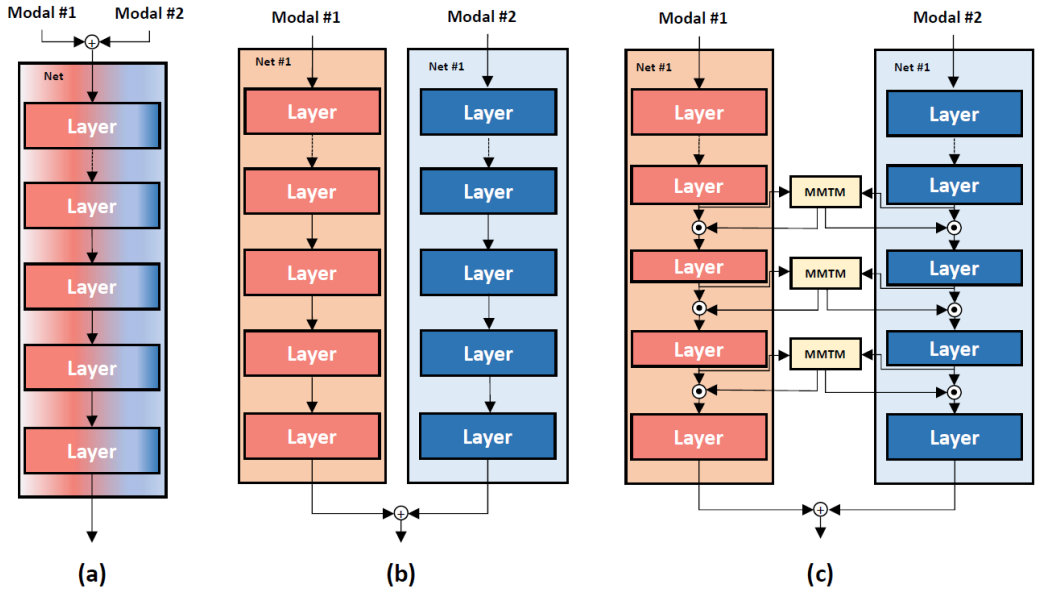
- Joze, H. R. V., Shaban, A., Iuzzolino, M. L., & Koishida, K. (2020). MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 13289-13299).
- Ma, M., Ren, J., Zhao, L., Tulyakov, S., Wu, C., & Peng, X. (2021, March). SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 3, pp. 2302-2310).

Lufei, Gao
Aug.25, 2021

MMTM: Multimodal transfer module for CNN fusion

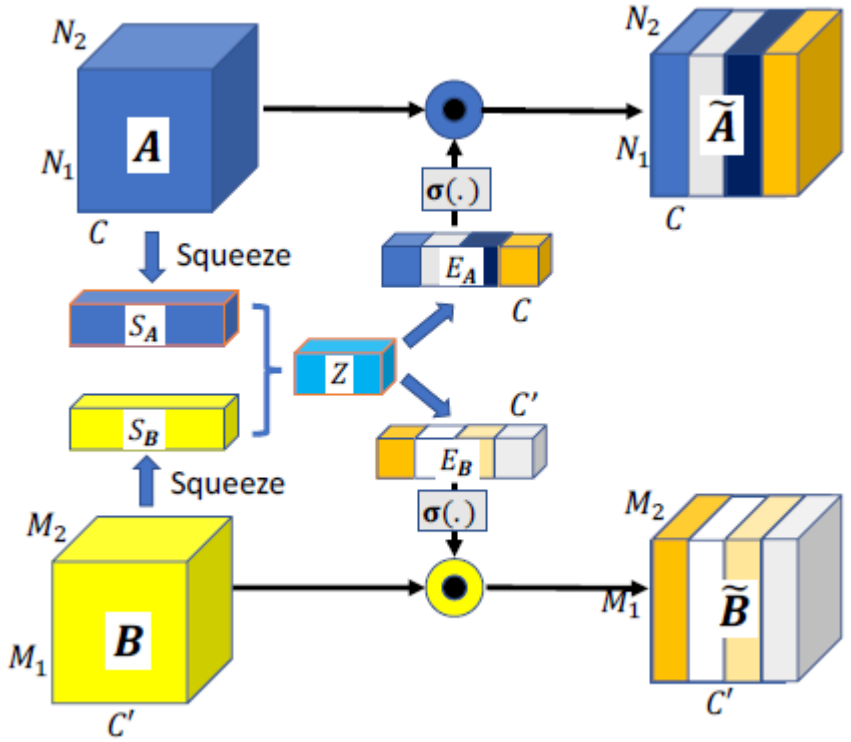
- Multimodal fusion is the act of extracting and combining relevant information from the different modalities that leads to improved performance over using only one modality.
- Fusion can be achieved at the input level (i.e. early fusion), decision level (i.e. late fusion) or intermediately.
- Late Fusion (state-of-the-art): each modality is processed in a separate unimodal CNN stream and the scores are fused at the end.
- Intermediate fusion usually requires major changes in the base network architecture, which complicates the use of pretrained weights.



Multimodal Transfer Module

- Using squeeze and excitation operations to recalibrate the channel-wise features in each CNN stream.
  - A. a multimodal squeeze unit that receives the features from all modalities at a level.
  - B. an excitation unit that uses this joint representation to adaptively emphasize on more important features and suppress less important ones in all modalities.

- Can be added at different levels of the feature hierarchy.
- It also enables learning a joint representation from modalities with different spatial dimensions.
- It could be added among unimodal branches with minimum changes in their network architectures.
- Allowing each branch to be initialized with existing pretrained weights



Multimodal Transfer Module

Squeeze

- C, C': 通道数 $\mathbf{A} \in \mathbb{R}^{N_1 \times \cdots \times N_K \times C}$

  $\mathbf{B} \in \mathbb{R}^{M_1 \times \cdots \times M_L \times C'}$

- （除了C通道，将各个通道内的所有元素进行相加，然后除以总共的个数，即可得到每个通道的信息）

$$S_{\mathbf{A}}(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{n_1,\ldots,n_K} \mathbf{A}(n_1, \ldots, n_K, c) \quad (1)$$

