Fusing Multimodal Signals on Hyper-complex Space for Extreme Abstractive Text Summarization (TL;DR) of Scientific Contents

Yash Kumar Atri
yashk@iiitd.ac.in
IIIT Delhi

Vikram Goyal
vikram@iiitd.ac.in
IIIT Delhi

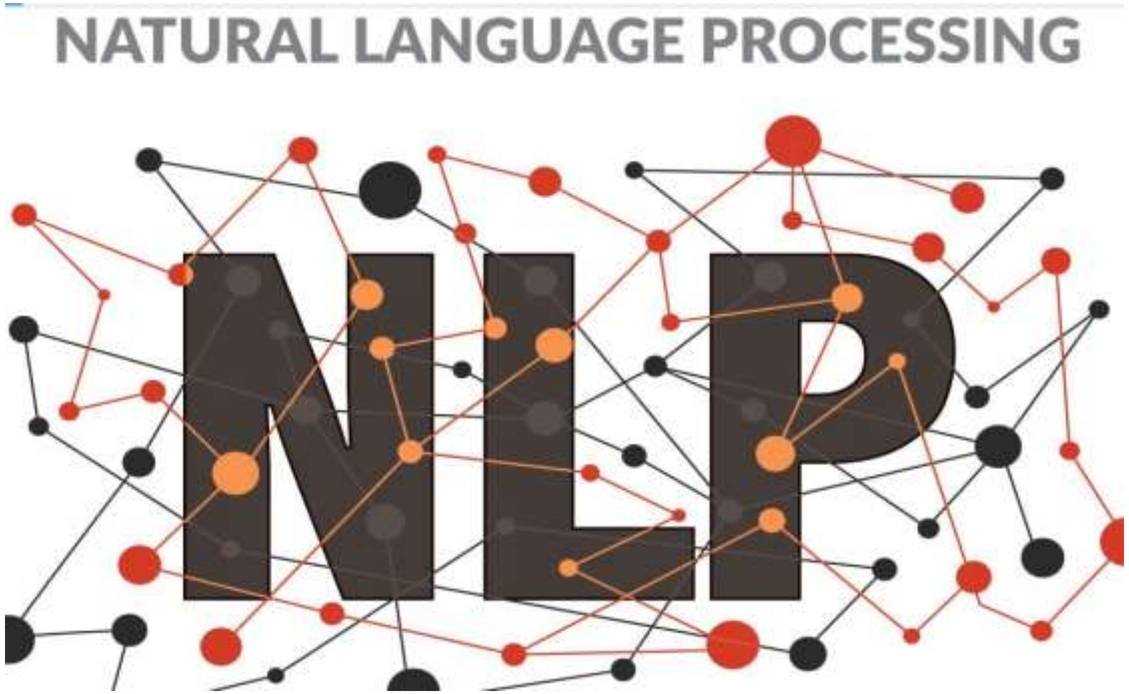
Tanmoy Chakraborty
tanchak@iitd.ac.in
IIT Delhi

https://github.com/LCS2-IIITD/mTLDRgen.

