Multimodal Learning

DMQA Lab, Korea Univ.

Seokho Moon

Nov 5, 2021

Presenter

❖ 문석호 (Seokho Moon)

- 고려대학교 산업경영공학과 대학원 재학 중
- Data Mining & Quality Analytics Lab (김성범 교수님)
- 석박사통합과정 (2019.09 ~)

❖ 관심 연구 분야

- Multimodal learning
- Self-supervised learning / Semi-supervised learning
- Anomaly detection

❖ E-mail

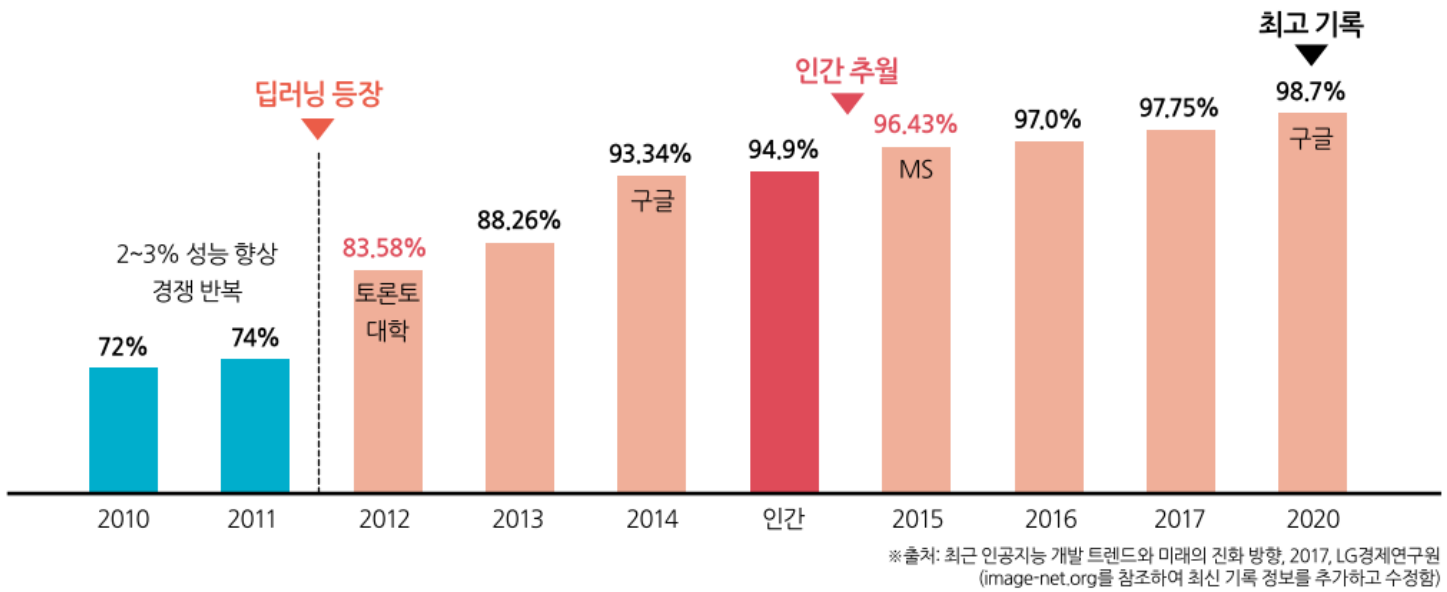
- danny232@korea.ac.kr

Contents

Introduction

Background

딥러닝 알고리즘의 발전과 컴퓨팅 성능의 확보로 인해 세부적인 작업에 대한 성능은 인간 수준을 넘어서고 있음

