M3ER: Multiplicative Multimodal Emotion Recognition using Facial, Textual, and Speech Cues

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, Dinesh Manocha (2020, AAAI)

김유리

Index

Introduction

Introduction
Contribution

- facial(video), text, audio feature 사용한 Emotion recognition model 제안
- Efficiently feature 구분 위해서 CCA 검증 사용
- 일부 modality feature 누락 시 effective feature을 주기위한 feature transformation method 제공 (proxy vector)

Introduction

Unimodal vs. Multimodal?

Unimodal : 소리, 시각, 언어, 지식, …

Multimodal : 시각+언어, 소리+언어, …

Introduction

왜 Multimodal을 사용하는가?

1. Richer information : 다른 modality끼리 보완 가능
2. Robustness to Sensor Noise : 데이터가 손상되거나 누락될 가능성이 있음

Introduction

Multimodal 사용시 해결해야할 문제

1. 어떤 modality를 결합할 것인가
2. modality를 어떻게 결합할 것인가

Introduction
Contribution

다시 본 논문의 Contribution으로 돌아와서!

- facial(video), text, audio feature 사용한 Emotion recognition model 제안
- Efficiently feature 구분 위해서 CCA 검증 사용
- 일부 modality feature 누락 시 effective feature을 주기위한 feature transformation method 제공 (proxy vector)

Introduction
Contribution

다시 본 논문의 Contribution으로 돌아와서!

- facial(video), text, audio feature 사용한 Emotion recognition model 제안
- Efficiently feature 구분 위해서 **CCA 검증** 사용
- 일부 modality feature 누락 시 effective feature을 주기위한 feature **transformation method** 제공 (proxy vector)

Introduction
Contribution

CCA(Canonical Correlational Analysis) 검증?

두 변수 집단의 연관성(association)을 각 변수 집단에 속한 변수들의

선형 결합(linear combination)의 상관계수를 이용하여 설명

Introduction

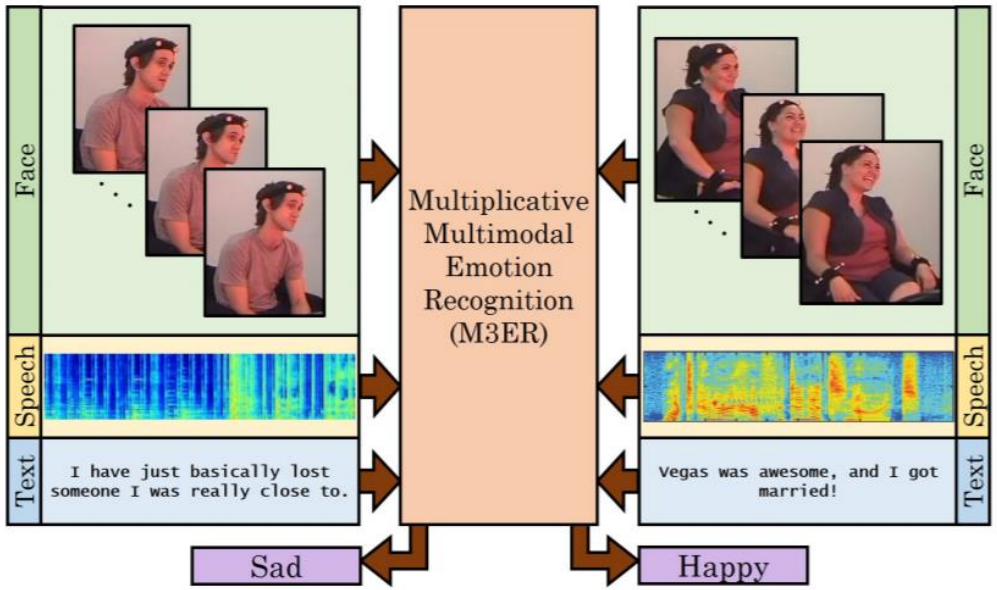


Figure 1: **Multimodal Perceived Emotion Recognition:** We use multiple modalities to perform perceived emotion prediction. Our approach uses a deep learning model along with a multiplicative fusion method for emotion recognition. We show results on two datasets, IEMOCAP and CMU-MOSEI both of which have face, speech and text as the three input modalities. Above is one sample point extracted from the IEMOCAP dataset.

Introduction

Dataset?