2023.10.29 • ChongQing

2023_KDD

Reported by Jinyuan Zhang



1.Introduction

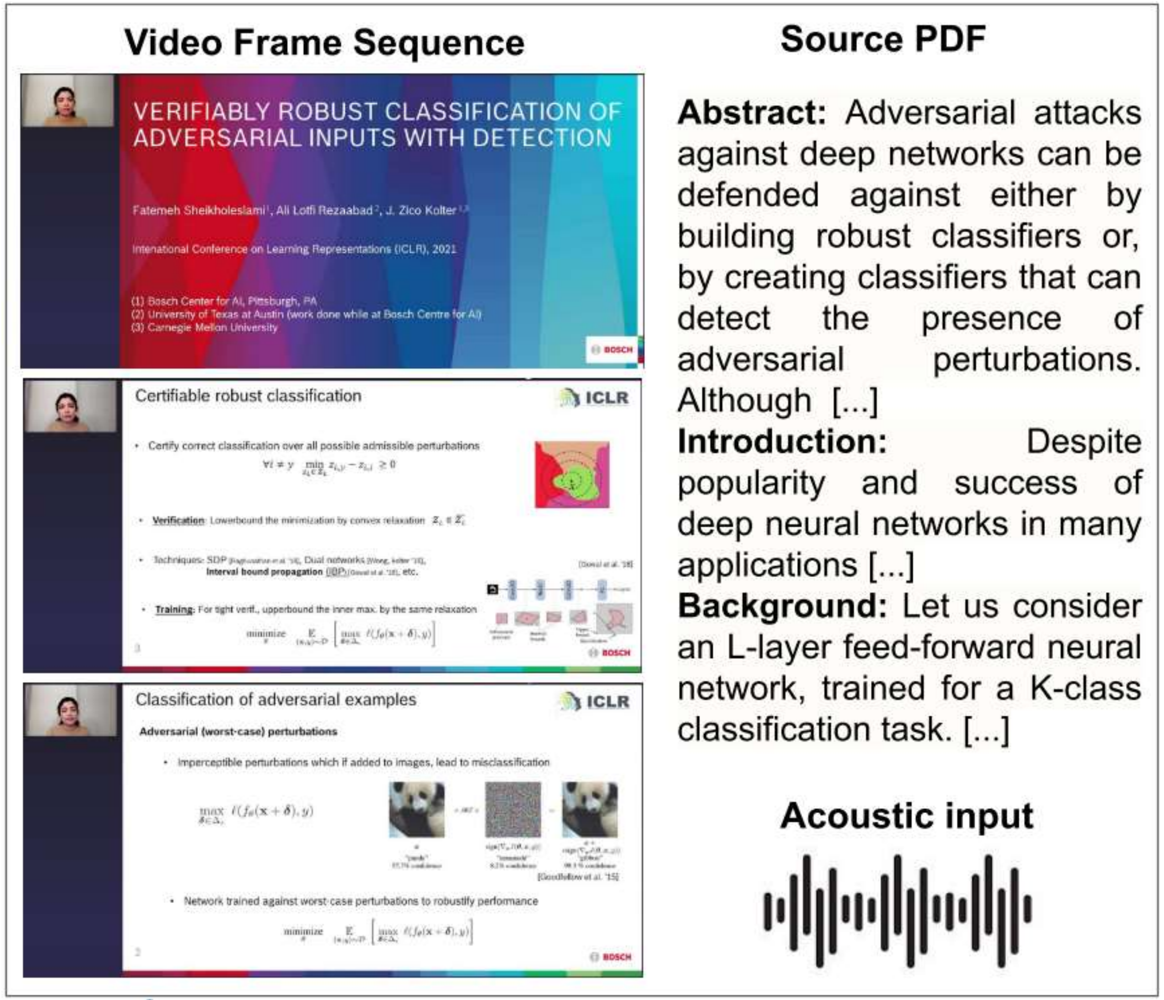
Motivation

3.Method

4.Experiments

gesis
Leibniz-Institut für Sozialwissenschaften

Introduction



· extraction-based summarization

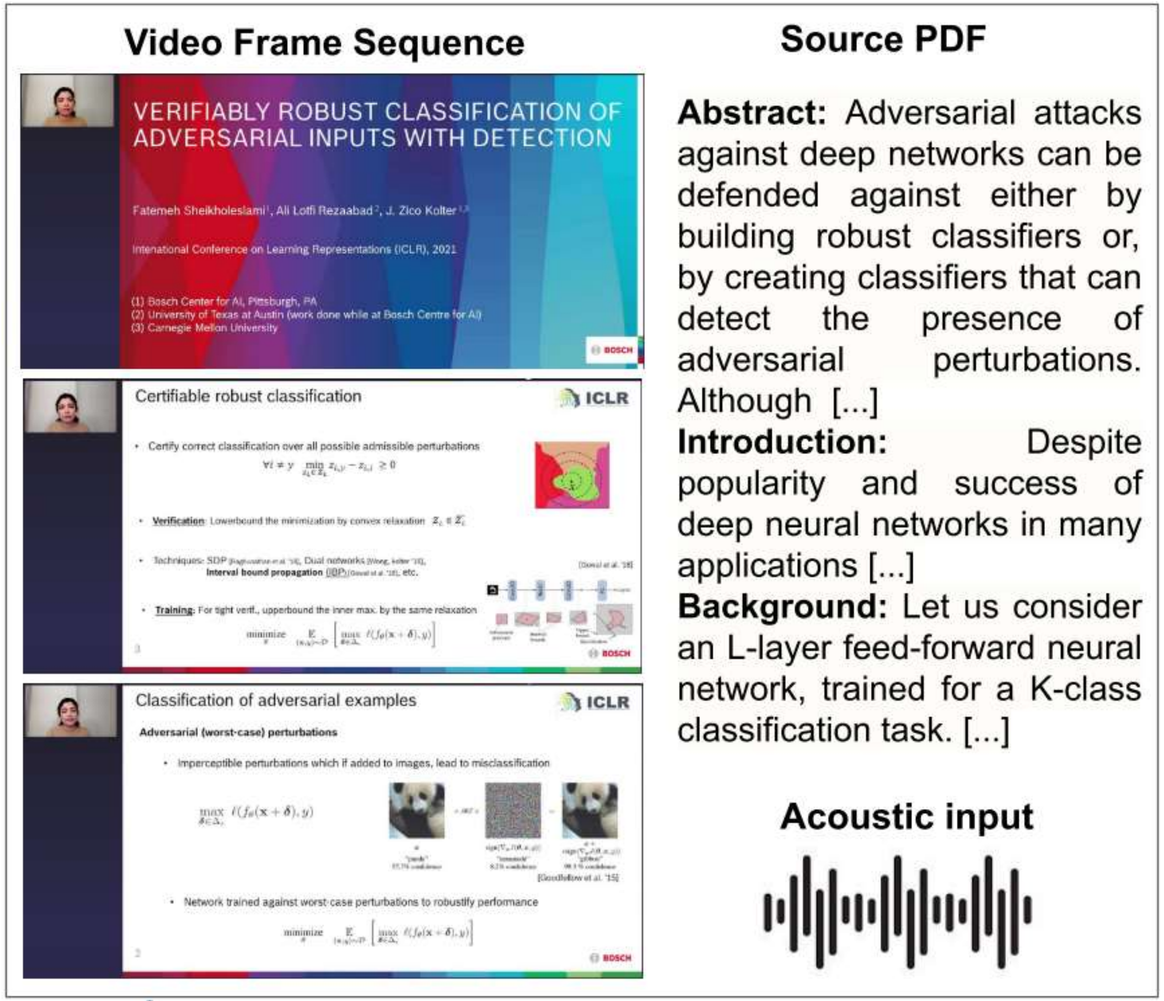
Give each sentence in the original text a binary label (0 or 1), where 0 means that the sentence does not belong to the abstract and 1 means that the sentence belongs to the abstract. The final summary consists of all sentences labeled 1.

· abstraction-based summarization

Generative summary, which attempts to generate a summary by understanding the meaning of the original text.

Based on code-decoding generative digest, text is encoded in semantic vector space, and then word by word digest is generated by decoding network

Motivation



· Limitations

Hard to keep up with the current literature by going through every piece of text in a research article.