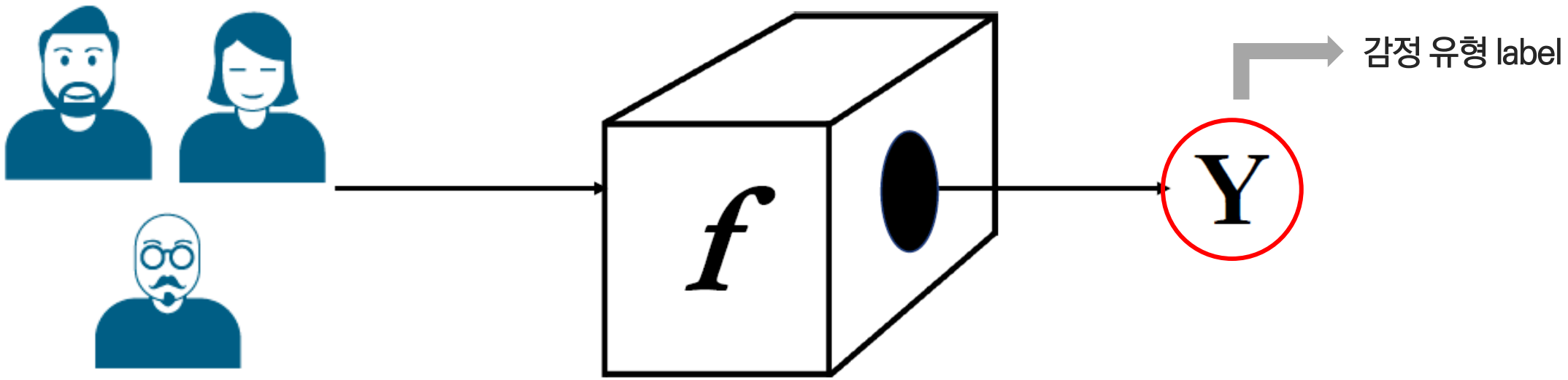


[이미지넷 인식 대회(ILSVRC) 성능]

Introduction

Background

인간 행동 인식이나 감정 인식 문제에 적용하고자 하면?



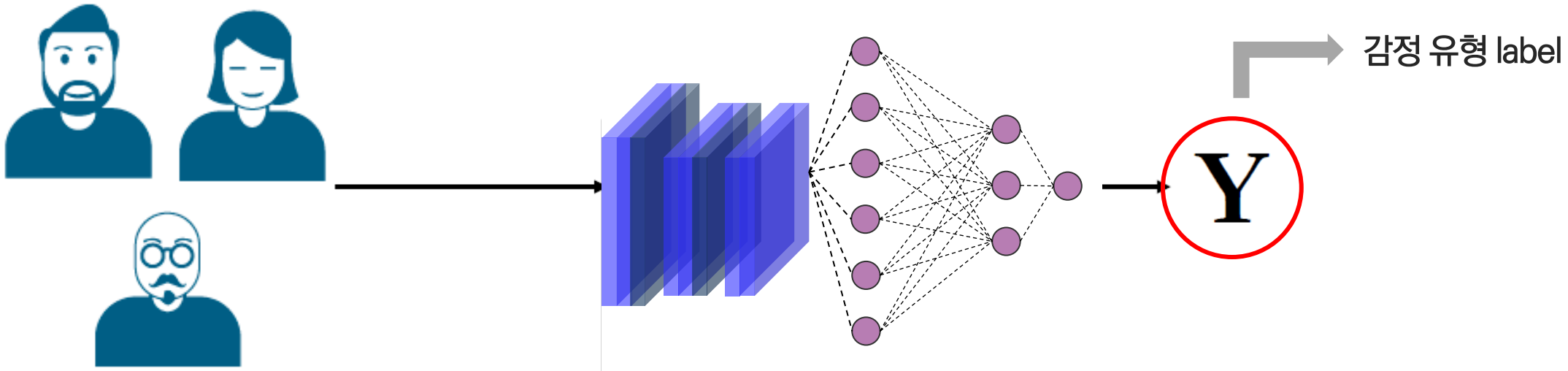
사람을 촬영한 **이미지** 데이터

[인공지능 (AI) 기반 감정 분류 모델]

Introduction

Background

인간 행동 인식이나 감정 인식 문제에 적용하고자 하면?



사람을 촬영한 **이미지** 데이터

[CNN 기반 감정 분류 모델]

Introduction

Background

그러나 인간 행동 인식이나 감정 인식 문제에서는

단순히 이미지를 잘 분류한다고 해서 성능이 확보되지 않음

Happiness ?

Neutral ?

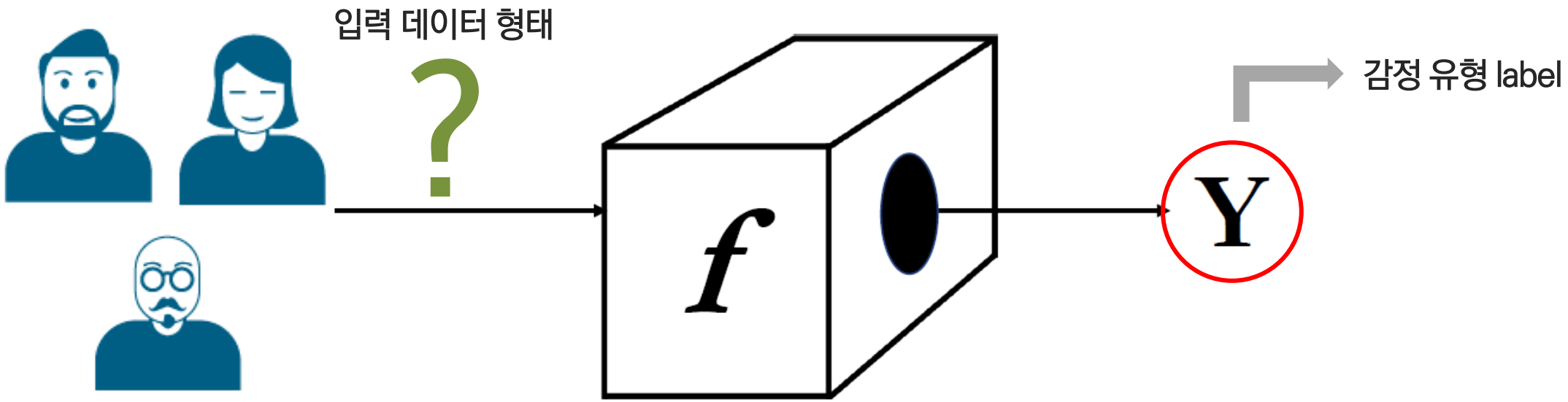
Sadness ?

