rest of the paper is structured as follows: Section 2 covers relevant work, Section 3 discusses the proposed architectures in detail, Section 4 describes the experimental setup, Section 5 shows results, and Section 6

contains our concluding remarks.

2 Related Work

In this section, we briefly review some previous works related to our task. Most earlier works in multimodal deep learning focused on traditional shallow classifiers such as support vector machines (Cortes and Vapnik, 1995) and Naive Bayes classifiers (Morade and Patnaik, 2015) to exploit bimodal data. Inspired by the success of deep learning over the last decade across multiple tasks, Ngiam et al. (2011) train end-to-end deep graph neural networks to reconstruct missing modalities at inference time. They demonstrate that better features for one modality can be learned if relevant data from different modalities is available at training time; however, they employ simple concatenation for fusion. Hence, the joint representation learned is shallow and is not guaranteed to learn inter-modal connections. Their findings were later verified by Srivastava and Salakhutdinov (2012), who use a Deep Boltzmann Machine (Salakhutdinov and Hinton, 2009) to generate/map data from the image and text modality.

Huang et al. (2018) construct a multilingual common semantic space to achieve better machine translation performance by extending correlation networks (Chandar et al., 2016). They use multiple non-linear transformations to repeatedly reconstruct sentences from one language to another and finally build a common semantic space for all the different languages. Fusion techniques, such as TFN (Zadeh et al., 2017) and LMF (Liu et al., 2018), were also proposed but the problem of efficiently modelling context in multimodal samples still remains unsolved.

3 Proposed methods

In this section, we describe the two methods developed for efficiently combining multimodal inputs.

3.1 Transfusion

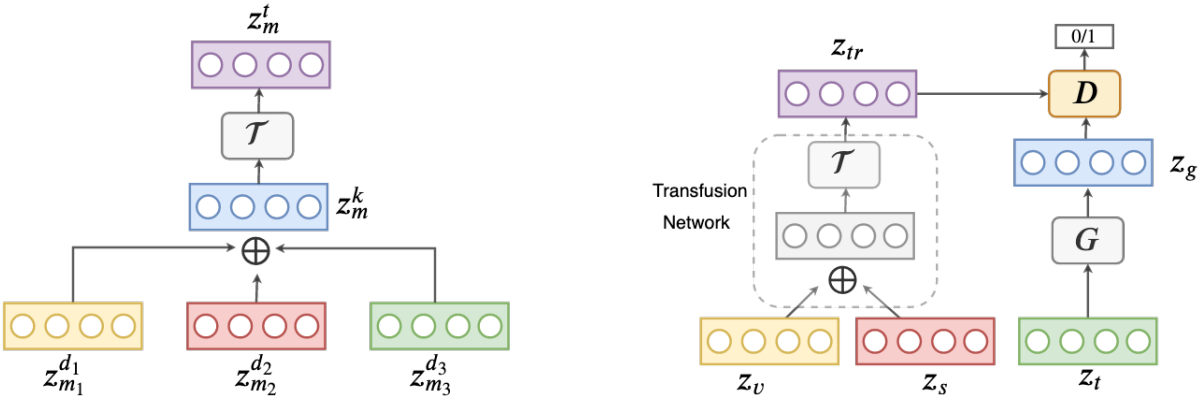
Most fusion methods proposed in the past either result in a shallow network, such as concatenation, where the network learns more intramodal features than intermodal features, or are computationally expensive, such as tensor fusion (Zadeh et al., 2017), and there is no intelligent feature extraction. In both cases, the fusion operation is specified by the user, and the network does not have the freedom to learn/relate intermodal features on its own.

