Adaptive Fusion Techniques for Multimodal Data

Gaurav Sahu, Olga Vechtomova

David R. Cheriton School of Computer Science, University of Waterloo

Objectives

Given inputs from different modalities (e.g. visuals, text, speech), we want to learn a meaningful joint representation to gain a better contextual understanding.

Introduction

A fusion mechanism has two main tasks:

- combining input from different modalities, *and*
- identifying important information, while filtering out the less useful signals from the input.

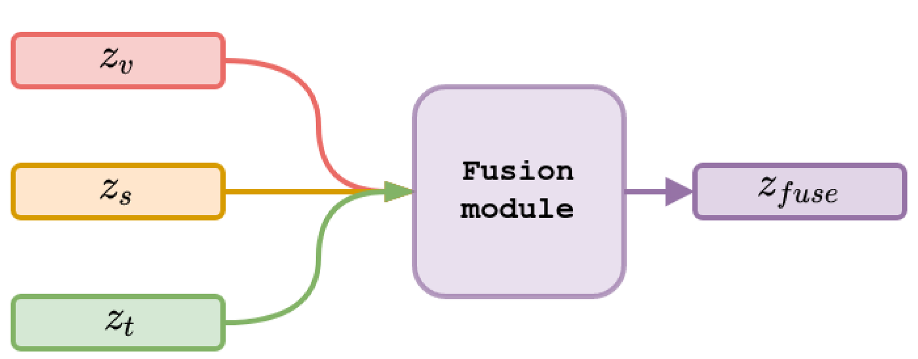


Figure: Fusion process: fusion module combines latent codes from three modalities and outputs a fused vector.

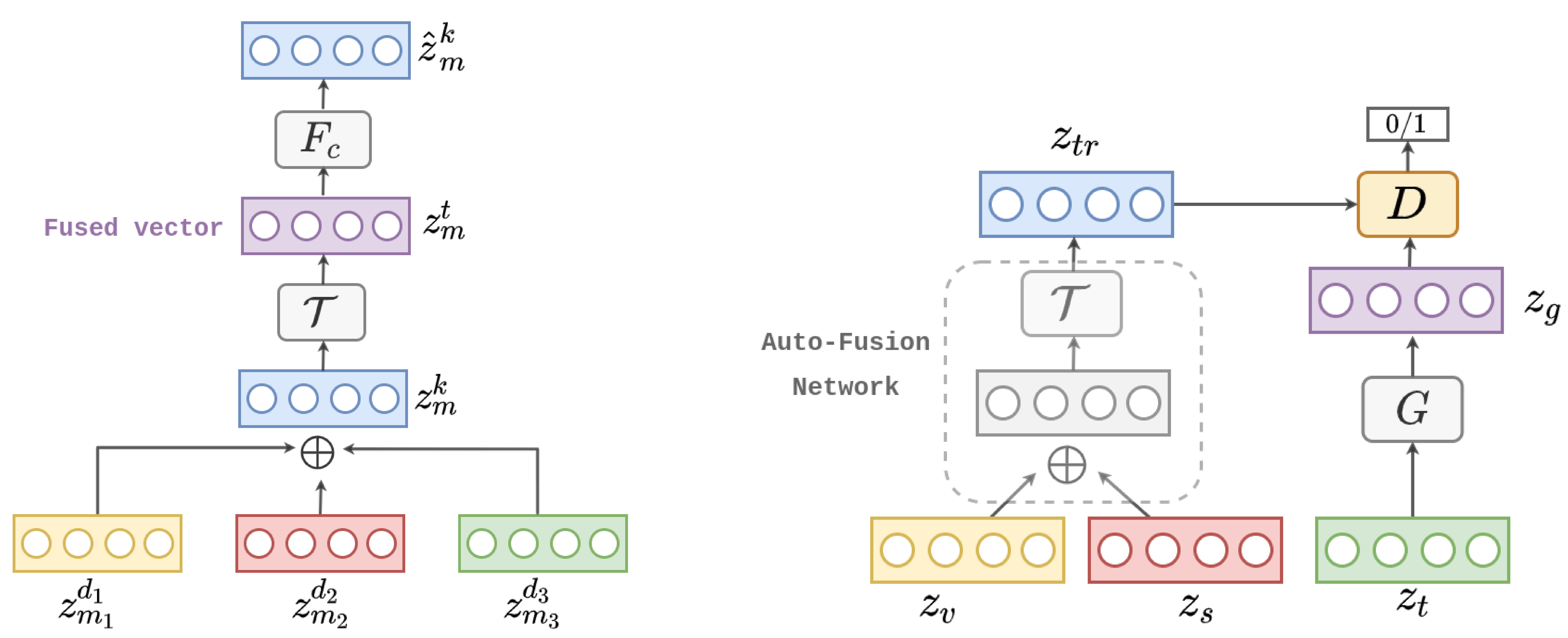
Approach

Proposed two end-to-end trainable fusion methods:

- **Auto-Fusion:** Train an autoencoder model to capture intermodal dynamics by maximizing correlation between multimodal inputs.
- **GAN-Fusion:** Train adversarial networks to align unimodal feature vectors with their complementary modalities. This helps in distinguishing between ambiguous inputs.

Given a multimodal sample with text, visual, and speech input (x_t, x_v, x_s), we first obtain their respective latent representations z_t, z_v, z_s. In GAN-Fusion, we learn aligned latent codes for every mode through an adversarial network. Finally, we combine their outputs to obtain a global fused vector z_{fuse}.

Proposed Fusion Techniques



(a) Auto-Fusion (b) GAN-Fusion

Figure: Proposed fusion methods.

Important Result

Using an adaptive techniques instead of “fixed" methods for fusion improves contextual understanding.

Results

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: 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 [1].

Loss Functions

- **Auto-Fusion:** The MSE loss is given by:

$$J_{tr} = \|\, \hat{\boldsymbol{z}}^{\boldsymbol{k}}_{\boldsymbol{m}} - \boldsymbol{z}^{\boldsymbol{k}}_{\boldsymbol{m}} \,\|^2 \quad (1)$$

- **GAN-Fusion:** Overall adversarial loss: $J_{adv} = J^t_{adv} + J^s_{adv} + J^v_{adv}$ where,

$$\min_G \max_D J^m_{adv}(D, G) = \mathbb{E}_{x\sim p_{z_{tr}}(x)}[\log D(x)] + \mathbb{E}_{\boldsymbol{z}\sim p_{z_m}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]\ \forall\, m \in \{t, v, s\} \quad (2)$$

Conclusion

- We propose two effective fusion strategies for multimodal data
- We make use of adversarial alignment to get a better contextual understanding of a multimodal sample
- Despite being significantly smaller than transformer-based baselines, our model achieves state-of-the-art results.

References

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

Contact Information

- Email: gaurav.sahu@uwaterloo.ca

UNIVERSITY OF WATERLOO