Anger ?

[AI hub에서 제공하는 감정 인식 비디오 데이터셋]

Introduction

Multimodal Learning

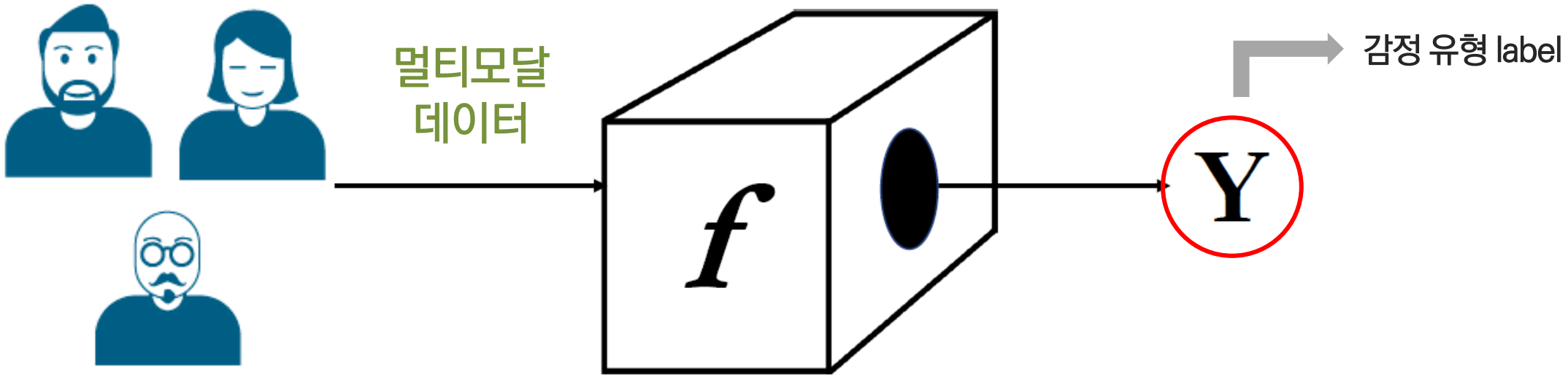


Introduction

Multimodal Learning

Audio, video, text 등 사람을 표현하는

여러 형태의 데이터(multimodal data)를 모델의 입력으로 사용



사람을 표현하는 데이터

(행동, 목소리, 표정, 대화 대본 등)

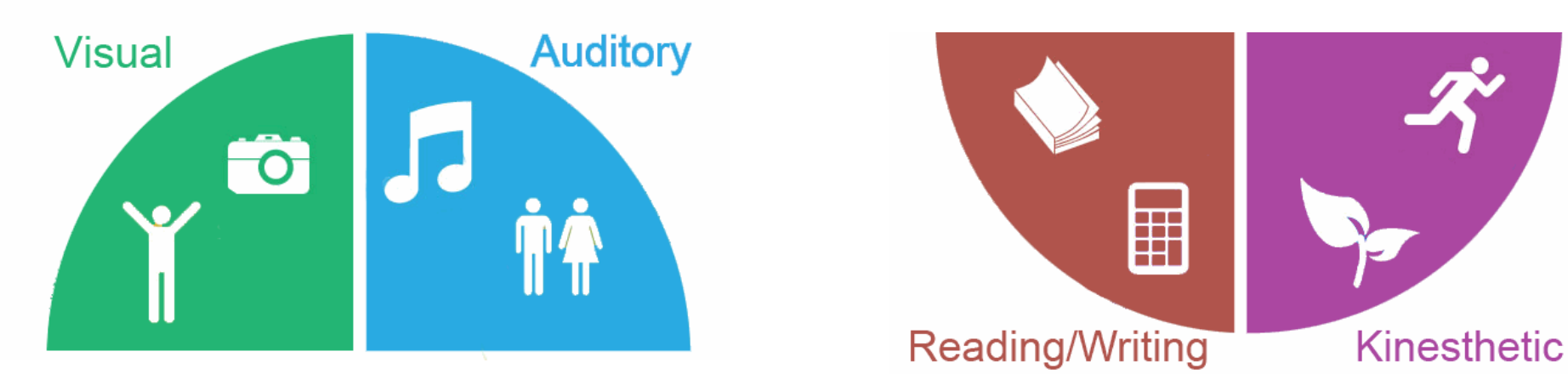
[멀티모달 감정 분류 모델]