IEMOCAP(busso et al. 2008), CMU-MOSEI(Zadeh et al. 2018)

Figure 4: **Qualitative Results on IEMOCAP:** We qualitatively show data points correctly classified by M3ER from all the 4 class labels of IEMOCAP. The labels as classified by M3ER in row order from top left, are *Angry, Happy, Neutral, Sad*.

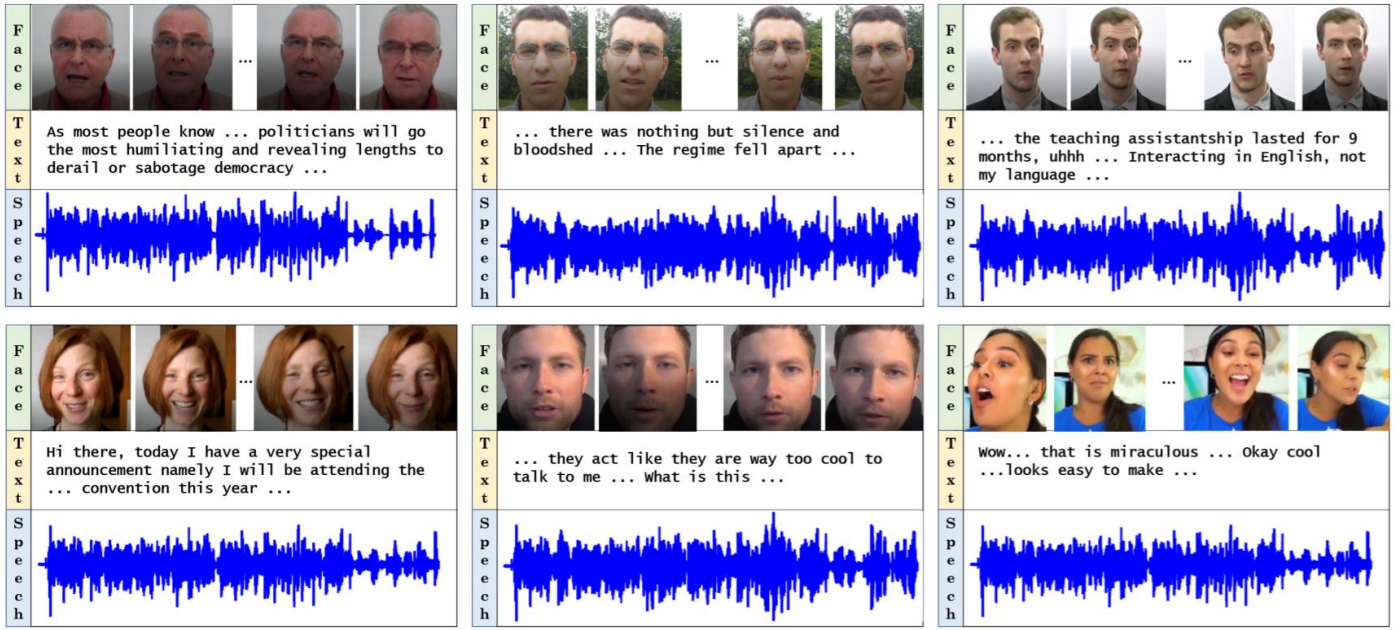


Figure 3: **Qualitative Results on CMU-MOSEI:** We qualitatively show data points correctly classified by M3ER from all the 6 class labels of CMU-MOSEI. The labels as classified by M3ER in row order from top left, are *Anger, Disgust, Fear, Happy, Sad, Surprise*.