In order to mitigate the "staticness" of previous methods, we propose a dynamic yet simple fusion technique, called *transfusion*, where the model learns to extract intermodal features by itself. In this method, we first concatenate the latent vectors from different modalities, and then pass them through a transformation layer to get a *transfused* latent vector whose dimension is much lower than the dimension of the input concatenated vector. We then minimize the euclidean distance between the transfused and the previously obtained concatenated vector. Note that in order to do so, we need to augment the transfused vector with a null vector of appropriate shape to match the concatenated vector's dimension. Augmentation with a null vector provides another important advantage: it makes sure that the transformation layer is not arbitrarily outputting signals from the previously concatenated latent vector. Instead, it is incentivized to "compress" the information without losing any important cues as much as possible. In other words, it increases correlation between the transfused and the concatenated latent vector. Such a method is applicable to any scenario where multiple features need to be combined. For example, they can be used combine the forward and backward hidden states of the LSTM, instead of pooling methods such as, 1D pooling, max pooling, sum pooling or even simple concatenation.

We now discuss the transfusion network in detail. We pose fusion of multimodal inputs as a compression problem, where we must retain as much information from the individual modalities as possible. Given n ($\leq$ 3 in our case) d-dimensional multimodal latent vectors, $z_{m_1}^{d_1}, z_{m_2}^{d_2}, \ldots, z_{m_n}^{d_n}$, we first concatenate them to obtain a vector, z_m^k, where $k = \sum_i^n d_i$. Then, we apply a transformation, $\mathcal{T}$, to z_{m_k}, reducing its dimension to t. Finally, we calculate the loss, J_{tr} , between the transfused latent vector, z_m^t, and z_m^k. We use MSE as our loss function for this case. These steps could be followed in Figure 1(a) and the loss for transfusion network is given by:

$$J_{tr} = \frac{([z_m^t; \mathbf{0}^{k-t}] - z_m^k)^2}{k} \quad (1)$$

Here ; represents the concatenation operator.



(a) Transfusion network

(b) Adversarial Network

Figure 1: Proposed architectures. (a) Transfusion network: Assuming that $z_{m_1}^{d_1}$, $z_{m_2}^{d_2}$, and $z_{m_3}^{d_3}$ represent the video, speech, and text latent vectors respectively, we first concatenate them to obtain z_m^k. It is then passed through the transfusion layer $\mathcal{T}$ which outputs the "transfused" vector z_m^t. Finally, we optimize the loss between z_m^t (augmented with a null vector of appropriate shape) and z_m^k. (b) GAN for latent-space regularization: Assuming that z_s, z_v, and z_t are the latent speech, video, and text vectors, respectively, we first transfuse z_s and z_v to give z_{tr}. Simultaneously, we pass z_t through the generator G to give z_g. When training the generator, we minimize the loss between z_{tr} and z_g and when training the discriminator D, we use z_{tr} and z_g as the two different sources of input. **Note:** $\bigoplus$ denotes concatenation.

3.2 GAN for Latent Space Regularization

In addition to the "staticness" of existing methods, there is also the challenge of distinguishing between ambiguous cases. For instance, the sentence "Kevin, this is hilarious," could be said in a funny, sarcastic, or angry manner. Currently, the existing methods, even when fed with the corresponding speech vector, cannot efficiently distinguish between similar but different emotions such as the sarcastic and the angry setting. We hypothesize that this is due to the fact that they are not learning the conditional distribution of sentiment given an utterance (an utterance includes input from all available modalities).

In order to mitigate this issue, we propose an adversarial training regime that is incentivized to learn the desired conditional distribution. For a task such as emotion recognition, this would be sentiment given an utterance, and for a more challenging generation task, the model could learn to model a more complex behaviour, such as the association of different sentences based on how similar they sound, in addition to their polarity. We show in our experiments that our GAN-based approach is better able to relate intermodal features as compared to the existing methods.

