Adaptive Fusion Techniques for Multimodal Data

Gaurav Sahu, Olga Vechtomova
University of Waterloo
{gsahu, ovechtom}@uwaterloo.ca

Abstract

Effective fusion of data from multiple modalities, such as video, speech, and text, is challenging due to the heterogeneous nature of multimodal data. In this paper, we propose adaptive fusion techniques that aim to model context from different modalities effectively. Instead of defining a deterministic fusion operation, such as concatenation, for the network, we let the network decide "how" to combine a given set of multimodal features more effectively. We propose two networks: 1) Auto-Fusion, which learns to compress information from different modalities while preserving the context, and 2) GAN-Fusion, which regularizes the learned latent space given context from complementing modalities. A quantitative evaluation on the tasks of multimodal machine translation and emotion recognition suggests that our lightweight, adaptive networks can better model context from other modalities than existing methods, many of which employ massive transformer-based networks.[1]

1 Introduction

Multimodal deep learning is an active field of research, where for a single event, one has information across multiple modalities, such as video, speech, and text. Human brains can easily and perpetually perceive the context of an event from such heterogeneous data; however, it is not a trivial task for a computer system. In order for the machine to gain a contextual understanding, heterogeneous inputs must be combined first. Combining, or more precisely, *fusing* multimodal inputs is, thus, a vital step for any multimodal task. Naturally, a better fusion method will help a multimodal system learn better, ultimately enhancing its performance for a given task.

The most common fusion technique used in the literature involves the concatenation of representations from all the available modalities. However, this results in a shallow network (Ngiam et al., 2011), and the network focuses more on learning intra-modal features, ignoring inter-modal dynamics altogether. Later, Zadeh et al. (2017) proposed Tensor Fusion Network (TFN), which models the unimodal, bimodal, and trimodal interactions using a 3-fold Cartesian product. TFN performs better than simple concatenation; however, it imposes high computational requirements since it projects all the information from input modalities to a dense 3-D space as-is, without any prior information extraction. The computational overhead grows exponentially with respect to the dimensionality of unimodal features. Liu et al. (2018) proposed a low-rank multimodal fusion technique (LMF) to address the previous problem. Such fusion techniques are useful but often result in a complex architecture. Moreover, the fusion methods mentioned above focus only on combining individual unimodal features rather than combining *and* extracting useful information simultaneously. This means that the final predictor module (decoder in a Seq2Seq network (Sutskever et al., 2014), for example) bears an additional responsibility of identifying useful signals to focus on.

This paper addresses these issues by proposing adaptive fusion techniques that allow the model to decide "how" to combine multimodal data more effectively for an event. The first technique, Auto-Fusion, learns to compress multimodal information while preserving as much meaning as possible. The second technique, GAN-Fusion, employs an adversarial network that regularizes the learned latent space for a given target modality complying with the information presented by complementary modalities. Since our models are generic, the need to specify a pre-determined fusion operation

---

[1]Code for our experiments: https://github.com/demfier/philo/

such as Cartesian product is alleviated, and this further incentivizes the network to model multimodal interactions by itself. Moreover, our techniques are lightweight relative to the existing heavier counterparts (Vaswani et al., 2017; Grönroos et al., 2018), thereby preventing unnecessary computational load.

We evaluate our models on three benchmark datasets:

1. the How2 dataset (Sanabria et al., 2018) with multimodal input for English-Portuguese translation.

2. the Multi30K dataset (Elliott et al., 2016), which contains parallel corpora for multimodal machine translation, and

3. the IEMOCAP dataset (Busso et al., 2008) which contains multimodal data for emotion detection.

A quantitative evaluation shows that our models outperform the existing state-of-the-art methods in terms of BLEU scores (Papineni et al., 2002) for machine translation and Precision, Recall, and F1-score for emotion recognition. Our ablation studies also indicate that the learned multimodal representations are robust; they perform reasonably well even after removing information from a target modality. We now summarize our main contributions as follows:

1. We propose two lightweight, adaptive techniques for better multimodal fusion of data: Auto-Fusion and GAN-Fusion.

2. We propose a multi-task framework for end-to-end training of multimodal networks (for both classification and generation).