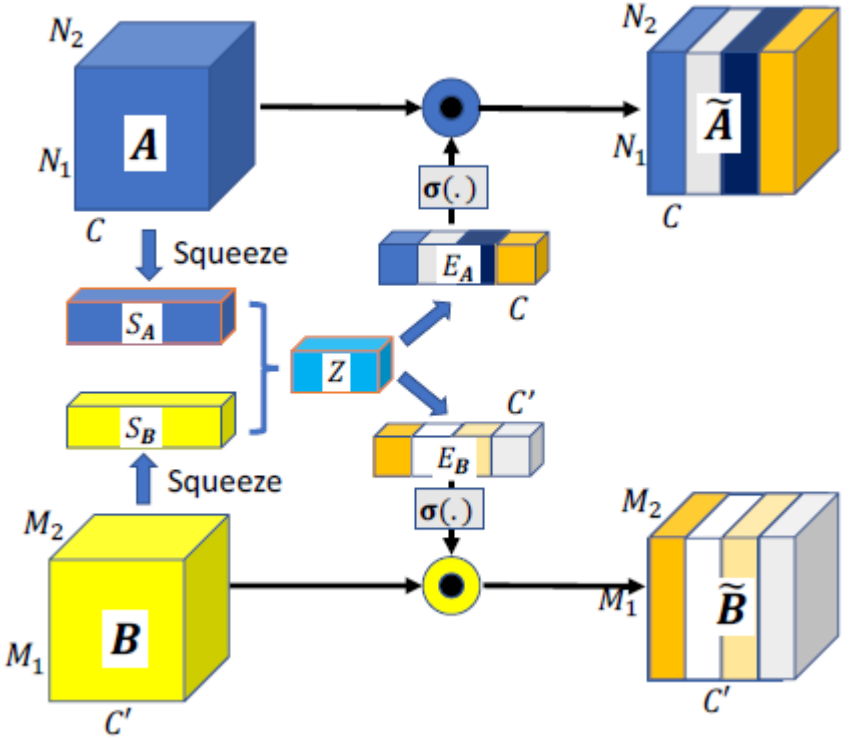
$$S_{\mathbf{B}}(c) = \frac{1}{\prod_{i=1}^{L} M_i} \sum_{m_1,\ldots,m_L} \mathbf{B}(m_1, \ldots, m_L, c). \quad (2)$$

Multimodal Excitation

- S_A, S_B经过concat操作之后再送入一个全连接层得到融合的特征Z
- 然后对于每个模态都通过独立的全连接层分别得到E_A, E_B
- 得到的E_A, E_B经过以一个sigmoid函数得到对应通道的权重后再与原来的特征图相乘
- σ(.)代表sigmoid函数，⊙ 是通道的点乘操作，以此对每个通道进行抑制或激活。

$$Z = \mathbf{W}[S_{\mathbf{A}}, S_{\mathbf{B}}] + b,$$

$$E_{\mathbf{A}} = \mathbf{W}_{\mathbf{A}} Z + b_{\mathbf{A}}, \quad E_{\mathbf{B}} = \mathbf{W}_{\mathbf{B}} Z + b_{\mathbf{B}}.$$



$$C_Z = (C + C')/4$$

限制模型容量，提高泛化能力

MMTM applications

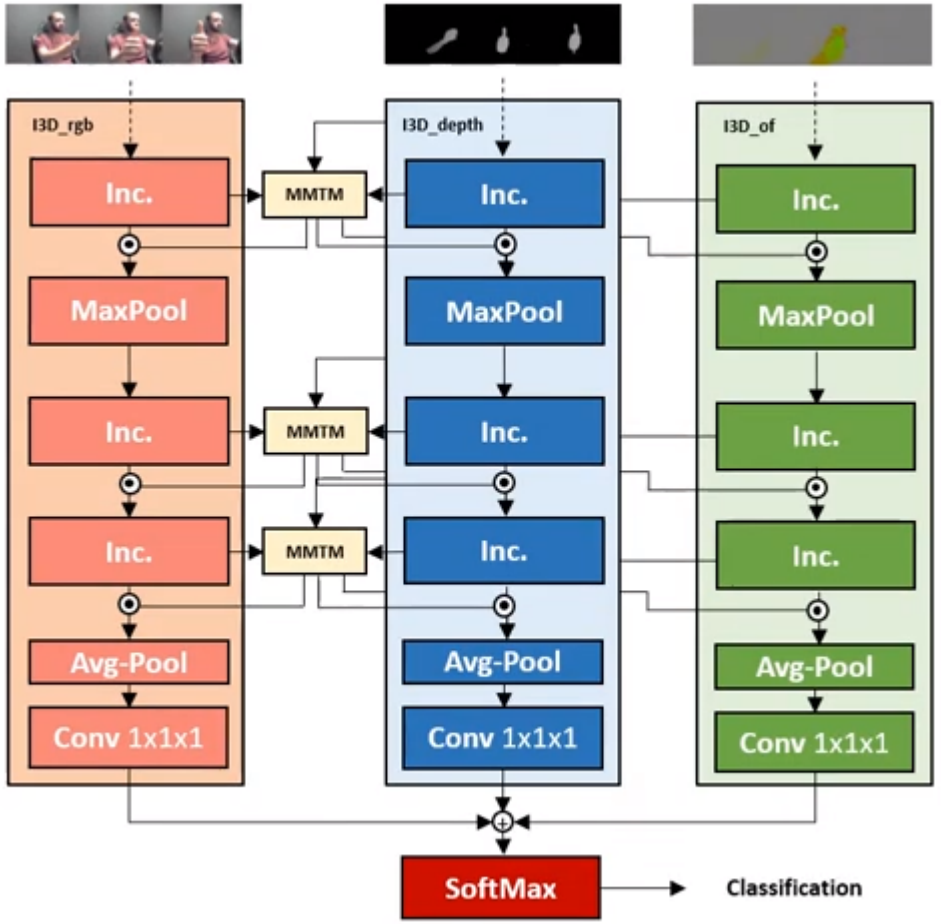
The module design is generic and could potentially be added at any level in the network hierarchy.

1. Dynamic hand gesture recognition
   - Modalities: RGB, optical flow and depth.
2. Audio-Visual Speech Enhancement
   - RGB and sound waveform.
3. Action Recognition
   - RGB and body joints.

Visual RGB
Depth
Optical Flow
Sound Waveform
Body joint key-points

Dynamic Hand Gesture Recognition

- It is shown that complementary sensory information, such as depth and optical flow, improves the performance of the gesture recognition.
- Design a gesture recognition network for fusing b video streams via MMTM
- To process the temporal inputs, use I3D network architecture with an inflated inception-v1 backbone for all the streams.
- Empirically find that the best performance is achieved when the squeeze operation is applied over all the dimensions except for the channel dimension.



Dynamic Hand Gesture Recognition

Method	Input Modalities	Accuracy
I3D [48]	RGB	90.33
I3D [48]	Depth	89.47
VGG16 [58]	RGB+Depth	66.5
VGG16 + LSTM [59]	RGB+Depth	81.4
C3D [60]	RGB+Depth	89.7
C3D+LSTM+RSTTM [41]	RGB+Depth	92.2
I3D late fusion [48]	RGB+Depth	92.78
Ours	RGB+Depth	**93.51**

Table 1. Accuracies of different multimodal fusion hand gesture methods on the EgoGesture dataset [41].