We now describe the GAN-based architecture in detail. For a given multimodal sample x, we first encode the inputs from each modality (speech, visual and text) to get the respective latent vectors, z_s, z_v, and z_t. Fixing a target modality (text in our case), we pass z_t through a generator to obtain $z_g = G(z_t)$, and transfuse the remaining latent vectors, z_s, and z_v simultaneously to obtain z_{tr}. In the event where we have input from only one modality in addition to text, we do not need any transfusion, and can simply treat the other modality's vector as z_{tr}. Finally, we train the network in adversarial fashion, labelling z_{tr} as positive samples and z_t as negative samples. The adversarial loss, J_{adv}, is given below:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x \sim p_{z_{tr}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_{z_t}(z)}[\log(1 - D(z_g))] \quad (2)$$

Overall, the generator G is incentivized to align features of the target modality (text, in our case) with features from the complimentary modalities (speech and video. in our case), and the discriminator tries to identify the type of its input. Learning the latent space in such an adversarial manner induces a clustering effect on the latent space, where texts attached to similar sounds and visuals are grouped together.

Model	Source modalities	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Sanabria et al. (2018)	t	-	-	-	54.4
Sanabria et al. (2018)	s-v-t	-	-	-	54.4
Lal et al., (2019)	s-v-t	-	-	-	51.0
Raunak et al., (2019)	t	-	-	-	55.5
Wu et al., (2019)	s-v-t	-	-	-	56.2
Seq2Seq	t	48.32	30.63	20.79	14.60
	s	20.11	7.01	3.12	1.57
	v	19.28	6.35	2.33	1.03
Seq2Seq + attn.	t	79.21	67.34	52.67	47.34
Transfusion Net (Ours)	s-t	56.31	33.82	24.63	21.45
	s-v-t	57.18	34.71	25.15	22.10
Transfusion Net + attn (Ours)	s-t	80.34	67.83	61.27	55.01
	s-v-t	85.23	71.95	69.54	57.80
Adversarial Net (Ours)	s-t	60.65	37.43	30.01	28.87
	s-v-t	61.23	38.76	31.23	29.31
Adversarial Net + attn (Ours)	s-t	82.25	69.43	64.33	56.5
	s-v-t	**89.66**	**74.48**	**71.29**	**59.83**

Table 1: Results for machine translation on `How2` dataset. 't', 's', 'v' represent the text, speech, and video modalities, respectively.

Model	BLEU 4	Meteor
Baseline	36.3	56.9
Grönroos et al. (2018)	44.1	**64.3**
Tranfusion + attn (Ours)	42.31	61.7
Adv. Network (Ours)	**44.23**	63.8

Table 2: Results for machine translation on `Multi30K` dataset. **Note:** All methods use 'v' and 't' as the source modalities except the baseline.

4 Experimental Setup

We evaluate our methods on the tasks of machine translation and emotion recognition. In this section we describe the datasets used and the details of the training process.

4.1 Datasets

4.1.1 Emotion Recognition: IEMOCAP

We use the IEMOCAP dataset (Busso et al., 2008) released by researchers from the University of Southern California (USC). It contains five recorded sessions of conversations from ten speakers and amounts to nearly 12 hours of audio-visual information along with transcriptions. It is annotated with eight categorical emotion labels, namely, angry, happy, sad, neutral, surprised, fear, frustrated and excited. It also contains dimensional labels such as values of the activation and valence from 1 to 5; however, they are not used in this work. The dataset is already split into multiple utterances for each session and we further split each utterance file to obtain wav files for each sentence. This was done using the start timestamp and end timestamp provided for the transcribed sentences. This results in a total of ∼10K audio files which are then used to extract features.

We identify the task as an emotion recognition problem, where, given a sentence and its audio, we wish to predict the correct emotion.

4.1.2 Machine Translation: How2

We evaluate our models on the multimodal `How2` dataset (Sanabria et al., 2018), which is comprised of 79,114 instructional videos in addition to word-level time alignments to the ground-truth English subtitles and their respective crowd-sourced Portuguese translations. A brief description of the video clip is also included to encourage future work on image captioning. This dataset was created by scraping videos along with their metadata from YouTube using a keyword based spider, and manually extracting and processing visual, auditory and textual features.