The rest of the paper is structured as follows: Section 2 covers relevant work, Section 3 discusses the proposed methodologies and overall architecture, 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 previous work related to our task. Most earlier works in multimodal deep learning focus 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 during inference. 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 model intermodal interactions. Their findings were later verified by Srivastava and Salakhutdinov (2012), who use a Deep Boltzmann Machine (Salakhutdinov and Hinton, 2009) to generate 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 reconstruct sentences from one language to another repeatedly and finally build a common semantic space for all the different languages. To address the shallowness exhibited by some earlier fusion methods, techniques such as TFN (Zadeh et al., 2017), LMF (Liu et al., 2018) and T2FN (Liang et al., 2019), were proposed that aim to capture *both* intra- and inter-modal dynamics simultaneously; however, the problem of effectively modelling context in multimodal samples remains unsolved.

More recently, Multimodal Transformer (MulT) (Tsai et al., 2019) was proposed to align data from different modalities implicitly. On a high-level, MulT leverages cross-modal attention modules for each modality, each of which is responsible for aligning (or attending to) the target modality vector with the complementary modalities'. It also imposes substantial computational overhead due to the use of transformer networks (Vaswani et al., 2017). Our methods, as discussed in detail in Section 3, use much simpler components. For instance, we use at most one attention module compared to multiple self-attention heads in a transformer. Variational Mixture-of-Experts Autoencoders (Shi et al., 2019), a class of deep generative multimodal frameworks, were employed to learn a synergic shared representation for multiple modalities; however, scaling of such a model for all the modalities (video, speech, and text) simultaneously and for a more complex task as multimodal machine transla-
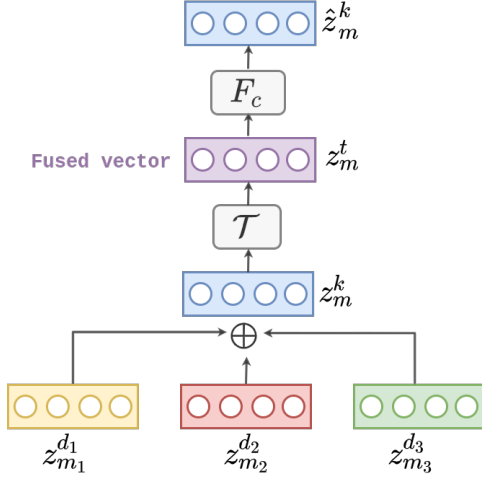
(a) Auto-Fusion network
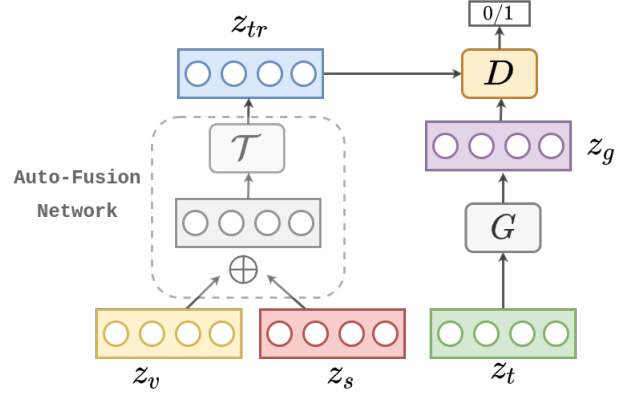
(b) GAN-Fusion Network

Figure 1: Proposed architectures. (a) Auto-Fusion 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 $\mathcal{T}$ which outputs the "autofused" vector z_m^t. We then obtain the reconstructed concatenated vector $\hat{z}_m^k$ by passing the autofused vector through F_c, another transformation layer. Finally, we optimize the loss between $\hat{z}_m^k$ and z_m^k. (b) GAN-Fusion module for the text modality: Assuming that z_s, z_v, and z_t are the latent speech, video, and text vectors, respectively, we first autofuse z_s and z_v to give z_{tr}. Simultaneously, we pass z_t through the generator G, along with some noise, to get z_g. The generator loss tries to match z_{tr} and z_g and discriminator D tries to distinguish between z_{tr} and z_g, the two sources of input. **Note:** $\oplus$ denotes concatenation.

tion is currently unexplored.

3 Proposed methods

This section will discuss the proposed methodologies for effectively fusing inputs from multiple modalities and describe the overall architecture of our models for classification and generation. Most fusion techniques proposed in the literature, such as concatenation, and TFN, involve a deterministic operation for constructing the joint multimodal representation. For instance, in TFN, the 3-fold Cartesian product of unimodal features is used for prediction. The method focuses more on learning rich unimodal features. However, there is no such "learning" procedure for joint representation; they are simply constructed by combining unimodal features in a specific fashion (here, by Cartesian product.) In this paper, we will refer to such techniques as *static* fusion techniques. Since there is no particular learning procedure for the joint representation, it becomes challenging for the final predictor module to model the complex dynamics of multimodal features. In other words, the model is unable to utilize multimodal information effectively.