Introduction

Multimodal Learning

❖ 멀티모달 딥러닝(multimodal deep learning)

- 단일 모달(single modal) 데이터의 한계를 극복하고자 여러 모달(multimodal)의 데이터를 사용하여 주어진 문제 (e.g., 감정 분류)를 해결하는 모델을 구축하는 방법론
- 모달(modality)라는 것은 **데이터의 형태**를 의미



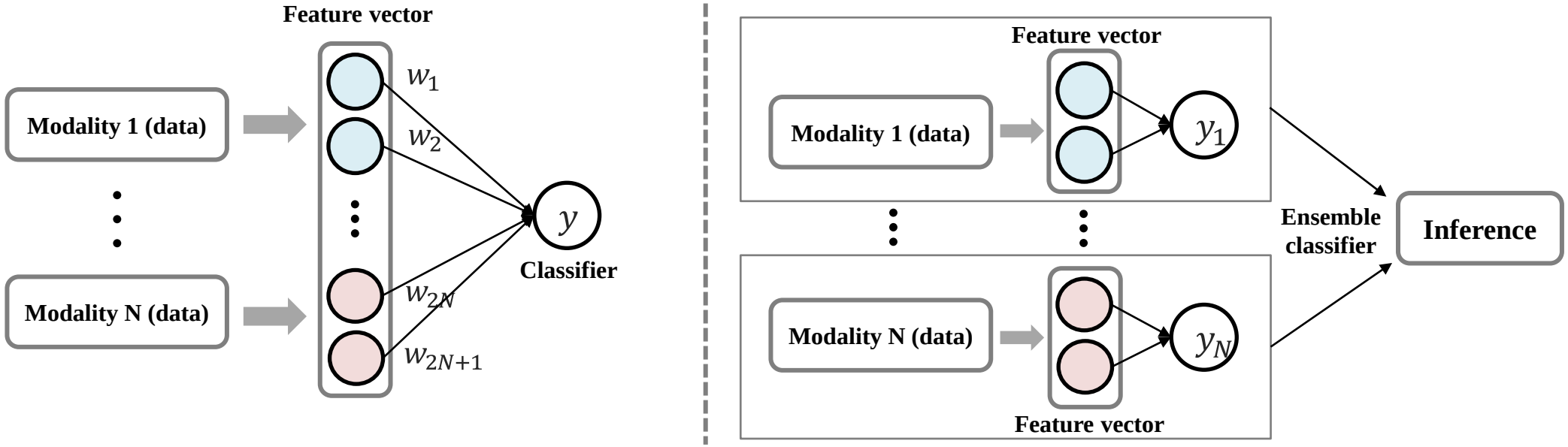
※ https://muda.kompas.id/baca/2016/12/11/knowledgeoflearningstyles/

Introduction

Multimodal Learning

❖ 멀티모달 딥러닝(multimodal deep learning)

- 각 모달에 적합한 딥러닝 구조를 사용하여 특징 벡터를 추출
- 모달을 통합하는 방식에는 대표적으로 (a) feature concatenation (b) ensemble classifier 두 가지 방법이 존재
- 최근 멀티모달 관련 application 연구 흐름은 (a) 방식에서 transformer 계열까지 적용한 구조까지 발전 중