Method	Input Modalities	Accuracy
I3D [48]	RGB	78.42
I3D [48]	Opt. flow	83.19
I3D [48]	Depth	82.28
HOG+HOG2 [64]	RGB+Depth	36.9
I3D late fusion [48]	RGB+Depth	84.43
Ours	RGB+Depth	**86.31**
Two Stream CNNs [14]	RGB+Opt. flow	65.6
iDT [62]	RGB+Opt. flow	73.4
R3DCNN [37]	RGB+Opt. flow	79.3
MFFs [44]	RGB+Opt. flow	84.7
I3D late fusion [48]	RGB+Opt. flow	84.43
Ours	RGB+Opt. flow	**84.85**
R3DCNN [37]	RGB+Depth+Opt. flow	83.8
I3D late fusion [48]	RGB+Depth+Opt. flow	85.68
Ours	RGB+Depth+Opt. flow	**86.93**
Human [37]		88.4

Table 2. Accuracies of different multimodal fusion hand gesture methods on the NVGesture dataset [37].

Audio-Visual Speech Enhancement

- The predominant method for AV speech enhancement combines audio and visual signals via channel-wise con-catenation (CWC) using the late fusion approach.
- Explore AV fusion for speech enhancement tasks using MMTM instead of the CWC-based late fusion
- Use a 2D ResNet-18 for visual network and an autoencoder with skip connections for audio network.

Method	Fusion Method	PESQ	STOI
Target	-	4.64	1.000
Mixed	-	2.19	0.900
AVSE [6]†	CWC	2.59	0.650
AO Baseline	-	2.43	0.930
AV Baseline	CWC	2.67	0.938
Ours	MMTM	**2.73**	**0.941**

Table 3. Speech enhancement evaluations on the *VoxCeleb2* dataset [54] for 3 simultaneous speakers. CWC: Channel-wise concatenation. † for approximate reference only.

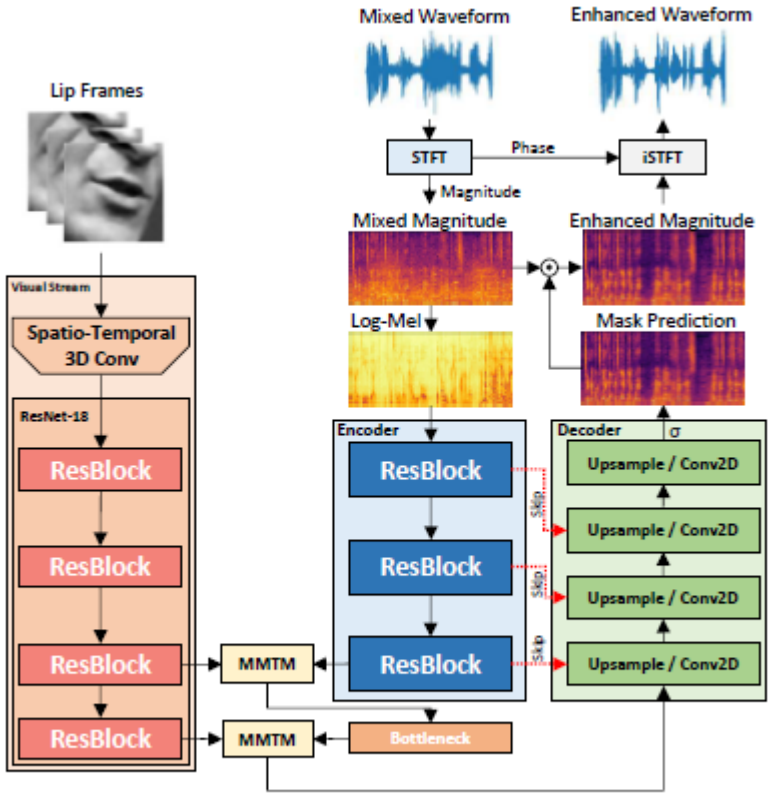
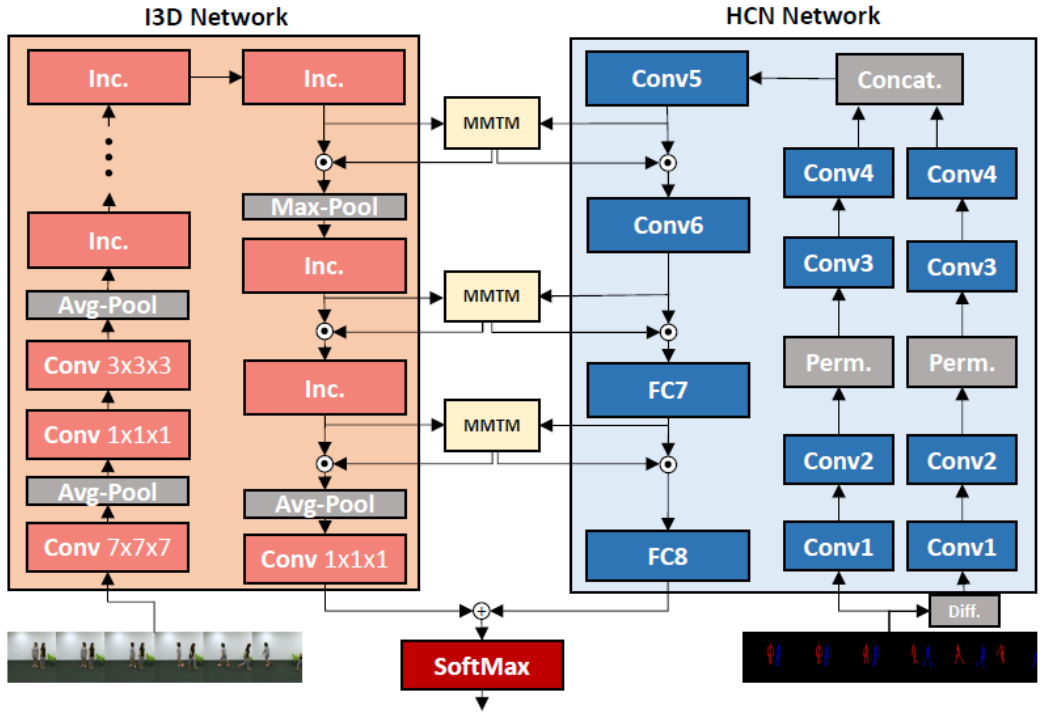


Figure 3. An overview of our AVSE architecture.