On the other hand, fusion methods such as LMF and MulT are *adaptive* because they involve a cog-

nitive feature processing step to construct the joint representation. In LMF, it is the decomposition module, and in MulT, it is the final feed-forward fusion mechanism. We refer the reader to Liu et al. (2018) and Tsai et al. (2019) for more detailed explanation of the models.

Our fusion methods involve the concatenation of unimodal embeddings as an initial step. To avoid any conflicts with past works, we will only consider steps *after* concatenation as a part of our fusion method because we do not use the concatenated vector for final prediction; it is only a preliminary step. Therefore, in order to mitigate the "staticness" of existing fusion methods, we propose two adaptive yet simple fusion techniques, *Auto-Fusion* and *GAN-Fusion*. They aim to effectively combine multimodal inputs and mitigate the problem of shallowness and computational overhead exhibited by prior fusion techniques.

3.1 Auto-Fusion

This method encourages the model to extract intermodal features by maximizing the correlation between multimodal inputs. In this method, we first concatenate individual unimodal features and then pass them through a transformation layer to get a
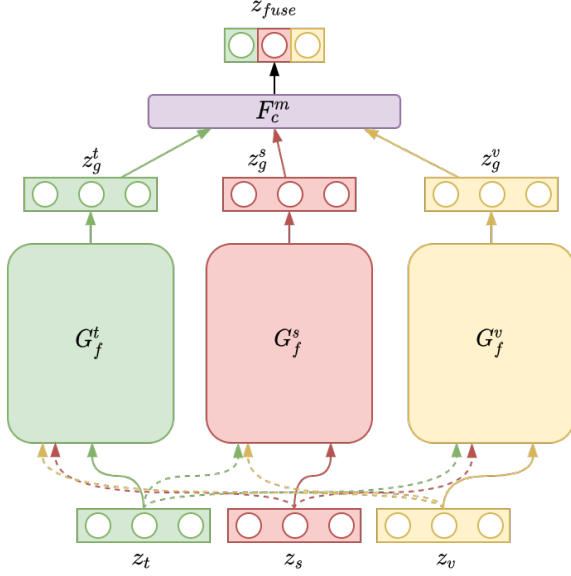
Figure 2: Overall architecture of the GAN-Fusion module. Here, the solid and dashed lines at the bottom part represent input from the target and the complementary modalities, respectively, for G_f^m, the GAN-Fusion module with target modality $m \in \{t, s, v\}$. Furthermore, F_c^m represents the feed-forward layer that produces the fused multimodal representation z_{fuse}, which is, in turn, fed to the target decoder for generation networks and to the fully-connected network for classification networks.

autofused latent vector. We use appropriate learners for individual modalities (See Section 4). We then try to reconstruct the originally concatenated vector from the autofused latent vector. Finally, we minimize the Euclidean distance between the original and reconstructed concatenated vector. This process ensures that the learned autofused vector does not contain arbitrary signals from the input concatenated latent vector. Furthermore, training the model for a downstream task such as emotion recognition incentivizes it to "compress" information without losing any essential cues. In other words, it increases the correlation between the autofused and the concatenated latent vector. This generic procedure applies to any scenario where multiple features need to be combined. For example, it can even be used to combine the forward and backward hidden states of LSTMs (Hochreiter and Schmidhuber, 1997), instead of pooling methods such as 1D pooling, max pooling, sum pooling or even simple concatenation.

We now discuss the Auto-Fusion network in detail. We pose fusion of multimodal inputs as a com-

pression problem, where we must retain as much information from the individual modalities as possible. Given n (≤ 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 number of dimensions to t. Then, we use z_m^t to reconstruct the originally concatenated vector $\hat{z}_m^k$. Finally, we calculate the loss, J_{tr}, between $\hat{z}_m^k$, and z_m^k. The simplest version of Auto-Fusion network employs the mean squared error (MSE) loss function, which aligns with our motivation to compress multimodal features: filter out the less useful signals. These steps could be followed in Figure 1(a) and the MSE loss for Auto-Fusion network is given by:

$$J_{tr} = || \, \hat{z}_m^k - z_m^k \, ||^2 \qquad (1)$$

For Auto-Fusion, we consider the intermediate vector, z_m^t, as the fused multimodal representation.