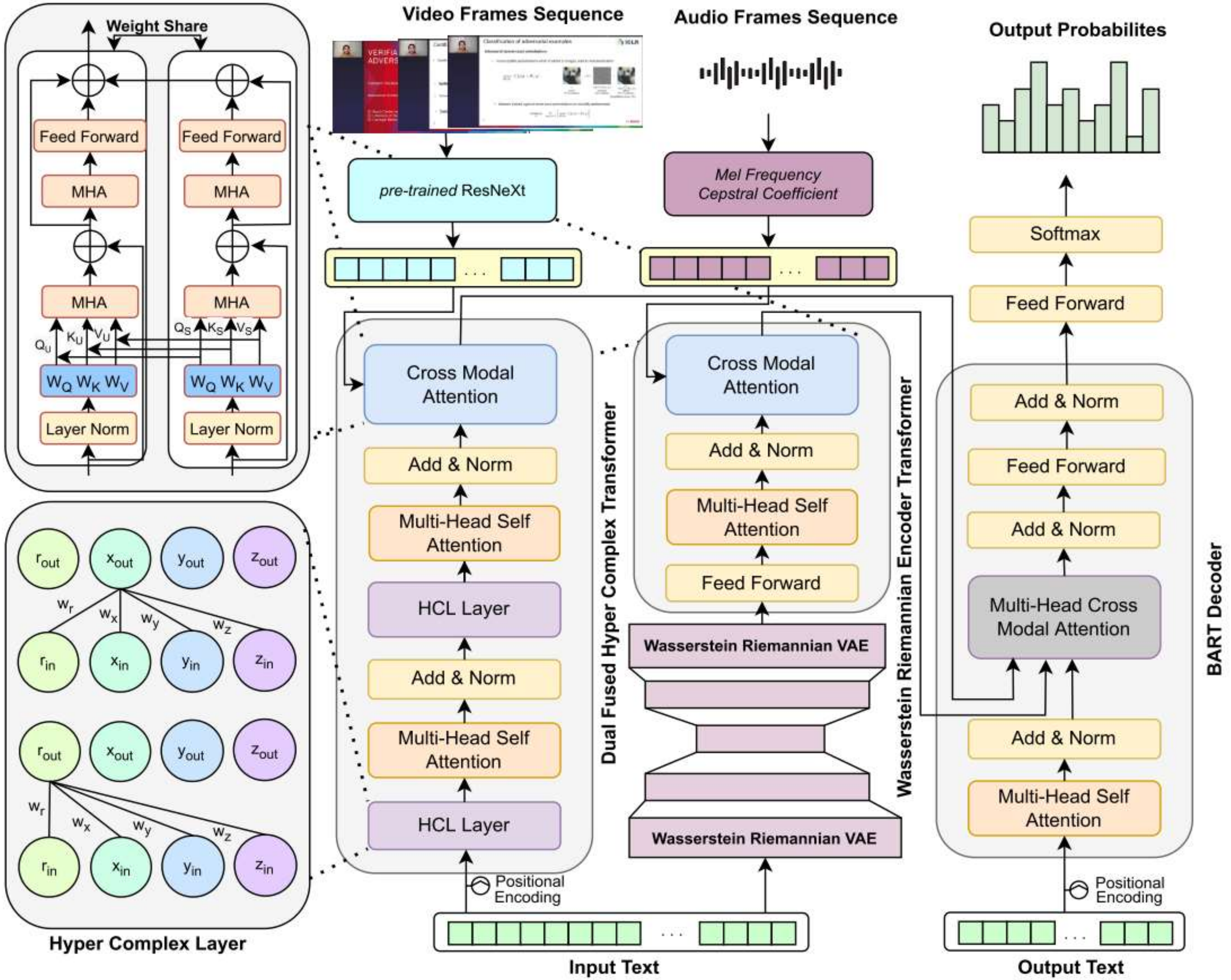
The text alone can not comprehend the entire gist of the research article.