Action Recognition

- Utilize MMTM for intermediate fusion between a visual and a skeleton based network.
- Use I3D for the RGB video stream and HCN for
- Add 3 MMTMs that receive inputs from last three inception modules of the I3D and last 3 layers of the skeletal stream.

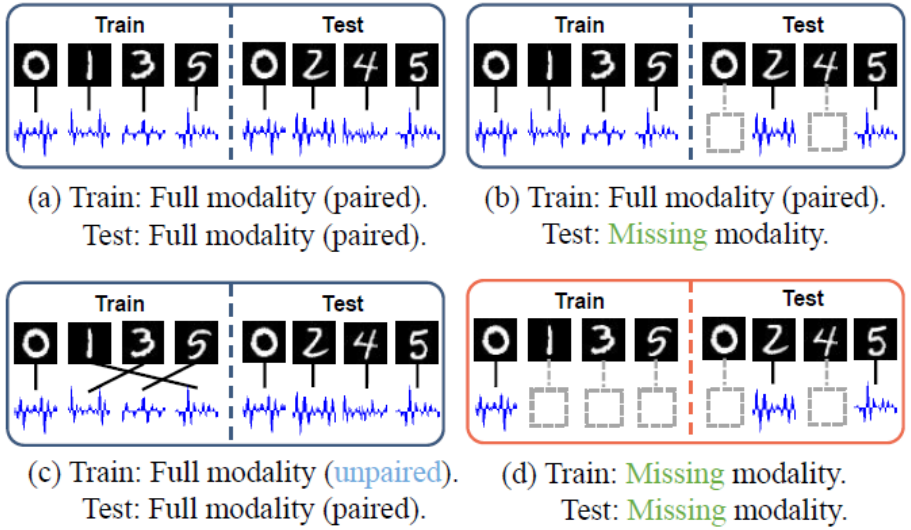


Method	Input Modalities	Accuracy
HCN_{ours}	Pose	77.96
I3D [48]	RGB	89.25
DSSCA - SSLM [75]	RGB+Pose	74.86
Bilinear Learning [29]	RGB+Pose	83.0
2D/3D Multitask [28]	RGB+Pose	85.5
PoseMap [11]	RGB+Pose	91.71
Late Fusion (I3D + HCN_{ours})	RGB+Pose	91.56
Ours	RGB+Pose	**91.99**

Table 4. Accuracies of different multimodal fusion action recognition methods on the NTU-RGBD dataset [55].

SMIL: Multimodal learning with severely missing modality

- Problem: Severely Missing Modality Learning (SMIL) (90%的训练样本包含不完整模态)
- 从灵活性（训练、测试或两者兼有）和效率（大多数训练数据都有不完全模态）两个方面研究了缺失模态的多模态学习.
- SMIL：利用贝叶斯元学习统一实现了两个目标

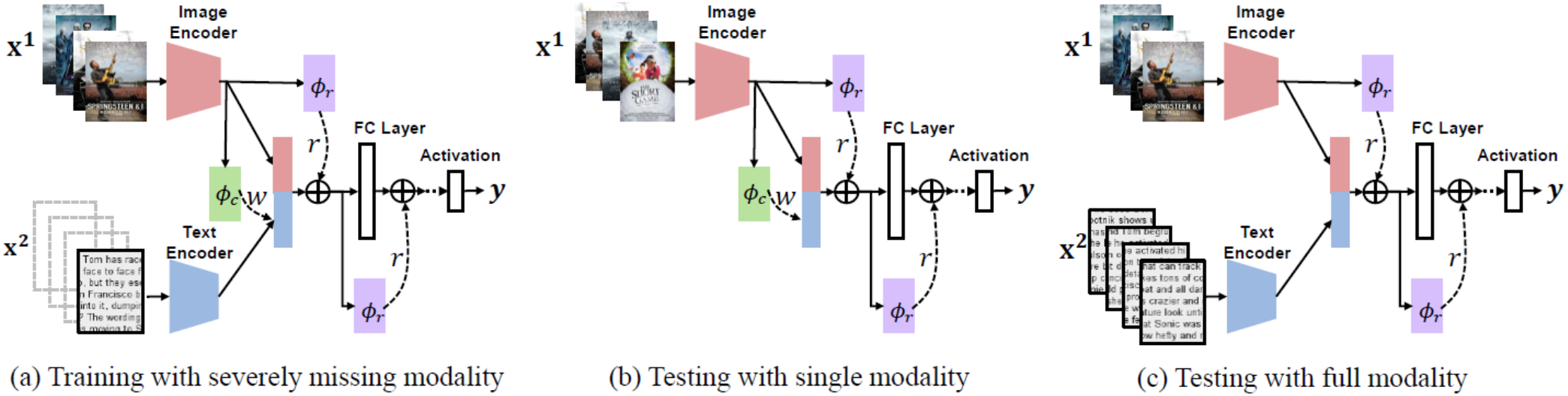


(a) Train: Full modality (paired). Test: Full modality (paired).
(b) Train: Full modality (paired). Test: Missing modality.
(c) Train: Full modality (unpaired). Test: Full modality (paired).
(d) Train: Missing modality. Test: Missing modality.

(a)完整和配对模态的训练和测试(Ngiam et al., 2011);
(b)缺失模态的测试(Tsai et al., 2019);
(c)无配对模式培训(Shi et al., 2020);
(d)训练、测试或两者中严重缺失模态的最具挑战性的配置。

Proposed method

- Multimodal-dataset: $D = \{D^f, D^m\}$
  - $D^f = \{x_i^1, x_i^2, y_i\}_i$: modality-complete dataset. $D^m = \{x_j^1, y_j\}_j$: modality-incomplete dataset.
- Target: to leverage both modality-complete and modality-incomplete data for model training
- Two perspectives:
  - Flexibility: how to uniformly handle missing modality in training, testing, or both?
  - Efficiency: how to improve training efficiency when major data suffers from missing modality?