Methods

Methods
Architecture

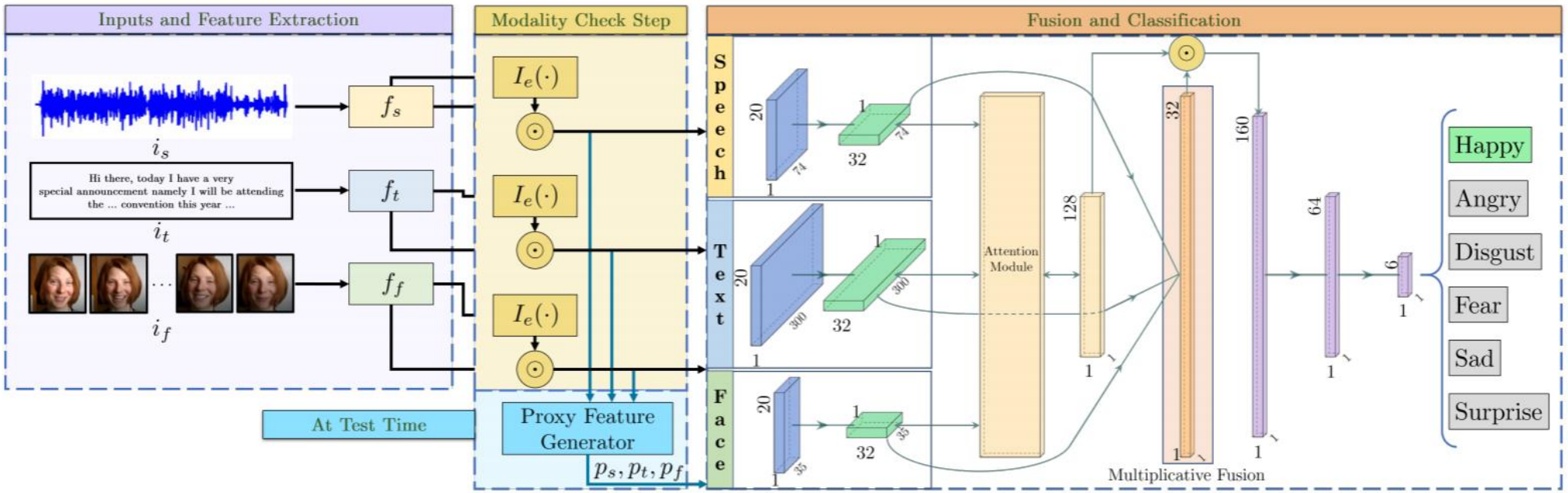


Figure 2: **M3ER:** We use three modalities, speech, text and the facial features. We first extract features to obtain f_s, f_t, f_f from the raw inputs, i_s, i_t and i_f (purple box). The feature vectors then are checked if they are effective. We use a indicator function I_e (Equation 1) to process the feature vectors (yellow box). These vectors are then passed into the classification and fusion network of M3ER to get a prediction of the emotion (orange box). At the inference time, if we encounter a noisy modality, we regenerate a proxy feature vector (p_s, p_t or p_f) for that particular modality (blue box).

Methods
Modality Check Step

Modality Check Step

$$f_i' = H_{i,j}^i f_i$$

$$f_j' = H_{i,j}^j f_j$$

$$\rho(f_i', f_j') = \frac{cov(f_i', f_j')}{\sigma_{f_i'}\sigma_{f_j'}}$$

$$\rho(f_i', f_j') < \tau$$

$$\boldsymbol{I_e(f_i)} = \begin{cases} \mathbf{0} \;\; \boldsymbol{\rho(f_i', f_j') < \tau, (i,j) \in \mathcal{M}, i \neq j} \\ \mathbf{1} \;\; \boldsymbol{else} \end{cases}$$

Methods
Modality Check Step

Modality Check Step

$$f_i' = H_{i,j}^i f_i$$

$$f_j' = H_{i,j}^j f_j$$

i, j는 modality 중 2가지
해당 과정을 통해 same lower dim으로 reduce
(output dim = 100)

H : 모든 modality 쌍에 대해 training set의 상관계수 미리 계산

$$\rho(f_i', f_j') = \frac{cov(f_i', f_j')}{\sigma_{f_i'}\sigma_{f_j'}}$$

$$\rho(f_i', f_j') < \tau$$

$$\boldsymbol{I_e(f_i)} = \begin{cases} \mathbf{0} & \boldsymbol{\rho(f_i', f_j') < \tau, (i,j) \in \mathcal{M}, i \neq j} \\ \mathbf{1} & \boldsymbol{else} \end{cases}$$

Methods
Modality Check Step

Modality Check Step

$$f_i' = H_{i,j}^i f_i$$

$$f_j' = H_{i,j}^j f_j$$

$$\rho(f_i', f_j') = \frac{cov(f_i', f_j')}{\sigma_{f_i'}\sigma_{f_j'}}$$

$$\rho(f_i', f_j') < \tau$$

두개의 feature에 대한 correlation score 계산 후 threshold(τ) 기준으로 판단

$$\boldsymbol{I_e(f_i)} = \begin{cases} \mathbf{0} \;\; \boldsymbol{\rho(f_i', f_j') < \tau, (i,j) \in \mathcal{M}, i \neq j} \\ \mathbf{1} \;\; \boldsymbol{else} \end{cases}$$

Methods
Proxy Feature Vectors

Generating feature vector가 challenging한 이유?