3.2 GAN-Fusion

In addition to the "staticness" of existing methods, there is also the challenge of distinguishing between ambiguous cases. For instance, the sentence "Your joke blew my mind away, Kevin," could be said in a funny or sarcastic manner. Resolving ambiguity becomes especially important when working on social problems such as hate speech detection. Even when fed with the corresponding speech vector, existing methods cannot effectively distinguish between similar but different emotions such as happiness and calmness. We hypothesize that this is because they do not learn the conditional distribution of sentiment given an utterance (an utterance includes input from all available modalities).

To address this issue, we propose an adversarial training regime that is incentivized to learn the desired conditional distribution. For a task such as emotion recognition, the objective would be sentiment given an utterance. For a more challenging generation task, the model could learn a more complex behaviour, such as the association of different sentences based on how similar they sound and their polarity. Our experiments show that our GAN-based approach is better able to learn such multimodal dynamics compared to other methods.

We now describe GAN-Fusion's architecture in detail for target modality text (denoted by t.) For a given multimodal sample x, we first encode the in-

puts from each modality (speech, visual and text) to get the respective latent vectors, $z_s, z_v,$ and z_t.T Choosing a target modality such as text, we pass z_t (along with random normal noise,) through a generator to obtain $z_g = G(z_t)$ and autofuse 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 Auto-Fusion, and can simply treat the other modality's vector as z_{tr}. Finally, we train the network in an adversarial fashion, labelling z_{tr} as positive samples and z_g as negative samples. The adversarial loss, J_{adv}^t, is given below:

$$\min_{G} \max_{D} J_{adv}^t(D, G) = \mathbb{E}_{x \sim p_{z_{tr}}(x)}[\log D(x)]$$
$$+ \mathbb{E}_{z \sim p_{z_t}(z)}[\log(1 - D(z_g))] \quad (2)$$

Overall, the generator G tries to align features of the target modality with features from the complementary modalities, and the discriminator tries to discern the source of its input. Such a translation between latent vectors has been shown to learn an "intermediate" latent vector denoting their joint representation (Pham et al., 2019; Gao et al., 2019). Learning the latent space in such an adversarial manner induces a clustering effect on the latent space, where texts associated with similar sounds and visuals are grouped together. We conjecture that the adversarial training helps the model learn the relative topology of the complementary modalities' latent space, which improves sampling from the target modality.

We elucidate this effect through the following example. For the sake of simplicity, let us consider only one complimentary modality (video) in this case and let text be the target modality. First, we make a reasonable assumption that video embeddings for, say, Soccer and Golf–falling under the general category of Sports–will be mapped closer to each other and farther from video embeddings from an unrelated topic such as Cooking. When learning the intermediate representation between video and text, the text latent space is constructed such that its relative topology partially reflects the video latent space. So, token embeddings in the text latent space related to videos of similar events (Soccer and Golf) will adopt similar relative positioning as followed by the video embeddings for Soccer and Golf in the video latent space. For multimodal machine translation, if the model is fed with a Golf
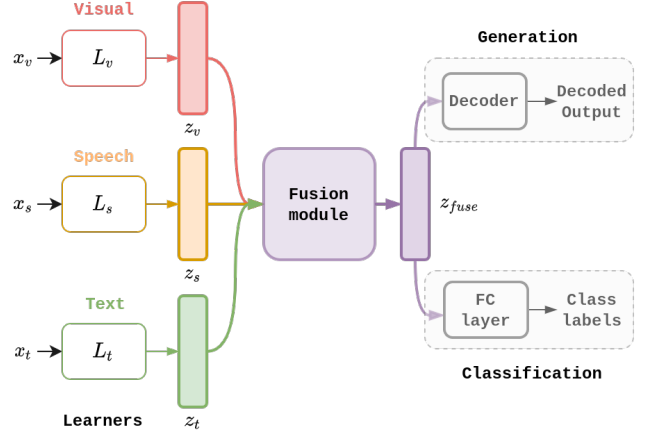

Figure 3: Using proposed fusion techniques for generation/classification. Unimodal inputs x_v, x_s, x_t are passed through their respective learners L_v, L_s, L_t to obtain unimodal representations z_v, z_s, z_t. Here, v, s, t correspond to visual, speech, and textual modalities respectively. The individual unimodal representations are then passed through the fusion module (either Auto-Fusion or GAN-Fusion,) which outputs the fused multimodal representation z_{fuse}. For generation, z_{fuse} is then passed through a compatible decoder, which generates outputs for the desired target modality. For classification, z_{fuse} is passed through a fully-connected layer instead, which predicts appropriate the class labels.

video and the source text as input, it may be better able to sample jargon words for Golf from the text latent space due to this topology inheritance. This ultimately improves translation-quality. This is also depicted in Figure 4.