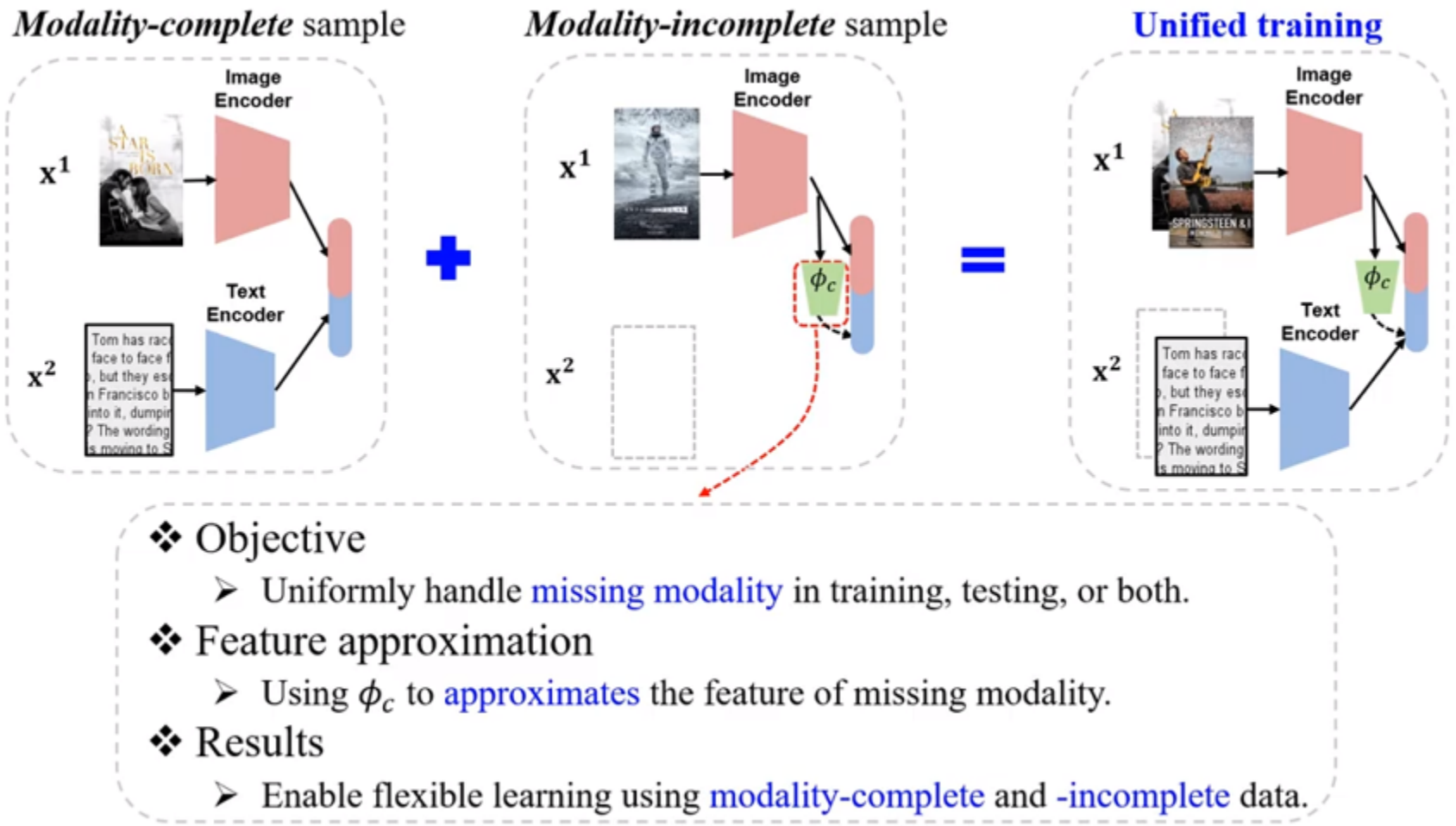


(a) Training with severely missing modality

(b) Testing with single modality

(c) Testing with full modality

To Address Flexibility



- ❖ Objective
  - ➢ Uniformly handle missing modality in training, testing, or both.
- ❖ Feature approximation
  - ➢ Using ϕ_c to approximates the feature of missing modality.
- ❖ Results
  - ➢ Enable flexible learning using modality-complete and -incomplete data.

特征重构网络：利用可用的模态高效地生成缺失模态特征的近似

Missing Modality Reconstruction ϕ_c

Given the observed modality x^1, in order to obtain the reconstruction x^2 of the missing modality, optimize the following objective for the reconstruction network:

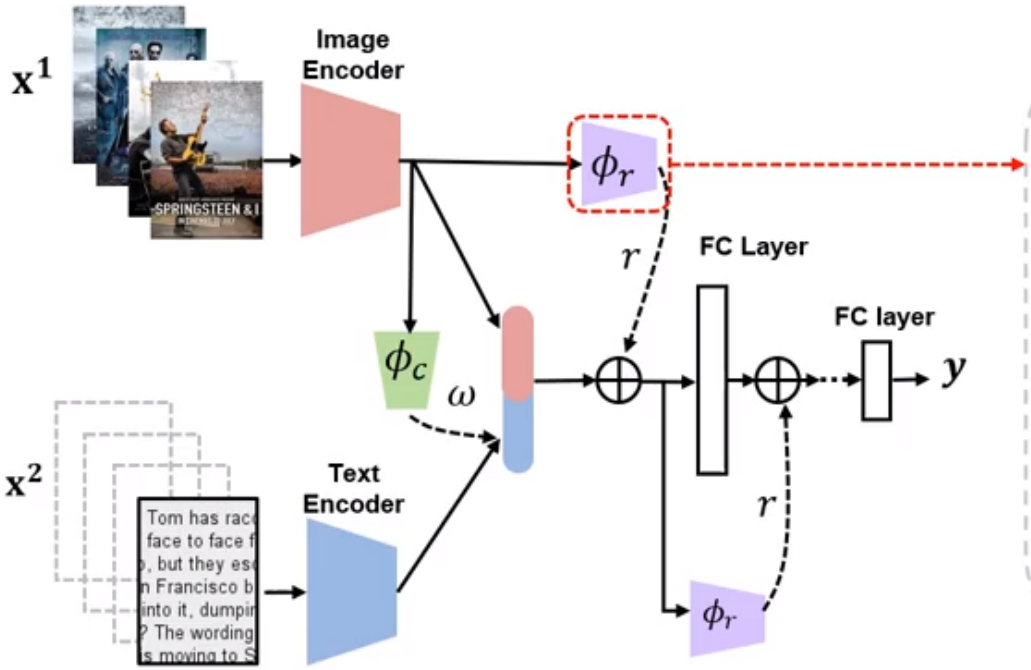
$$\phi_c^* = \arg\min_{\phi_c} \mathbf{E}_{p(\hat{\mathbf{x}}^1, \mathbf{x}^2)}(-\log p(\hat{\mathbf{x}}^2|\mathbf{x}^1; \phi_c)).$$

Approximate the missing modality using a **weighted sum of modality priors** learned from the modality-complete dataset.
ϕ_c are trained to predict weights of the priors:

- learning **a set of modality priors $\mathcal{M}$** which can be clustered among all modality-complete samples using K-means or PCA.
- let ω represent the weights assigned to each modality prior.
- model ω as a multivariate Gaussian with fixed means and changeable variances as $\mathcal{N}(I, \sigma)$. $\sigma = f_{\phi_c}(x^1)$.
- Given the weights ω, reconstruct the missing modality $\hat{x}^2$ by calculating the weighted sum of the modality priors.
- reconstructed missing modality can be achieved by:

$$\hat{\mathbf{x}}^2 = \langle \boldsymbol{\omega}, \mathcal{M} \rangle, \text{where } \boldsymbol{\omega} \sim \mathcal{N}(\mathbf{I}, \boldsymbol{\sigma}).$$

To Address Efficiency

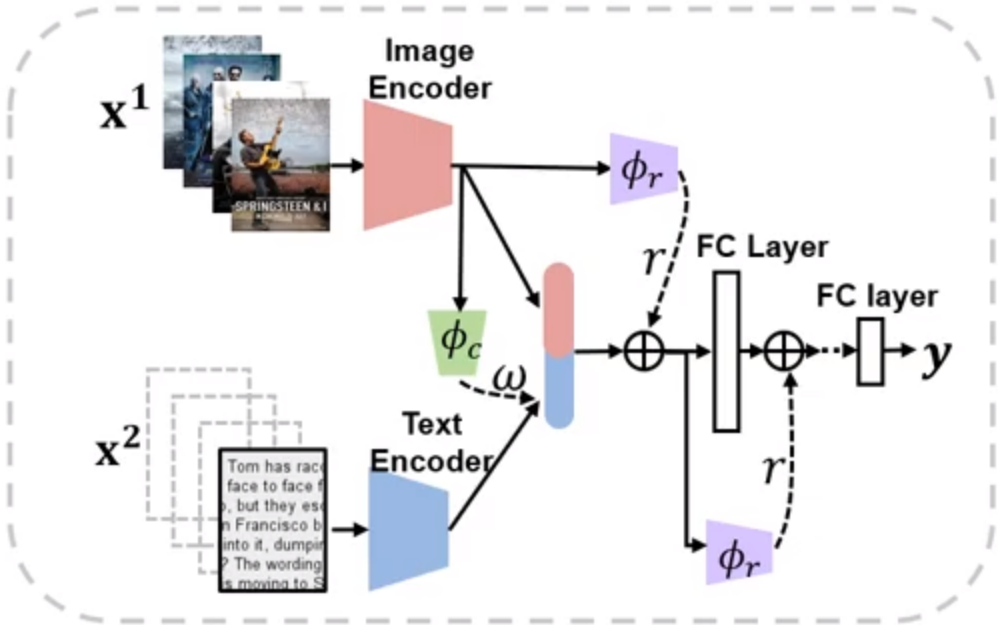


- ❖ Objective
  - ➢ Regularize the feature to avoid biased and low-quality approximation.
- ❖ Feature regularization
  - ➢ Using ϕ_r to regularize the latent features.
- ❖ Results

$$\mathbf{r} = f_{\phi_r}(\mathbf{h}^{l-1})$$ h^l: latent feature of l-th layer.

$$\mathbf{h}^l := \mathbf{h}^l \circ \text{Softplus}(\mathbf{r}), \text{ where } \mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$

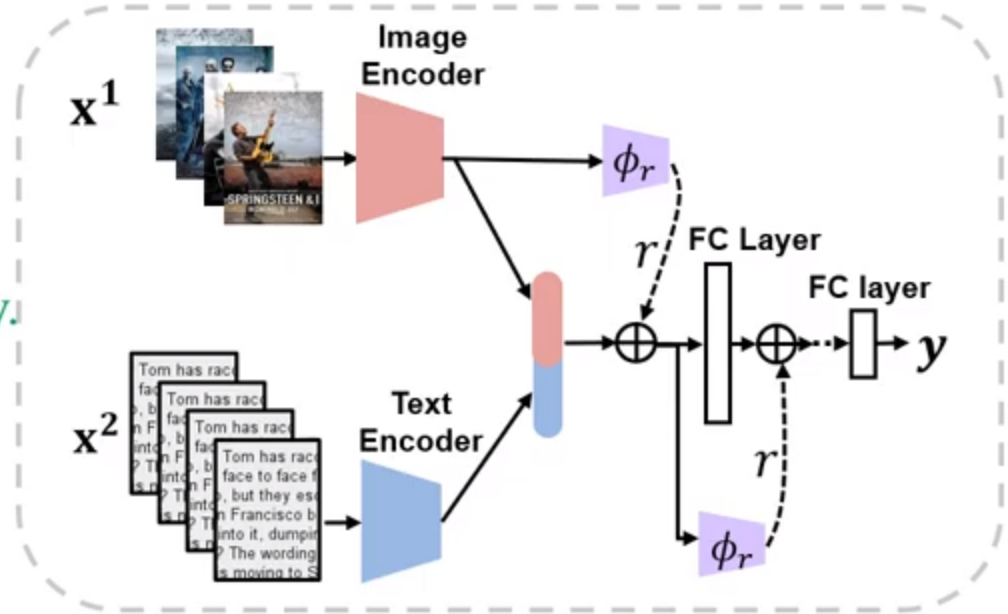
Overall Framework: training & testing setup



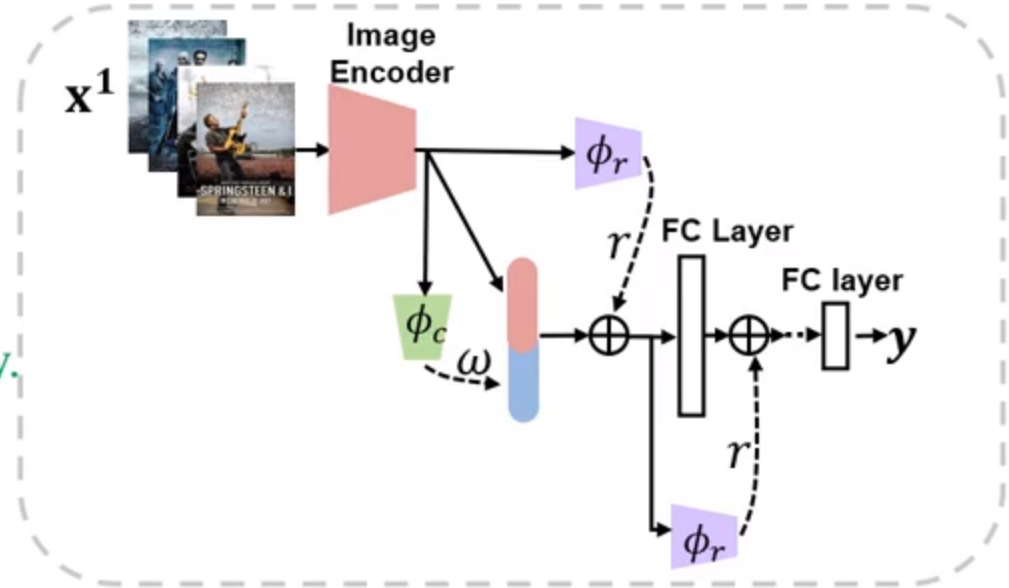
Unified training:

- Modality-incomplete.
- Modality-complete.