· the current paper offers

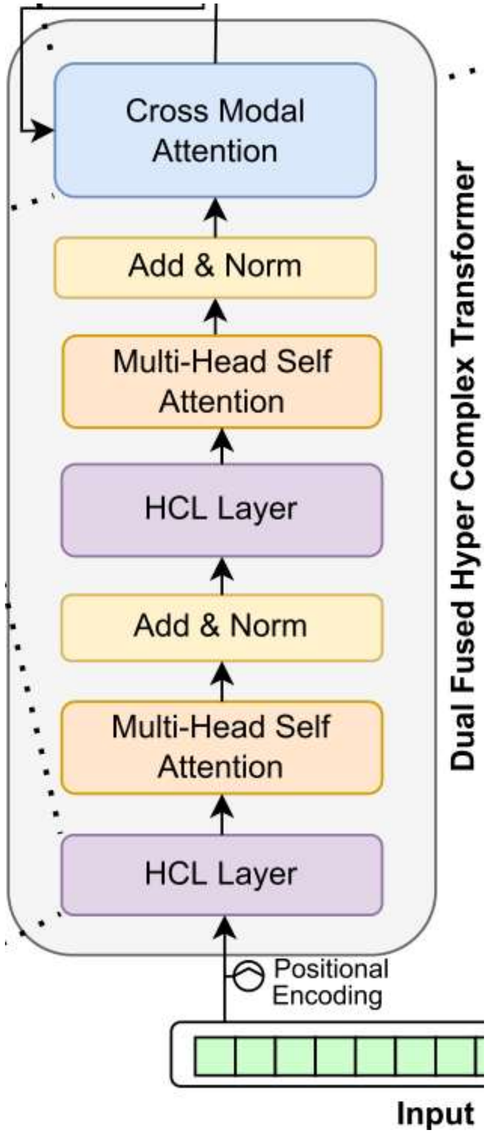
Multimodal information

· the visual modality to capture the visual elements,
· the audio modality to capture the tonal-specific details of the presenter
· the text modality to help the model align all three modalities.

Method


Weight Share
Feed Forward
MHA
MHA
Feed Forward
MHA
MHA
Q_U K_U V_U
Q_S K_S V_S
W_Q W_K W_V
Layer Norm
r_{out} x_{out} y_{out} z_{out}
w_r w_x w_y w_z
r_{in} x_{in} y_{in} z_{in}
Hyper Complex Layer
Video Frames Sequence
pre-trained ResNeXt
Audio Frames Sequence
Mel Frequency Cepstral Coefficient
Output Probabilites
Softmax
Feed Forward
Cross Modal Attention
Add & Norm
Multi-Head Self Attention
HCL Layer
Add & Norm
Multi-Head Self Attention
HCL Layer
Dual Fused Hyper Complex Transformer
Positional Encoding
Input Text
Cross Modal Attention
Add & Norm
Multi-Head Self Attention
Feed Forward
Wasserstein Riemannian VAE
Wasserstein Riemannian VAE
Wasserstein Riemannian Encoder Transformer
Add & Norm
Feed Forward
Add & Norm
Multi-Head Cross Modal Attention
Add & Norm
Multi-Head Self Attention
BART Decoder
Positional Encoding
Output Text

Method



$$HCL(X) = Hx + b \quad (1)$$

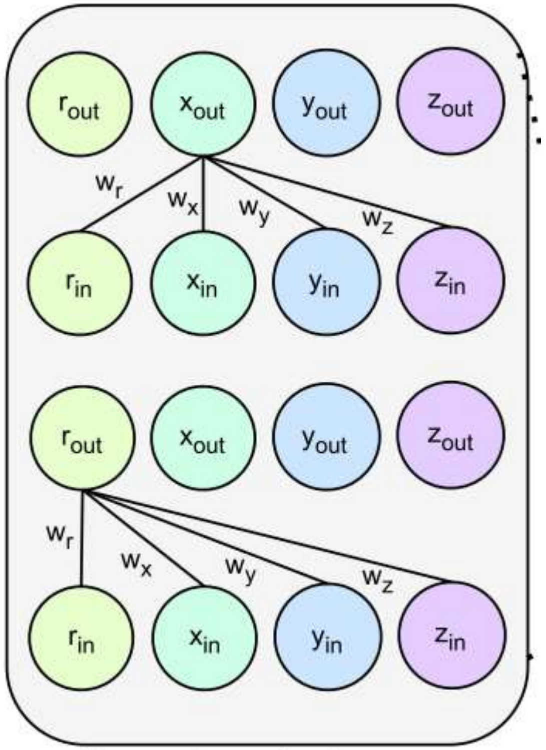
$$H = \sum_{i=1}^{n} P_i \otimes Q_i$$ Kronecker products