Figure 1(b) shows GAN-Fusion module for the text modality. The GAN-Fusion module, overall, has one such module for every modality. Total adversarial loss is, therefore, given by:

$$J_{adv} = J_{adv}^t + J_{adv}^s + J_{adv}^v \quad (3)$$

where losses J_{adv}^s and J_{adv}^v for speech and video, respectively, are defined similarly as J_{adv}^t.

Figure 2 shows an overall architecture of the GAN-Fusion module, which consists of $G_f^t, G_f^s,$ and G_f^v, the respective fusion modules for text, speech, and video modalities. We pass the outputs of these modules through a feed-forward layer to obtain the final fused multimodal representation z_{fuse}.

3.3 Overall Architecture

In this section, we describe the end-to-end training process for using the proposed fusion methods for
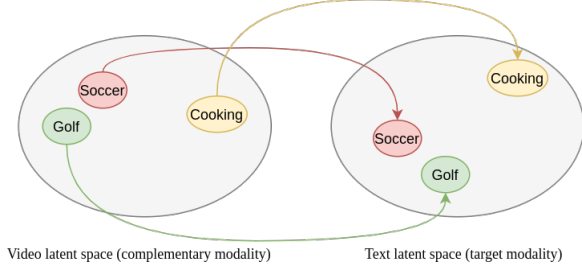
Figure 4: Visualizing the induced clustering effect of GAN-Fusion. The circles represent a cluster of words related to the indicated topic (text inside the circles)

1) Generation tasks (e.g. visual question answering, multimodal machine translation) and 2) Classification tasks (e.g. speech emotion recognition, hate speech detection.)

Generation: Figure 3 shows the end-to-end pipeline integrating proposed fusion techniques for generation. We first pass raw inputs from different modalities through their respective learners to obtain their respective latent representations. They are then passed through the fusion module that outputs a fused representation to be used for decoding. We only validate this process for generating text, but this method could very well be used for generating outputs for different target modalities. Notably, our pipeline for generation looks very similar to a Seq2Seq network. We simply introduce a fusion module between the encoder and the decoder module. It should also be noted that all GAN-Fusion modules for all the target modalities (G_f^t, G_f^s, G_f^v) are trained simultaneously with the rest of the network.

Classification: Figure 3 also shows how to adapt the previously described generation network for classification: simply replace the decoder with a fully-connected layer to predict appropriate class labels.

It is important to note that $z_{fuse} = z_m^t$ for Auto-Fusion, and it is obtained as shown in Figure 2 for GAN-Fusion. The overall loss function for our networks can be generalized as follows:

$$J_{total} = \lambda_1 J_{fusion} + \lambda_2 J_{task} \qquad (4)$$

Here, J_{fusion} refers to the loss function of the fusion network. It equals J_{tr} (from equation 1) when using Auto-Fusion, and J_{adv} (from equation 3) when using GAN-Fusion. Furthermore, J_{task} refers to the task-specific loss, i.e., classification loss (such as max-margin loss) or generation loss (such as cross-entropy loss for Seq2Seq network).

λ_1 and λ_2 are hyperparameters to tune.

4 Experimental Setup

We measure our models' effectiveness on two tasks: 1) multimodal machine translation and 2) multimodal emotion recognition. The subsequent sections describe our complete experimental setup, including datasets and baselines used, implementation details, and evaluation metrics.

4.1 Datasets

We choose three datasets for our experiments, which are described as follows:

IEMOCAP: We use the benchmark Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008) for emotion recognition. We only use the textual and speech modalities for our emotion recognition experiments. The dataset is originally split into multiple utterances for each session, and we further split each utterance file based on the provided start and end timestamps to obtain wav files for each sentence. This results in a total of ~10K audio files, which are then used to extract features for predicting a given utterance's emotion. Concretely, we identify the task as an emotion recognition problem, where, given a sentence and its audio, we aim to infer the correct emotion for that utterance.