(a) Feature Concatenation

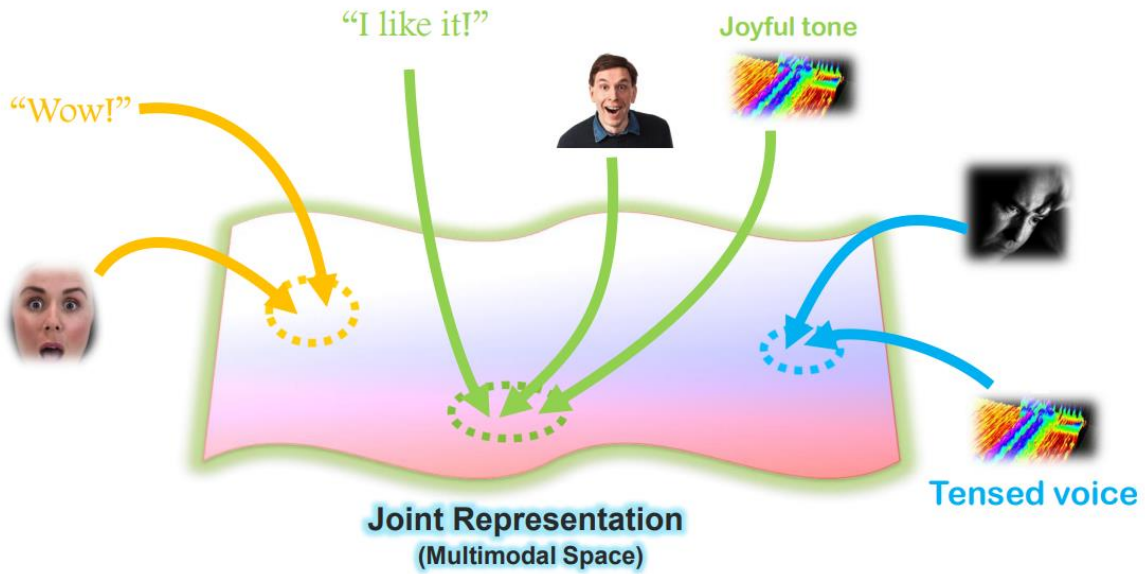
(b) Ensemble Classifier

Introduction

Multimodal Learning

❖ 멀티모달 딥러닝(multimodal deep learning)

- 각 모달에 적합한 딥러닝 구조를 사용하여 특징 벡터를 추출
- 모달을 통합하는 방식에는 대표적으로 (a) feature concatenation (b) ensemble classifier 두 가지 방법이 존재
- 최근 멀티모달 관련 application 연구 흐름은 (a) 방식에서 transformer 계열까지 적용한 구조까지 발전 중

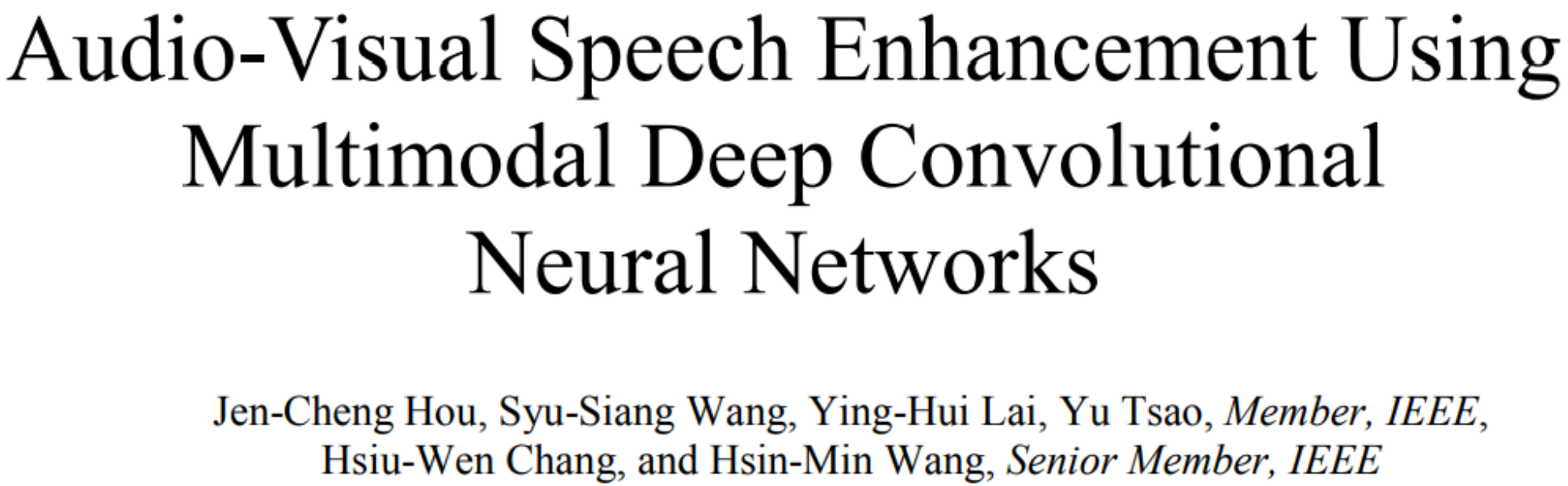


Related paper

Audio-visual speech enhancement using multimodal deep convolutional neural networks, 2018

❖ Feature vector concatenation 방식 소개

- 본 논문은 음성 향상(Speech enhancement)를 통해 음성 신호의 노이즈를 줄이는 것을 목표로 하고 있음


Audio-Visual Speech Enhancement Using Multimodal Deep Convolutional Neural Networks

Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, *Member, IEEE*,
Hsiu-Wen Chang, and Hsin-Min Wang, *Senior Member, IEEE*