각 modality 간의 관계성이 non-linear하기 때문

Methods
Proxy Feature Vectors

Regenerating Proxy Feature Vectors

$$p_i = \mathcal{T} f_i, where\ i\ \in\ \mathcal{M}\ and\ \mathcal{T}\ if\ any\ linear\ transformation$$

1. The first step is $to\ find\ v_j = argmin_j d(v_j, f_f), where\ d\ is\ any\ distance\ metric. (L2\ norm)$

   $2. Compute\ constants\ a_i \in \mathbb{R}\ by\ solving\ the\ following\ linear\ system, f_f = \sum_{i=1}^{p} a_i v_i$

$$\boldsymbol{p_s} = \mathcal{T} f_f = \sum_{i=1}^{p} a_i \mathcal{T} v_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{a_i w_i}$$

Methods
Proxy Feature Vectors

Regenerating Proxy Feature Vectors

$$p_i = \mathcal{T} f_i, where\ i\ \in\ \mathcal{M}\ and\ \mathcal{T}\ if\ any\ linear\ transformation$$

1. The first step is $to\ find\ v_j = argmin_j d(v_j, f_f), where\ d\ is\ any\ distance\ metric.\ (L2\ norm)$

   $2.\ Compute\ constants\ a_i \in \mathbb{R}\ by\ solving\ the\ following\ linear\ system, f_f = \sum_{i=1}^{p} a_i v_i$

$$\boldsymbol{p_s} = \mathcal{T} f_f = \sum_{i=1}^{p} a_i \mathcal{T} v_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{a_i w_i}$$

우리가 생성 해야 할 facial feature(f_f)와 알고있는 j 번째 facial feature(v_j) 사이의 L2 norm(d)을 최소화하는 v_j를 찾음

Methods
Proxy Feature Vectors

Regenerating Proxy Feature Vectors

$$p_i = \mathcal{T} f_i, where\ i\ \in\ \mathcal{M}\ and\ \mathcal{T}\ if\ any\ linear\ transformation$$

1. The first step is $to\ find\ v_j = argmin_j d(v_j, f_f), where\ d\ is\ any\ distance\ metric. (L2\ norm)$

   $2. Compute\ constants\ a_i \in \mathbb{R}\ by\ solving\ the\ following\ linear\ system, f_f = \sum_{i=1}^{p} a_i v_i$

$$\boldsymbol{p_s} = \mathcal{T} f_f = \sum_{i=1}^{p} a_i \mathcal{T} v_i = \sum_{\boldsymbol{i=1}}^{\boldsymbol{p}} \boldsymbol{a_i w_i}$$

$\mathcal{T}: F_b \rightarrow S_b$ linear transformation
위 단계를 각 effective feature에 대해 계산해보고 결과의 평균을 사용

Methods
Multiplicative Modality Fusion

Multiplicative Modality Fusion

$$define\ the\ loss\ for\ the\ i^{th}\ modality\ as\ follws$$

$$c^{(y)} = -\sum_{i=1}^{M}\prod_{j\neq i}(1-p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}$$

modified **loss** is as follows

$$\boldsymbol{c^{(y)} = -\sum_{i=1}^{M}(p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}}$$

[Liu et al. 2018] Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to combine modalities in multimodal deep learning. arXiv:1805.11730.

Methods
Multiplicative Modality Fusion

Multiplicative Modality Fusion

$$define\ the\ loss\ for\ the\ i^{th}\ modality\ as\ follws$$

$$c^{(y)} = -\sum_{i=1}^{M}\prod_{j\neq i}(1-p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}$$

modified **loss** is as follows

$$\boldsymbol{c^{(y)} = -\sum_{i=1}^{M}(p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}}$$

original 수식 [Liu et al. 2018]

M : modality 개수
y : true class label
β : hyper parameter
(down-weights the unreliable modalities and $p_j^{(y)}$)

[Liu et al. 2018] Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to combine modalities in multimodal deep learning. arXiv:1805.11730.