Test with full modality.



Test with missing modality.



Meta-Learning Framework

Directly train the θ, ϕ_c, and ϕ_r is not feasible.

Input: Multimodal training dataset $\mathcal{D} = \{D^f, D^m\}$.

Output: Learned backbone θ^*, auxiliary $\psi = \{\phi_c, \phi_r\}$.

Training scheme:

- *Meta-train*

  Update backbone $\theta \rightarrow \theta^*$ on D^m.

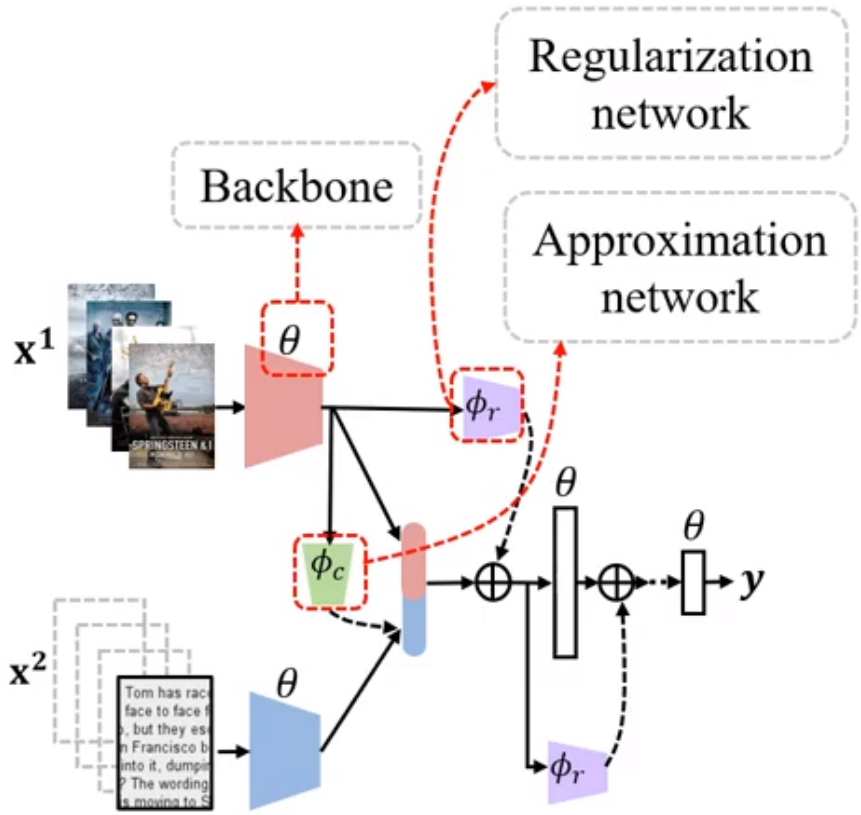
  Using the approximated feature and regularization.

- *Meta-test*

  Evaluate θ^* on D^f.

- *Meta-update*

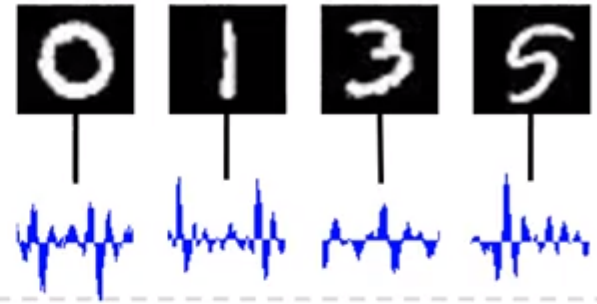
  Update θ and ψ. by gradient descent



Datasets

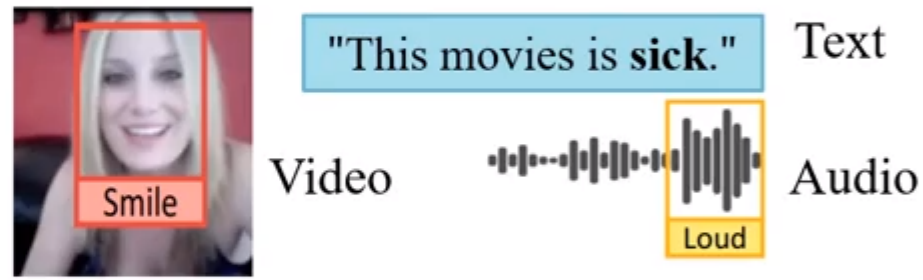
av-MNIST

- Modalities: image, audio.
- Task: digits classification.
- Experiment: flexibility.



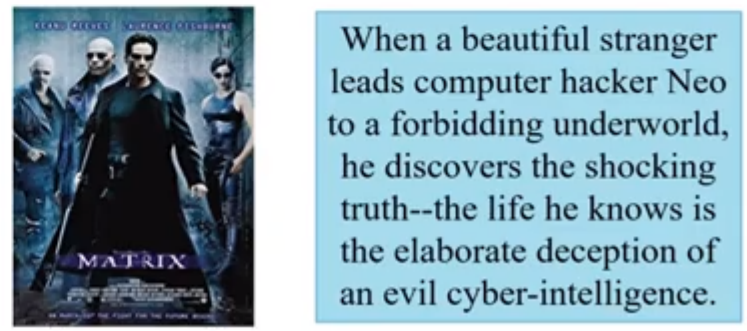
CMU-MOSI

- Modalities: video, audio, and text.
- Task: emotion classification.
- Experiment: efficiency.