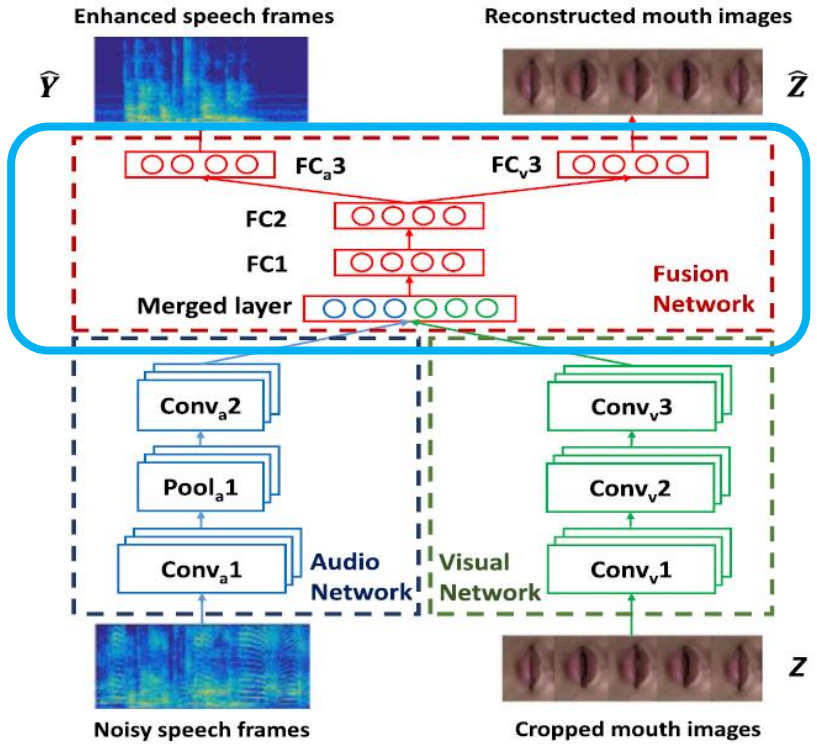
Hou, J. C., Wang, S. S., Lai, Y. H., Tsao, Y., Chang, H. W., & Wang, H. M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117-128.

Related paper

Audio-visual speech enhancement using multimodal deep convolutional neural networks, 2018

❖ Feature vector concatenation 방식 소개 논문

- 아래와 같이 feature vector에 대한 병합 구조를 가지고 있음



[AVDCNN 구조]

각 network에서 나온 feature vector를 simple concatenation 후에 merged layer로 변환

fully connected layer를 추가하여 object function을 통해 전체 parameter를 학습

$$\min_{\theta}\left(\frac{1}{K}\sum_{i=1}^{K}\left\|\hat{Y}_i - Y_i\right\|_2^2 + \mu\left\|\hat{Z}_i - Z_i\right\|_2^2\right)$$

Hou, J. C., Wang, S. S., Lai, Y. H., Tsao, Y., Chang, H. W., & Wang, H. M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117-128.

Related paper

Audio-visual speech enhancement using multimodal deep convolutional neural networks, 2018

❖ Feature vector concatenation 방식 소개 논문

- 아래 표처럼 특징 벡터들의 변환 사이즈를 볼 수 있음

Layer Name	Kernel	Activation Function	Number of Filters or Neurons
$Conv_a1$	12×2	Linear	10
$Pool_a1$	2×1		
$Conv_a2$	5×1	Linear	4
$Conv_v1$	15×2	Linear	12
$Conv_v2$	7×2	Linear	10
$Conv_v3$	3×2	Linear	6
Merged Layer			2804
FC1		Sigmoid	1000
FC2		Sigmoid	800
FC_a3		Linear	600
FC_v3		Linear	1500

[AVDCNN 모델 파라미터 수]