Methods
Multiplicative Modality Fusion

Multiplicative Modality Fusion

$define\ the\ loss\ for\ the\ i^{th}\ modality\ as\ follws$

$$c^{(y)} = -\sum_{i=1}^{M}\prod_{j\neq i}(1-p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}$$

modified **loss** is as follows

$$\boldsymbol{c^{(y)} = -\sum_{i=1}^{M}(p_j^{(y)})^{\beta/(M-1)}\log p_i^{(y)}}$$

original : wrong prediction을 명시적으로 사용

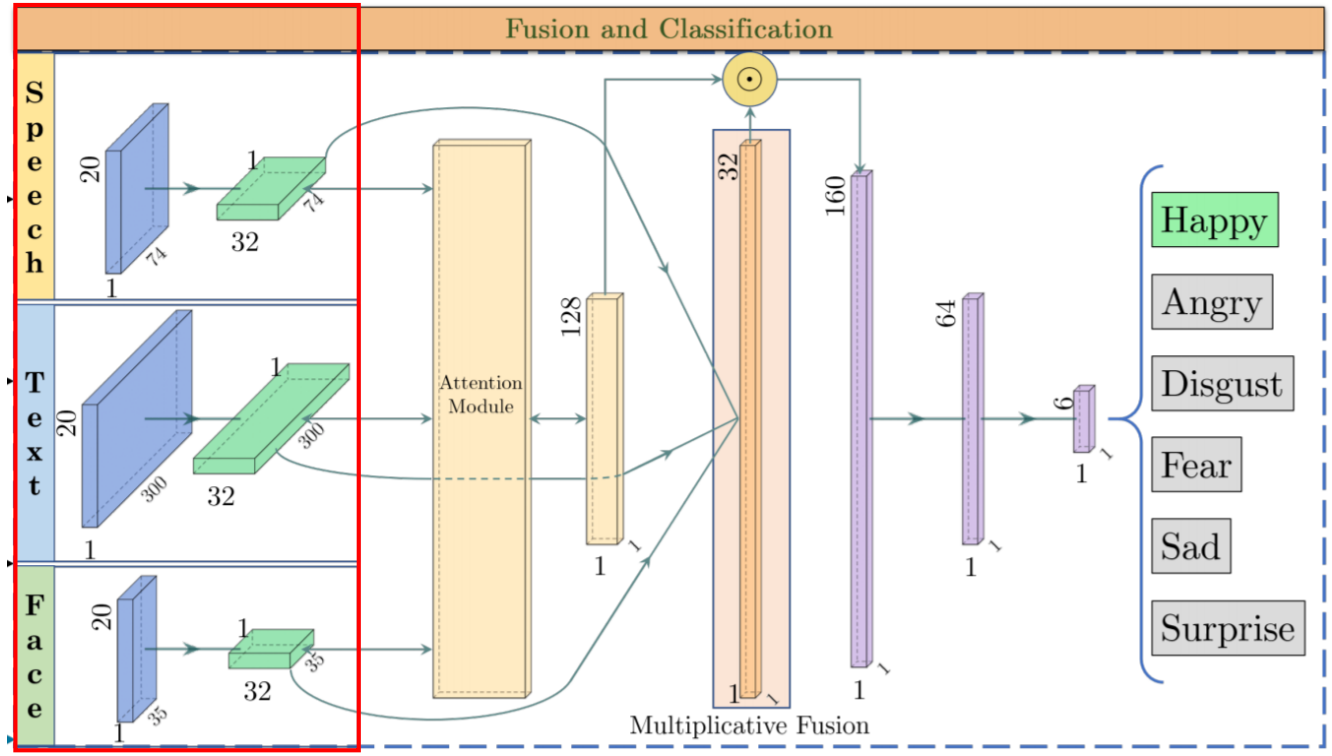
modified : ignore the wrong predictions by simply not addressing them

[Liu et al. 2018] Liu, K.; Li, Y.; Xu, N.; and Natarajan, P. 2018. Learn to combine modalities in multimodal deep learning. arXiv:1805.11730.