Unlike other popular multimodal datasets that are frequently featured in multimodal deep learning literature, such as `CUAVE` (Patterson et al., 2002) and `AVLetters` (Matthews et al., 2002),

Model	Precision	Recall	F1-score	Accuracy
LSTM + attn (t)	53.2	40.6	43.4	43.6
LSTM + attn ([s;t])	66.1	65.0	64.7	64.2
Yoon et al. (2018)	-	-	-	71.8
Yoon et al. (2019)	-	-	-	76.5
Transfusion + attn (Ours)	75.3	77.4	76.3	77.8
Adversarial + attn (Ours)	**77.3**	**79.1**	**78.2**	**79.2**

Table 3: Results for speech emotion recognition on IEMOCAP dataset.

the How2 data is in fact trimodal, therefore making it suitable to evaluate the contribution of each modality towards different tasks. The speech features are 43-dimensional vectors extracted from 16 kHz raw speech using the toolkit Kaldi[1]. A 2048-dimensional video feature vector is also derived per 16 frames in each video. Further, as a large-scale multilingual dataset, it enables a convenient medium for neural machine translation in our project.

4.1.3 Machine Translation: Multi30K

We also run experiments on the Multi30K dataset (Specia et al., 2016). The dataset contains pairs of sentences in English and many languages such as French, German, and Czech. Each sample in the dataset has an image, its description in the source language and its translated version. We only run experiments on the En-Fr version of the dataset.

4.2 Hyperparameters and Training Details

We use Bidirectional LSTM units (Hochreiter and Schmidhuber, 1997) of size 256 to encode text and a unidirectional LSTM of size 256 as the decoder. We preprocess the text where we lowercase and normalize all the words to remove any punctuations and non-ascii characters. We train a word2vec model on all our datasets with embeddings of dimension 300. We finetune a pre-trained VGG (Simonyan and Zisserman, 2014) to encode images in the Multi30K dataset. For experiments on the How2 dataset, we use the already provided feature vectors for speech and video.

For training, the desired fusion network is used before the final step of each of the tasks, i.e., before the final classification layer for emotion recognition, and before the decoder for machine translation. All the models are trained in an end-to-end manner.

[1]https://github.com/kaldi-asr/kaldi

5 Results

Results of our experiments on the How2, Multi30K and IEMOCAP dataset are shown in Tables 1, 2 and 3, respectively. For the relatively easier task of emotion recognition, we observe that our models perform well across all the evaluation metrics. For the more difficult task of machine translation, we note that our best performing model beats the existing methods in terms of BLEU scores, despite being much lighter than the previous models.

6 Conclusions and Future Work

In this paper, we proposed two dynamic fusion techniques that allowed for better multimodal fusion with an added advantage of being very lightweight. Instead of using a pre-defined fusion operation, we let the model decide the most optimal way to extract signals from multiple modalities. Our results indicate that such adaptive models are better and computationally more efficient for the given task.

One interesting aspect to study would be that of multimodal feature alignment that could help reduce the heterogeneity in multimodal inputs.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural computation*, 28(2):257–285.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham,

Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Kevin Knight. 2018. Multi-lingual common semantic space construction via cluster-consistent word embedding. *arXiv preprint arXiv:1804.07875*.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Iain Matthews, Timothy F. Cootes, Andrew Bangham, Stephen Cox, and Richard Harvey. 2002. Extraction of visual features for lipreading. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24:198 – 213.

Sunil S. Morade and Suprava Patnaik. 2015. Comparison of classifiers for lip reading with cuave and tulips database. *Optik - International Journal for Light and Electron Optics*, 126(24):5753–5761.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Eric K Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N Gowdy. 2002. Cuave: A new audio-visual database for multimodal human-computer interface research. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages II–2017. IEEE.

Ruslan Salakhutdinov and Geoffrey Hinton. 2009. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Lucia Specia, Stella Frank, Khalil Simaan, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826. IEEE.

Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung. 2018. Multimodal speech emotion recognition using audio and text. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 112–118. IEEE.

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