Merged feature vector의 크기를

FC2 layer까지 줄여나가면서 학습

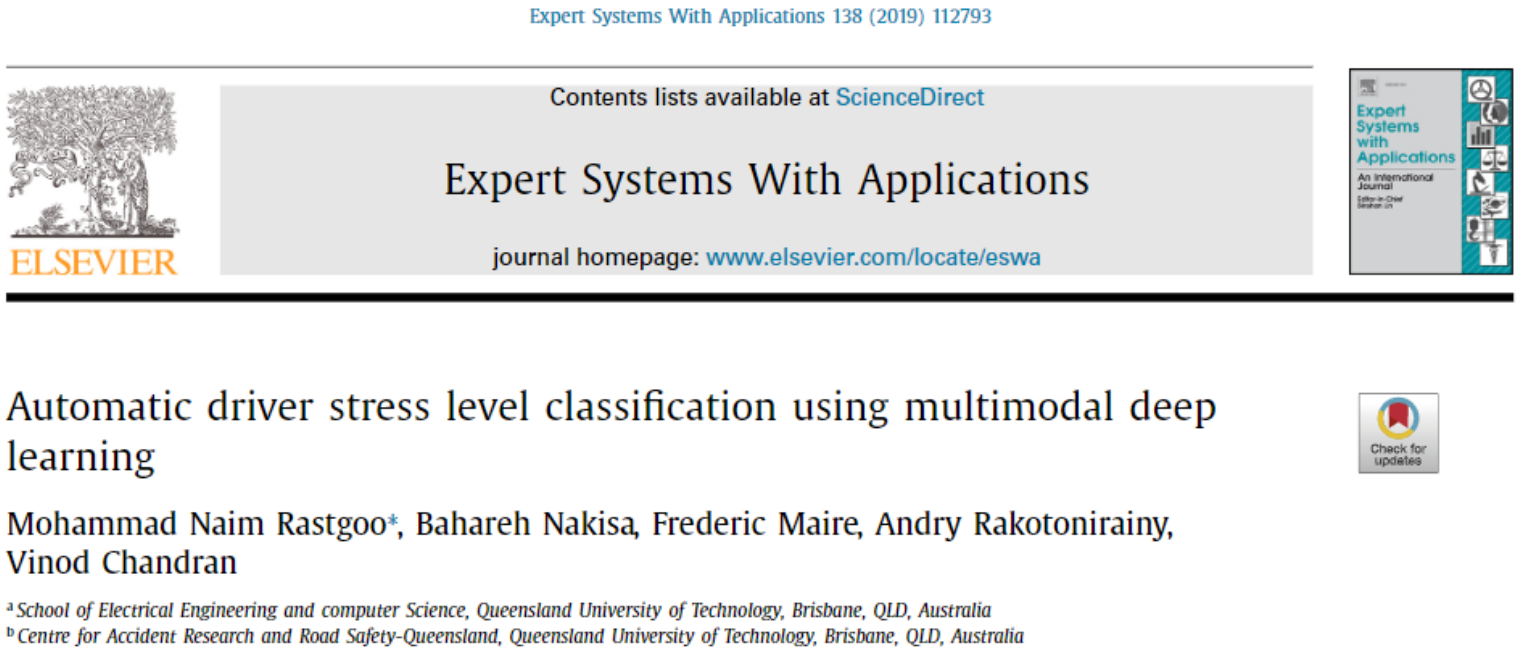
Hou, J. C., Wang, S. S., Lai, Y. H., Tsao, Y., Chang, H. W., & Wang, H. M. (2018). Audio-visual speech enhancement using multimodal deep convolutional neural networks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *2*(2), 117-128.

Related paper

Automatic driver stress level classification using multimodal deep learning, 2019

❖ Feature vector concatenation 방식 소개 논문

- 본 논문은 차량 사고에 큰 영향을 주는 주행 중 운전자의 스트레스 수준을 분석하고 감지하는 것을 목표로 함
- 운전자의 스트레스에 영향을 주는 데이터는 ECG 신호, 차량 데이터(스티어링 휠, 브레이크 페달 등), 상황 데이터(기상데이터 등)이며 이를 딥러닝 방법론으로 융합하여 우수한 성능 확보



Expert Systems With Applications 138 (2019) 112793

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Automatic driver stress level classification using multimodal deep learning

Mohammad Naim Rastgoo*, Bahareh Nakisa, Frederic Maire, Andry Rakotonirainy, Vinod Chandran

[a] School of Electrical Engineering and computer Science, Queensland University of Technology, Brisbane, QLD, Australia
[b] Centre for Accident Research and Road Safety-Queensland, Queensland University of Technology, Brisbane, QLD, Australia

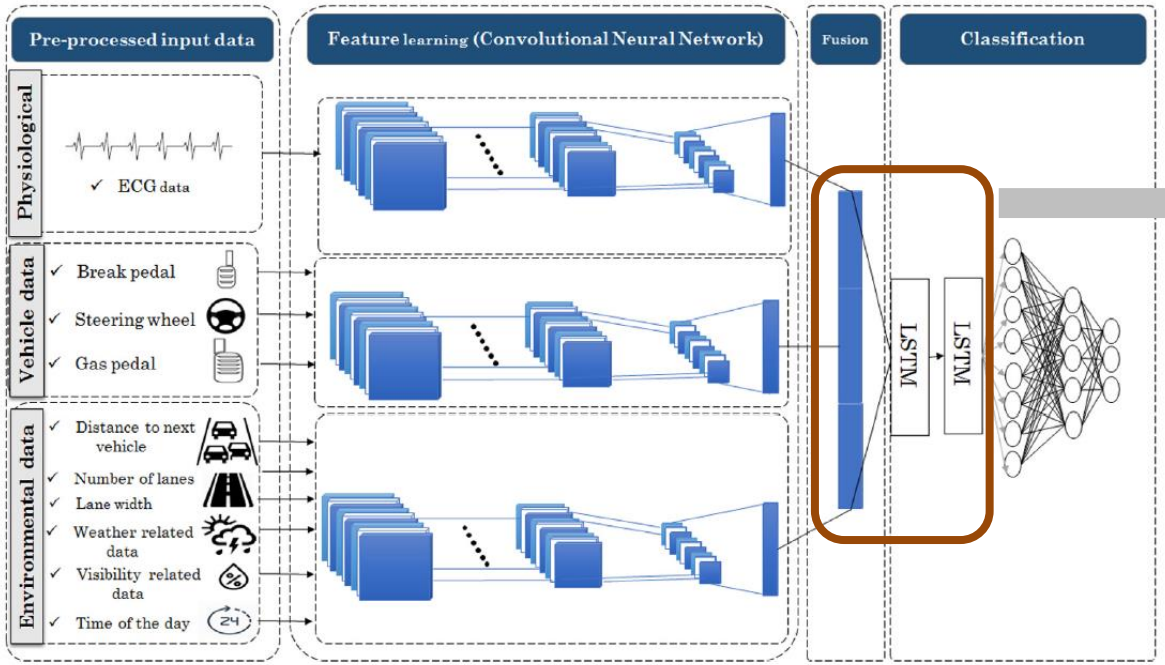
Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications*, *138*, 112793.