Methods
Classification Network

Classification Network

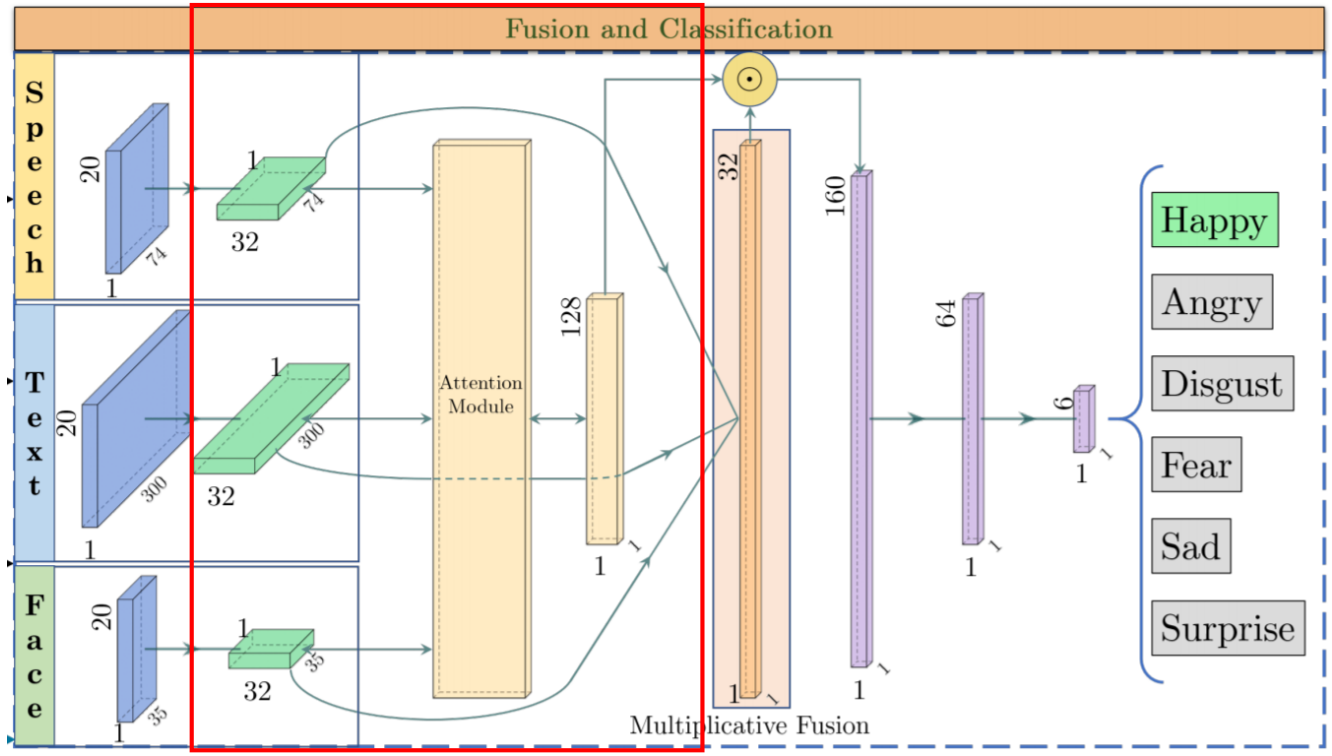
각 modality input은 single-hidden-layer LSTM 통과 (dim : 32)