How2: We evaluate our models on the multimodal How2 dataset (Sanabria et al., 2018), which comprises of 79,114 instructional videos, their Kaldi (Povey et al., 2011) audio features, and word-level time alignments of English-to-Portuguese translations. The How2 dataset is trimodal in comparison to other multimodal datasets (Matthews et al., 2002; Patterson et al., 2002). This makes it suitable to evaluate the contribution of each modality for different tasks. Further, as a large-scale multilingual dataset, it enables a convenient medium for neural machine translation in our work.
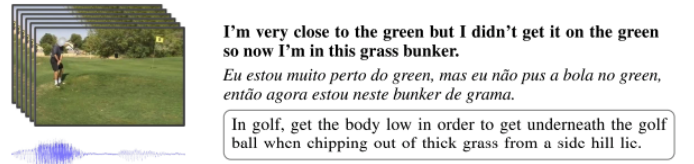

Figure 5: A multimodal sample from the How2 dataset (Sanabria et al., 2018)

Multi30K: In addition to the How2 dataset, we also run experiments on the bimodal Multi30K

Model	Source modalities	BLEU 1	BLEU 2	BLEU 3	BLEU 4
Unimodal S2S	t	-	-	-	54.4
Multimodal S2S	s-v-t	-	-	-	54.4
BPE Multimodal	s-v-t	-	-	-	51.0
Unimodal SPM Transformer	t	-	-	-	55.5
Attention over Image Features	s-v-t	-	-	-	56.2
Seq2Seq (w/o attn)	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	t	79.21	67.34	52.67	47.34
Auto-Fusion (Ours)	s-t	80.34	67.83	61.27	55.01
	s-v-t	85.23	71.95	69.54	57.80
GAN-Fusion (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. Here, 'attn' refers to the word-level attention (Luong et al., 2015).

Model	BLEU 4	Meteor
Unimodal Seq2Seq	36.3	56.9
MeMAD submission	44.1	**64.3**
Auto-Fusion (Ours)	42.31	61.7
GAN-Fusion (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 Unimodal Seq2Seq.

Model	P	R	F	A
LSTM (t)	53.2	40.6	43.4	43.6
LSTM ([s;t])	66.1	65.0	64.7	64.2
MDRE	-	-	-	71.8
MHA-2	-	-	-	76.5
Auto-Fusion (Ours)	75.3	77.4	76.3	77.8
GAN-Fusion (Ours)	**77.3**	**79.1**	**78.2**	**79.2**

Table 3: Precision (P), Recall (R), F1-score (F), and Accuracy (A) for emotion recognition on IEMOCAP.

dataset (Elliott et al., 2016)–a benchmark dataset for machine translation–extended for German, where each sample has an image, its description in the source language, and its translated version. We only run experiments on the En-Fr version, however.

4.2 Implementation details

Generation: We train the network shown in Figure 3 with Auto-Fusion and GAN-Fusion as fusion modules. We use an LSTM encoder with 256 hidden units as the learner for textual description in our generation experiments, unless otherwise stated. For the How2 dataset, we use the already provided 2048-dimensional feature vectors for video as raw input for the video learner, and we feed the Kaldi speech vectors to the speech learner, which is a simple feed-forward layer in this case. For the Multi30K dataset, we use a pre-trained VGG (Si-

monyan and Zisserman, 2015) to encode images, and we do not have a speech learner as there is no speech input in the dataset. The latent dimension is 100 for the fused vector.

Classification: For the task of speech emotion recognition, we train a multimodal classifier on the IEMOCAP dataset. We use LSTMs with 256 hidden units to encode text. For audio, we first pre-process the raw audio files to obtain a lower-dimensional feature vector and then use LSTMs with 50 hidden units as a learner. We predict emotion labels through a full-connected layer as shown in Figure 3. The latent dimension for the fused vectors is 50 in this case.

When training the GAN-Fusion network, we followed numerous tricks (Goodfellow, 2016) to ensure training stability. A few that helped the most include input normalization, batch normalization, Leaky ReLU activation function (Maas et al., 2013) and Adam optimizer (Kingma and Ba, 2015) for the generator and discriminator networks. The presence of multiple auxiliary losses in our networks also helped.

All the networks in our experiments are implemented in PyTorch (Paszke et al., 2019). To train the different classification networks and generation networks on the Multi30K dataset, we use an Nvidia RTX 2080Ti with 12GB of RAM. However, to train our trimodal networks on the How2 dataset, we use Nvidia P100 with 16 GB RAM to fit video feature vectors in memory.