$$Q, K, V = \Phi(\mathrm{HCL}(X)), \quad (2)$$

$$A = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \quad (3)$$

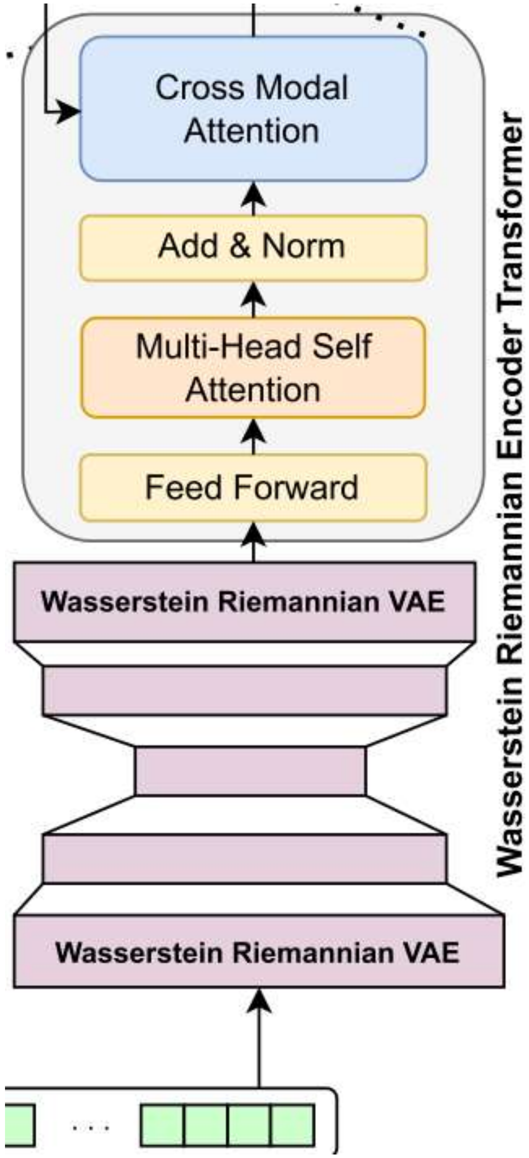
$$X = \mathrm{HCL}([H_1 + \ldots + H_{Num_h}]) \quad (4)$$