Methods
Classification Network

Classification Network

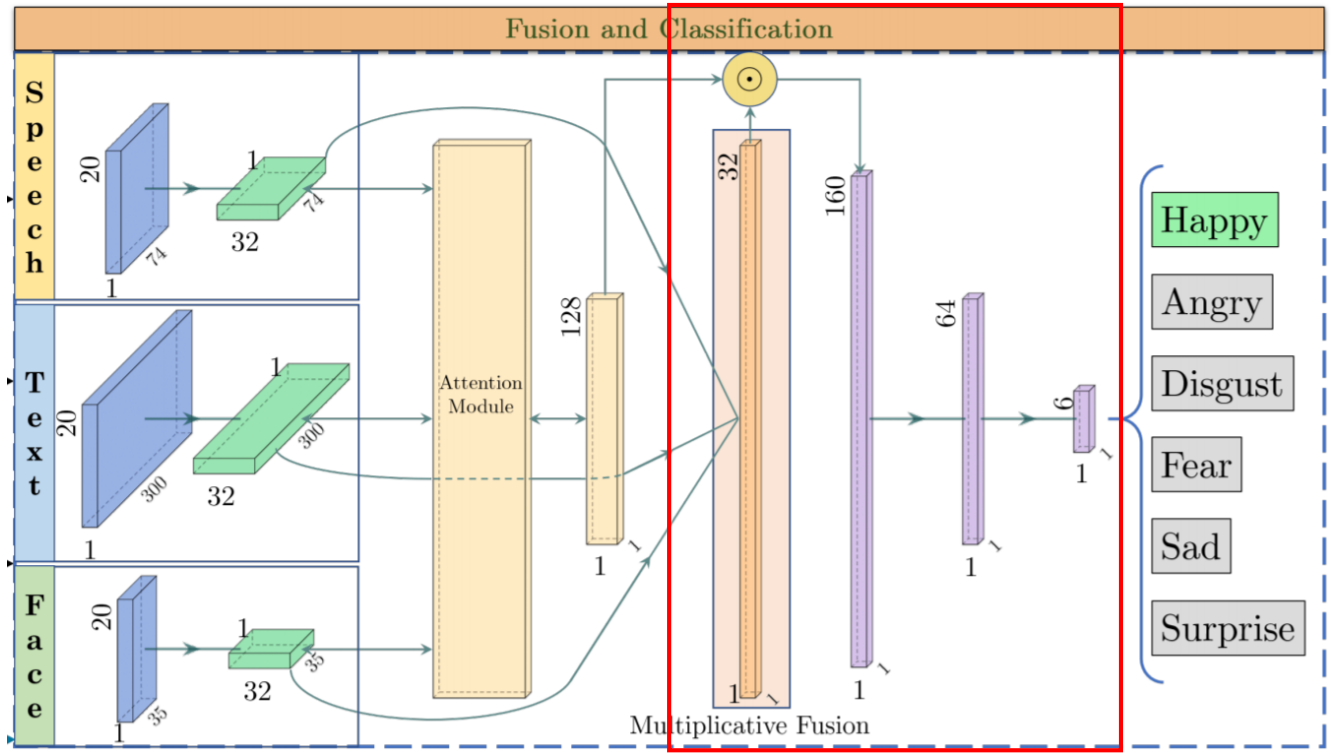
LSTM의 output은 0으로 초기화 된 128 dim memory variable과 함께 attention module 통과 (input modality의 max길이 t 만큼 반복)
attention module의 출력은 memory variable, LSTM 입력 모두 업데이트



Methods
Classification Network

Classification Network

3개의 LSTM 출력 값은 multiplicative fusion에 의해 32 dim feature vector로 변형
160 dim feature vector는 64 dim FC layer 통과



Experiments

Experiments

1. M3ER을 통한 Emotion recognition 성능 평가
2. 생성된 proxy vector의 유효성에 대한 평가

Experiments

Training Details

- Train : validation : test = 7 : 1 : 2
- batch size : 256
- epochs : 500
- Adam optimizer
- learning rate : 0.01
- NVIDIA GeForce GTX 1080 TI GPU

Experiments

(a) Ablation Experiments performed on IEMOCAP Dataset.

Ineffectual modalities?	Experiments	Angry		Happy		Neutral		Sad		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Liu et al. 2018)	0.794	80.6%	0.750	76.9%	0.695	68.0%	0.762	80.8%	0.751	76.6%
	M3ER	**0.862**	**86.8%**	**0.862**	**81.6%**	**0.745**	**74.4%**	**0.828**	**88.1%**	**0.824**	**82.7%**
Yes	M3ER – Modality Check Step – Proxy Feature Vector	0.704	71.6%	0.712	70.4%	0.673	64.7%	0.736	79.8%	0.706	71.6%
	M3ER – Proxy Feature Vector	0.742	75.7%	0.745	73.7%	0.697	66.9%	0.778	84.0%	0.741	75.1%
	M3ER	**0.799**	**82.2%**	**0.743**	**76.7%**	**0.727**	**67.5%**	**0.775**	**86.3%**	**0.761**	**78.2%**

(b) Ablation Experiments performed on CMU-MOSEI Dataset.