4.3 Baselines

How2: We use the following baselines for experiments on the How2 dataset:

- **Seq2Seq**: A sequence-to-sequence with attention mechanism (Luong et al., 2015). It

employs the previously described learners for each modality, and early fusion.

- **Unimodal and Multimodal S2S:** Unimodal and Multimodal S2S model as described in Sanabria et al. (2018).

- **How2-challenge baselines:** We use Attention over Image Features (Wu et al., 2019), Unimodal SPM Transformer (Raunak et al., 2019) and BPE Multimodal (Lal et al., 2019) from the How2-challenge leaderboard.[2]

Multi30K: We use the following baselines for the Multi30K dataset:

- **Unimodal Seq2Seq:** A text-only NMT system used by Elliott et al. (2017).

- **MeMAD submission** (Grönroos et al., 2018): The best performing model on Multi30K, a multimodal transformer network.

IEMOCAP: We use the following baselines for the IEMOCAP dataset:

- **LSTM (t):** A unimodal LSTM classifier with attention mechanism trained using only text.

- **LSTM ([s;t]):** A bimodal LSTM classifier with attention mechanism on text only. We use the concatenation of speech and text features as the joint multimodal representation.

- **MDRE:** Multimodal Dual Recurrent Encoder proposed by Yoon et al. (2018).

- **MHA-2:** A multimodal classifier with Multi-hop attention mechanism (Yoon et al., 2019).

We compare the above baselines' performance with our two main models: GAN-Fusion and Auto-Fusion, which replace early fusion in the Seq2Seq baseline. We report the results in Table 1, 2, and 3.

4.4 Evaluation metrics

We use Precision, Recall, F-Score, and classification accuracy to evaluate our classification networks trained for speech emotion recognition. For experiments on the How2 and Multi30K dataset, we use BLEU (Papineni et al., 2002) to evaluate the quality of translated sentences. For How2, we compute BLEU1-BLEU4 scores of the different models
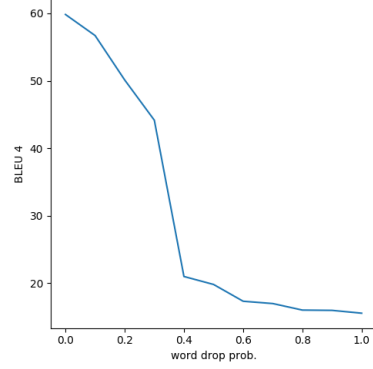

Figure 6: Ablation test on the How2 dataset. Word drop probability v/s BLEU 4. A sudden drop in BLEU scores as we move from 0.3 to 0.4 indicates that our model was able to compensate for $\sim 30\%$ of the missing text.

under consideration. For Multi30K, we also use METEOR (Banerjee and Lavie, 2005), which is a weighted harmonic mean of unigram precision and unigram recall providing a better indication of translation quality. For all the mentioned evaluation metrics, a higher number denotes better performance.

5 Results

Quantitative analysis: Results of our experiments on the How2, Multi30K and IEMOCAP dataset are shown in Tables 1, 2 and 3, respectively. For speech emotion recognition, we observe that our models consistently perform well across all the evaluation metrics. For the relatively challenging multimodal machine translation task, we observe that our model outperforms all existing baselines in terms of BLEU scores. Compared against the best performing baselines, GAN-Fusion improves BLEU4 scores by 3.63 points and 0.13 points on the How2 and Multi30K datasets, respectively. This shows that the fusion module was better able to extract signals from all modalities. On the Multi30K dataset, our model is competitive in terms of METEOR. Such performance becomes more pronounced considering our models have only one attention module, in contrast to multiple self-attention heads in the transformer-based baselines.

The size of our best-performing models (employing GAN-Fusion) is roughly 6M and 11M for classification and generation, respectively, which is significantly lower than traditional transformer models. Fewer trainable parameters reduces the

computational cost and the training time. This also indicates potential by-passing of a mechanism like distillation, which is used to reduce parameters in transformers (Sanh et al., 2019); however, more thorough experimentation is required to reach a concrete conclusion. Our models can also be used in conjunction with transformers, where we utilize the transformers to learn meaningful unimodal feature vectors initially and then employ the proposed fusion methods.

We also perform a comprehensive set of qualitative experiments on the How2 dataset to understand the capability of our fusion techniques. They are described as follows:

Effect of introducing more modalities: To understand the effect of introducing new modalities separately, we perform experiments with different combinations of source modalities, including individual unimodal baselines (Refer to Table 1.) The results reveal that both auditory and visual modalities *always* contribute towards enhanced translation, but the contribution of visual modality is slightly lower (indicated by the lower increase in BLEU scores in Table 1.) This is also consistent with the findings of Grönroos et al. (2018).