$$Y = \mathrm{HCL}(\mathrm{ReLU}(\mathrm{HCL}(X))) \quad (5)$$



Hyper Complex Layer

Method

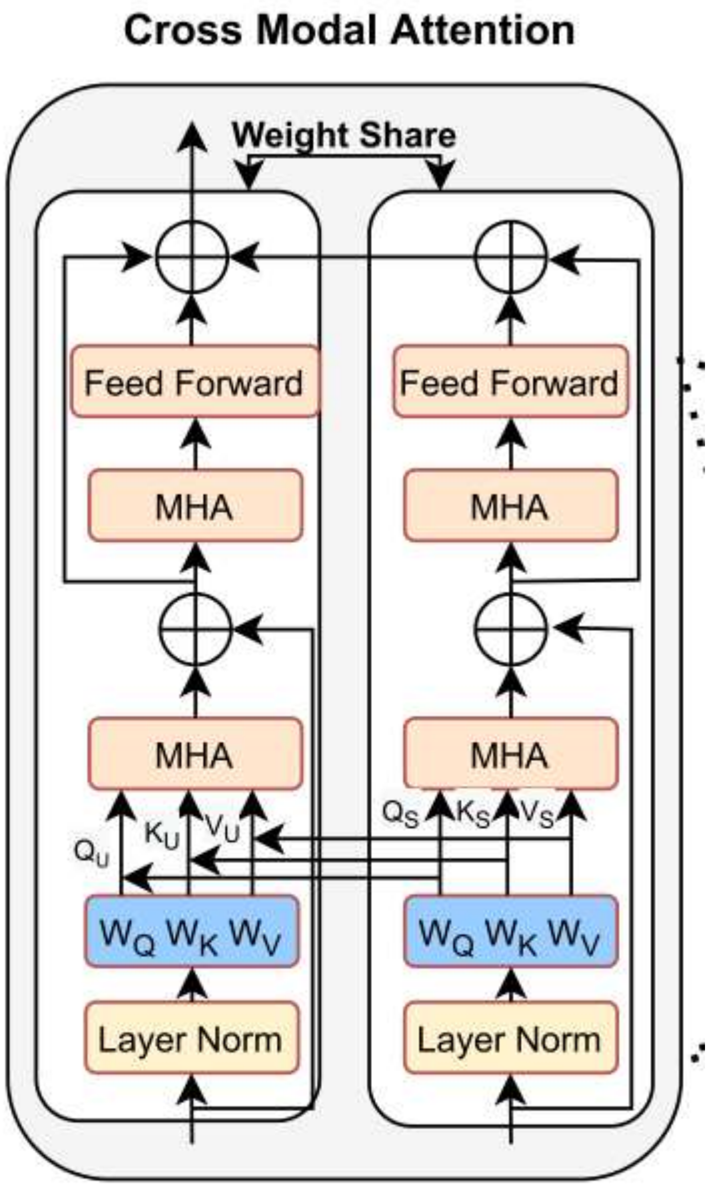


$$Dist(A_X, B_G) = \inf_{Q(Z|X)\in Q} E_{P_X} E_{Q(Z|X)}[c(X, G(Z))] + \lambda MMD(Q_Z, P_Z) \quad (6)$$

$$MMD_k(P_Z, Q_Z) = \left|\left| \int_{\mathcal{Z}} k(z,\cdot)dP_Z - \int_{\mathcal{Z}} k(z,\cdot)dQ_Z \right|\right| \quad (7)$$

$$Dist(A_X, B_G) = \inf_{Q(Z|X)\in Q} P_X Q(Z|X)[c(X, G(Z'))] + \lambda MMD(Q_{Z'}, P_{Z'}) + \alpha(KLD(q(z|x)||p(z)) - \sum \log|det \frac{\partial f'}{\partial z}|) \quad (8)$$

Method



$$Q = Z_t W_q;\ K = Z_v W_k;\ V = Z_v W_v$$

$$E_s = \mathrm{softmax}\left(\frac{X_\alpha W_{Q_\alpha} W_{K_\beta}^\top X_\beta^\top}{\sqrt{d_k}}\right) X_\beta W_{V_\beta} \qquad (9)$$