Ineffectual modalities?	Experiments	Angry		Disgust		Fear		Happy		Sad		Surprise		Overall	
		F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA	F1	MA
No	Original Multiplicative Fusion (Liu et al. 2018)	0.889	79.9%	0.945	89.6%	0.963	93.1%	0.587	55.8%	0.926	85.3%	0.949	90.0%	0.878	82.3%
	M3ER	**0.919**	**86.3%**	**0.927**	**92.1%**	**0.904**	**88.9%**	**0.836**	**82.1%**	**0.899**	**89.8%**	**0.952**	**95.0%**	**0.902**	**89.0%**
Yes	M3ER – Modality Check Step – Proxy Feature Vector	0.788	73.3%	0.794	80.0%	0.843	85.0%	0.546	55.7%	0.832	79.5%	0.795	80.1%	0.764	75.6%
	M3ER – Proxy Feature Vector	0.785	77.8%	0.799	83.2%	0.734	77.5%	0.740	77.1%	0.840	86.0%	0.781	83.5%	0.783	80.9%
	M3ER	**0.816**	**81.3%**	**0.844**	**86.8%**	**0.918**	**89.4%**	**0.780**	**75.7%**	**0.873**	**86.1%**	**0.932**	**91.3%**	**0.856**	**85.0%**

Table 2: **Ablation Experiments:** We remove one component of M3ER at a time, and report the F1 and MA scores on the IEMOCAP and the CMU-MOSEI datasets, to showcase the effect of each of these components. Modifying the loss function leads to an increase of 6-7% in both F1 and MA. Adding the modality check step on datasets with ineffectual modalities leads to an increase of 2-5% in F1 and 4-5% in MA, and adding the proxy feature regeneration step on the same datasets leads to a further increase of 2-7% in F1 and 5-7% in MA.

Experiments

original feature vector, ineffectual version, regenerated feature

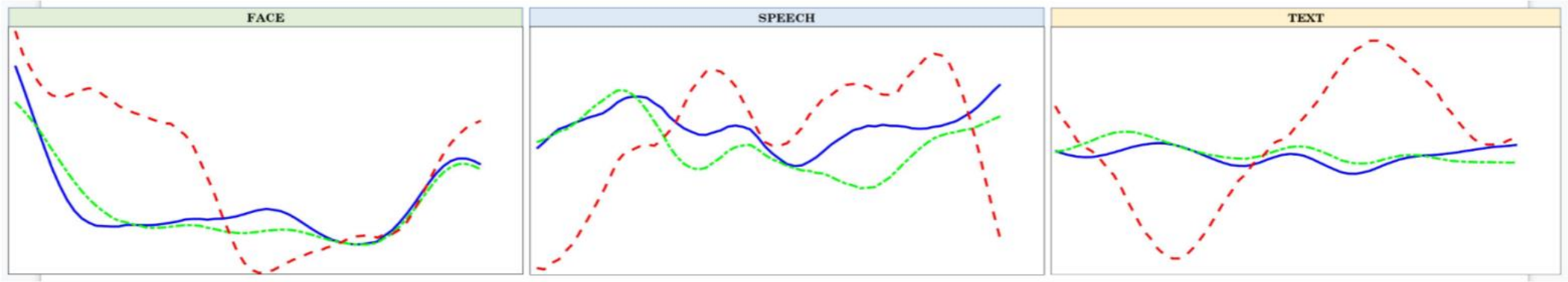


Figure 7: **Regenerated Proxy Feature Vector:** We show the quality of the regenerated proxy feature vectors for each of the three modalities. For the three graphs, we demonstrate the original feature vector (blue), the ineffectual version of the modality because of added white Gaussian noise (red) and the regenerated feature vector (green). The mean L_2 norm distance between the original and the regenerated vector for the speech, text and face modality are all around 0.01% of the L_2 norm of the respective data.

Conclusion

Conclusion

- multiplicative fusion layer를 사용하는 multimodal emotion recognition model(M3ER) 제안
- modality check step에서 무의미한 feature를 구분하기 때문에 다중 modality가 들어왔을 때 무의미한 feature 구분 가능
- multiplicative fusion을 통해 각 샘플 별로 어떤 modality에 가중치를 줄지 결정
- 3가지 modality(Facial, audio, text)를 사용함!

Limitations

Limitations

- 특정 label에 대해 성능이 낮음
- 현재 클래스 별 이진 분류 수행 → 인간의 감정은 주관적이기 때문에 이진 분류 보다 확률 분포와 더 유사(다중 분류를 고려해야함)

Thank you