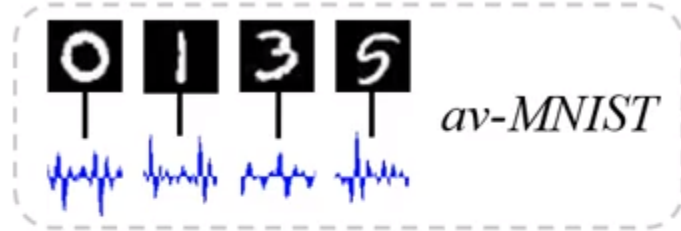


MM-IMDb

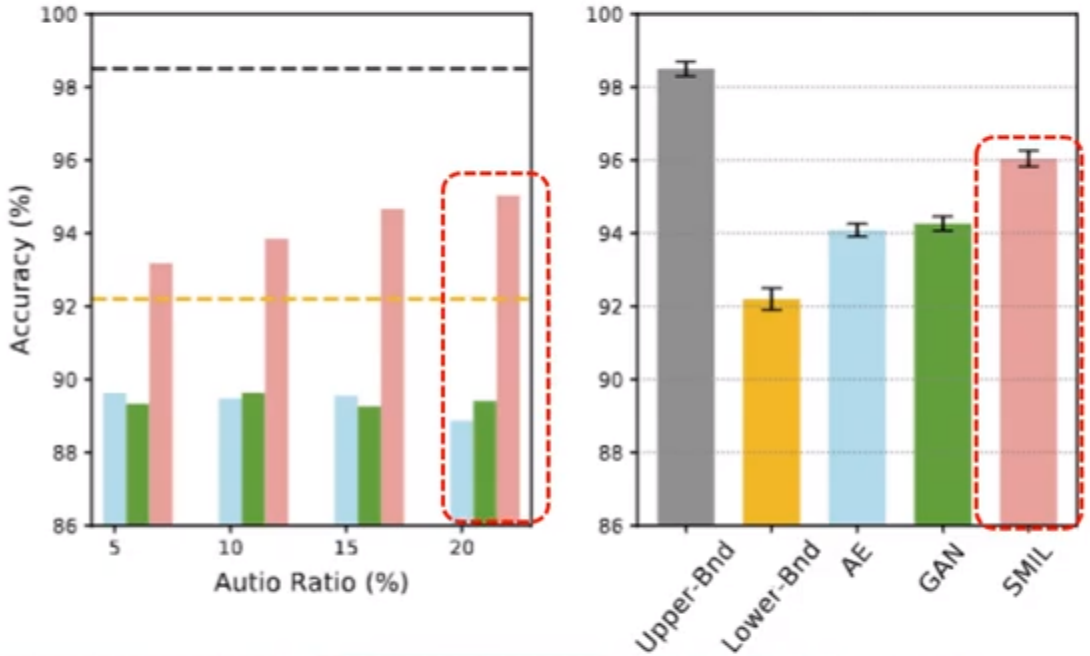
- Modalities: image and text.
- Task: movie genre classification.
- Experiment: efficiency.



Experiment Results: Flexibility



Left: training with 100% Image + η% Audio and testing with Image Only.
Right: training with 100%Image + η% Audio and testing with Image + Audio.



AE or GAN can not handle testing with missing modality.

- The performance of AE or GAN is inferior to Lower-bound (89.8% *vs.* 92.0%).

SMIL is flexible to handle missing or full modality.

- SMIL, trained using 100% Image and 20% Audio, gives consistent performance (94.9% *vs.* 96.0%). AE or GAN gives different performance (89.8% *vs.* 94.0%).

Generative model-based method.

Model trained using 100% Image + 100% Audio.

Model trained using 100% Image.

Experiment Results: Efficiency

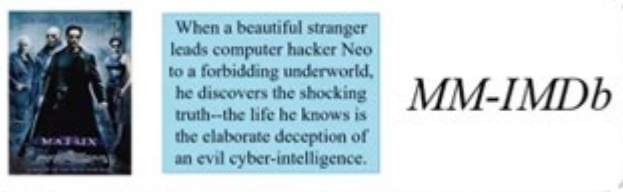
100% Image + 20% Text.

Method	Accuracy ↑		F1 score ↑	
	100%		100%	
Lower-bound	44.8		27.7	
Upper-bound	71.0		70.5	
MVAE	58.5		58.1	

Method	Accuracy ↑		F1 score ↑	
	10%	20%	10%	20%
AE	56.4	60.4	54.4	59.0
GAN	<u>56.5</u>	<u>60.6</u>	<u>54.6</u>	<u>59.1</u>
SMIL	**60.7**	**63.3**	**58.0**	**62.5**

- SMIL is efficient under severe missing modality.
- Multimodal training is better than single modality training.

[MVAE] Mike et al. In NeurIPS '18.



Method	F1 Samples ↑	F1 Micro ↑
	100%	100%
Lower-bound	47.6	48.2
Upper-bound	61.7	52.0
MVAE	48.4	48.6

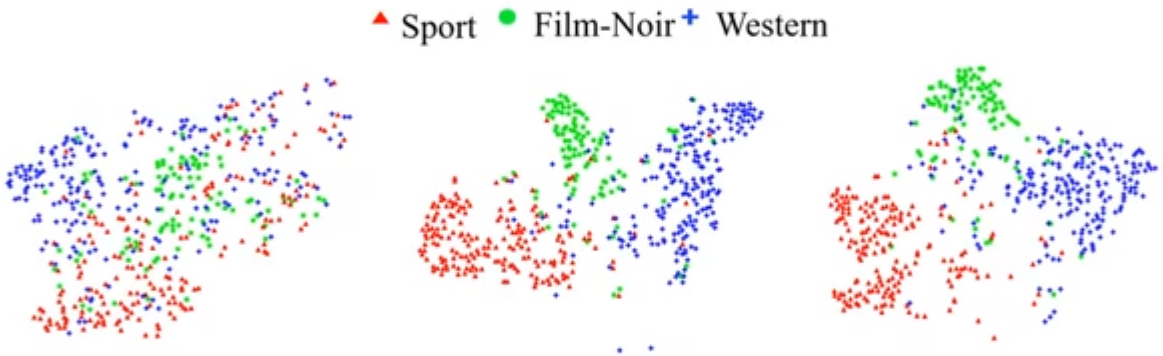
Method	F1 Samples ↑		F1 Micro ↑	
	10%	20%	10%	20%
AE	44.5	50.9	44.8	50.7
GAN	<u>45.0</u>	<u>51.1</u>	<u>44.6</u>	<u>51.0</u>
SMIL	**49.2**	**54.1**	**49.5**	**54.6**

Visualization & Ablation Study

When a beautiful stranger leads computer hacker Neo to a forbidding underworld, he discovers the shocking truth--the life he knows is the elaborate deception of an evil cyber-intelligence.

MM-IMDb

Visualization of embeddings.



(a) 100% Image (Lower-Bound)

(b) 100% Image + 20% Text (SMIL)

(c) 100% Image + 100% Text (Upper-Bound)

Ablation Study

Method	F1 samples		F1 Micro	
	10%	20%	10%	20%
SMIL w/o K-means	48.2	53.5	48.5	53.0
SMIL w/o Regularization	46.9	52.1	47.2	53.0
SMIL (Full)	**49.2**	**54.1**	**49.5**	**54.6**

Feature regularization is critical for severely missing modality.

- 4.9% ↓ without regularization (in 10% text case).
- 3.8% ↓ without regularization (in 20% text case).