Robustness of multimodal features: It is very important for the learned multimodal latent features to be robust, i.e., they should be able to exploit signals from complementary modalities to compensate for the presence of noise in one modality. Therefore, to gauge the robustness of learned multimodal features, we conduct an ablation test on the How2 dataset. We randomly replace some tokens in the test sentence with an <UNK> token and attempt to translate using our best performing model, GAN-Fusion. Figure 6 shows results of our ablation study. We can see that features from the complementary modalities are able to compensate for $\sim 30\%$ of the missing text as we see a sharp drop in BLEU scores beyond that point. This shows that the model does not rely on just the textual description for translation; it also tries to gain a contextual understanding. Hence, it follows that the learned joint representation indeed contains rich information from other modalities.

6 Conclusion and Future Work

In this paper, we propose two adaptive fusion techniques that allow for effective multimodal fusion. Instead of "fixing" a fusion operation a priori, we let the model decide "how" to extract and effec-

tively combine signals from different modalities. Moreover, the joint multimodal representations learned by such models are empirically shown to be robust, which allows the system to maintain good performance even in the absence of some information. Our results indicate that such adaptive models deliver without compromising performance than their more massive counterparts, such as transformers, which is a significant gain.

Our experiments indicate the importance of learning richer unimodal representations. This also suggests that using these methods in conjunction with transformers, which may learn richer unimodal representations, should further improve downstream tasks such as multimodal machine translation and speech emotion recognition. Currently, the attention mechanism is applied only on text. So, another simple way to improve performance would be to introduce visual and acoustic attention mechanisms as well; however, we would still need to address the core problem heterogeneity.

On training GANs: Training GANs is known to be difficult. However, we employed various tricks to ensure the training stability of models, especially the one's employing GAN-Fusion. In fact, a musing line of exploration towards learning a better adaptive model could be to probe the implicit assumptions of GANs themselves. GANs are known to exhibit numerous issues in practice (Arora et al., 2017; Sinn and Rawat, 2018). In Li and Malik (2018), the authors argue the need to return to the principle of maximum likelihood, insisting on full recall.

It is essential to note that much is unknown about these models. More concrete and sound reasoning for the success of these models will rely on two vital components: 1) understanding of the dynamics of the learned latent space, and 2) aligning multimodal features to address heterogeneity. Both these components require more interpretable representations of the otherwise black-box models.

7 Acknowledgments

References

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *ICML*, pages 224–232. PMLR.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

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*, page 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.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*.

Xiang Gao, Sungjin Lee, Yizhe Zhang, Chris Brockett, Michel Galley, Jianfeng Gao, and Bill Dolan. 2019. Jointly optimizing diversity and relevance in neural response generation. In *NAACL-HLT*.

Ian Goodfellow. 2016. NIPS 2016 Tutorial: Generative adversarial networks.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611.

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. In *EMNLP*, pages 250–260.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

Ke Li and Jitendra Malik. 2018. On the implicit assumptions of gans. *NeurIPS Workshop on Critiquing and Correcting Trends in Machine Learning*.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Learning representations from imperfect time series data via tensor rank regularization. In *ACL*, pages 1569–1576.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, pages 2247–2256.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421.

Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Iain Matthews, Timothy F Cootes, J Andrew Bangham, Stephen Cox, and Richard Harvey. 2002. Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):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 *ICML*, pages 689–696.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

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 *ICASSP*, volume 2, pages II–2017.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *AAAI*, volume 33, pages 6892–6899.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE workshop on automatic speech recognition and understanding*.

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. In *Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Yuge Shi, N Siddharth, Brooks Paige, and Philip Torr. 2019. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *NeurIPS*, pages 15718–15729.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Mathieu Sinn and Ambrish Rawat. 2018. Non-parametric estimation of jensen-shannon divergence in generative adversarial network training. In *AISTATS*, pages 642–651.

Nitish Srivastava and Ruslan R Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *NeurIPS*, pages 2222–2230.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NeurIPS*, pages 3104–3112.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *ACL*, volume 2019, page 6558.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, pages 5998–6008.

Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. Speech emotion recognition using multi-hop attention mechanism. In *ICASSP*, pages 2822–2826.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pages 1103–1114.