Experiment

Type	System	mTLDR					How2				
		R1	R2	RL	BERTSc.	FEQA	R1	R2	RL	BERTSc.	FEQA
Extr-text	Lead-2	22.82	4.61	15.47	61.27	32.45	43.96	13.31	39.28	71.56	32.28
	LexRank	27.18	6.82	17.22	63.23	34.21	27.93	12.88	16.93	64.52	31.89
	TextRank	27.43	6.86	17.41	63.34	34.29	27.49	12.61	16.71	64.55	31.92
	MMR	29.54	8.19	18.84	64.59	35.67	28.24	13.12	17.86	64.87	31.98
	ICSISumm	31.57	9.52	19.42	65.84	36.14	28.53	13.44	17.93	65.14	32.16
	BERTExtractive	31.52	9.49	19.31	65.83	36.13	27.18	12.47	15.38	63.47	31.67
Abst-text	Seq2Seq	23.54	5.61	15.48	62.47	31.57	55.37	23.08	53.86	76.15	36.48
	PG	23.59	5.78	16.21	62.71	31.84	51.68	22.63	50.29	73.47	35.37
	CopyTransformer	25.63	7.82	18.54	63.11	37.86	52.94	23.25	50.26	73.58	35.43
	Longformer	21.37	6.47	15.12	61.05	32.14	49.24	21.39	47.41	72.39	35.28
	BERT	24.87	8.85	18.33	62.91	31.89	53.74	23.86	48.06	73.45	35.62
	BART	26.13	9.69	19.62	64.24	38.64	53.81	23.89	48.15	73.51	35.68
	T5	25.87	9.24	18.63	64.13	38.45	53.21	22.51	47.48	73.42	35.65
	Pegasus	26.66	9.83	19.26	64.85	36.98	53.87	23.91	48.17	73.61	35.70
Video only	Action features only	26.38	6.47	15.37	62.48	30.41	45.24	24.42	38.47	69.74	31.28
	RNN (Action features)	26.73	6.51	15.75	63.14	31.35	48.27	27.74	46.37	72.32	35.11
Multimodal	HA	29.32	11.84	26.18	67.24	39.37	55.82	38.31	54.96	77.15	38.55
	FLORAL	31.69	13.54	31.55	69.56	41.19	56.84	39.86	56.93	79.84	39.14
	MFFG	33.19	18.88	33.28	71.54	43.13	61.49	44.61	57.21	80.16	41.59
	ICAF	*36.38*	*20.54*	*34.52*	*73.94*	*45.63*	*63.84*	*44.78*	*58.24*	*82.39*	*42.86*
	mTLDRgen	**41.62**	**22.69**	**37.87**	**78.39**	**48.46**	**67.33**	**48.71**	**61.83**	**84.11**	**44.82**
Δ_{mTLDRgen}	BEST	↑ 5.24	↑ 2.15	↑ 3.35	↑ 4.45	↑ 2.83	↑ 3.49	↑ 3.93	↑ 3.59	↑ 1.72	↑ 1.96

Table 3: Ablation study to show the efficacy of each module of mTLDRgen.

Experiment

System	mTLDR				How2			
	Rouge-1	Rouge-2	Rouge-L	BERTScore	Rouge-1	Rouge-2	Rouge-L	BERTScore
Transformer	25.63	7.82	18.54	63.11	52.94	23.25	50.26	73.58
+ DFHC	29.37	11.78	23.19	67.81	57.34	28.71	56.02	77.31
+ WRET	34.52	14.82	26.54	72.06	61.12	36.89	58.1	81.44
+ DFHC & WRET	37.34	18.32	32.49	74.58	64.23	42.61	59.02	82.45
mTLDRgen	41.62	22.69	37.87	78.39	67.33	48.71	61.83	84.11

Modality	Rouge-1	Rouge-2	Rouge-L
Text +Audio	27.46	7.47	19.62
Audio +Video	27.62	7.53	20.11
Text +Video	28.05	7.83	24.49
Text +Audio +Video	41.62	22.69	37.87

Experiment

Table 6: Human evaluation scores over the metrics – Informativeness (Infor.), Fluency, Coherence, and Relevance for the text-based baselines (BART and T5), multimodal baselines (MFFG, FLORAL, and mTLDRgen) on the mTLDRgen and How datasets.

Modality	System	mTLDR				How2			
		Infor.	Fluency	Coherence	Relevance	Infor.	Fluency	Coherence	Relevance
Abstractive-text	BART	2.81	2.51	2.94	2.85	2.34	2.37	2.46	2.54
Abstractive-text	T5	2.78	2.49	2.81	2.74	2.33	2.28	2.43	2.54
Multimodal	FLORAL	3.2	3.03	3.02	3.11	3.13	3.14	3.08	3.13
Multimodal	MFFG	3.21	3.05	3.09	3.11	3.17	3.21	3.04	3.11
Multimodal	mTLDRgen	**3.46**	**3.32**	**3.27**	**3.29**	**3.34**	**3.27**	**3.21**	**3.18**

Thanks!