Related paper

Automatic driver stress level classification using multimodal deep learning, 2019

❖ Feature vector concatenation 방식 소개 논문

- 본 논문은 차량 사고에 큰 영향을 주는 주행 중 운전자의 스트레스 수준을 분석하고 감지하는 것을 목표로 함
- 운전자의 스트레스에 영향을 주는 데이터는 ECG 신호, 차량 데이터(스티어링 휠, 브레이크 페달 등), 상황 데이터 (기상데이터 등)이며 이를 딥러닝 방법론으로 융합하여 우수한 성능 확보



Fusion된 feature에서 LSTM을 통해 연속적인 특징까지 추출하고자 함

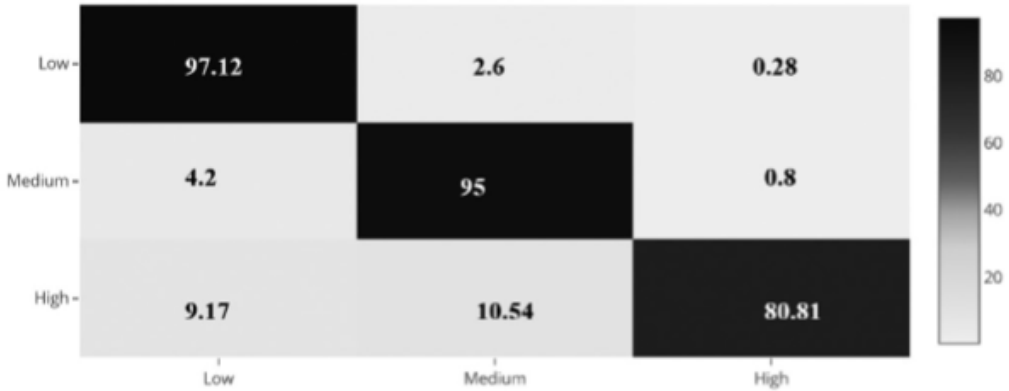
Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications*, *138*, 112793.

Related paper

Automatic driver stress level classification using multimodal deep learning, 2019

❖ Feature vector concatenation 방식 소개 논문

- 본 논문은 차량 사고에 큰 영향을 주는 주행 중 운전자의 스트레스 수준을 분석하고 감지하는 것을 목표로 함
- 운전자의 스트레스에 영향을 주는 데이터는 ECG 신호, 차량 데이터(스티어링 휠, 브레이크 페달 등), 상황 데이터(기상데이터 등)이며 이를 딥러닝 방법론으로 융합하여 우수한 성능 확보



[스트레스 수준에 따른 정오표]

Rastgoo, M. N., Nakisa, B., Maire, F., Rakotonirainy, A., & Chandran, V. (2019). Automatic driver stress level classification using multimodal deep learning. *Expert Systems with Applications*, *138*, 112793.

Related paper

Audio-visual emotion fusion (AVEF), 2019

❖ Feature vector concatenation 방식 소개 논문

- 본 논문은 audio, video 데이터를 통해 인간의 감정 인식 모델을 구축하고자 함
- 감정 인식 성능 향상을 위해 각 모달의 데이터마다 특징을 추출하여 병합한 뒤에 분류기를 학습시킴

Information Fusion 46 (2019) 184–192

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Audio-visual emotion fusion (AVEF): A deep efficient weighted approach

Yaxiong Ma[a], Yixue Hao[b], Min Chen[b], Jincai Chen[*,a,b,c], Ping Lu[a,b,c], Andrej Košir[d]

[a] Wuhan National Laboratory for Optoelectronics (WNLO), Huazhong University of Science and Technology, China
[b] Embedded and Pervasive Computing (EPIC) Lab, Huazhong University of Science and Technology, China
[c] Key Laboratory of Information Storage System (School of Computer Science and Technology, Huazhong University of Science and Technology), Ministry of Education of China, China
[d] Faculty of Electrical Engineering, University of Ljubljana, Tržaška cesta 25, Ljubljana 1000, Slovenia

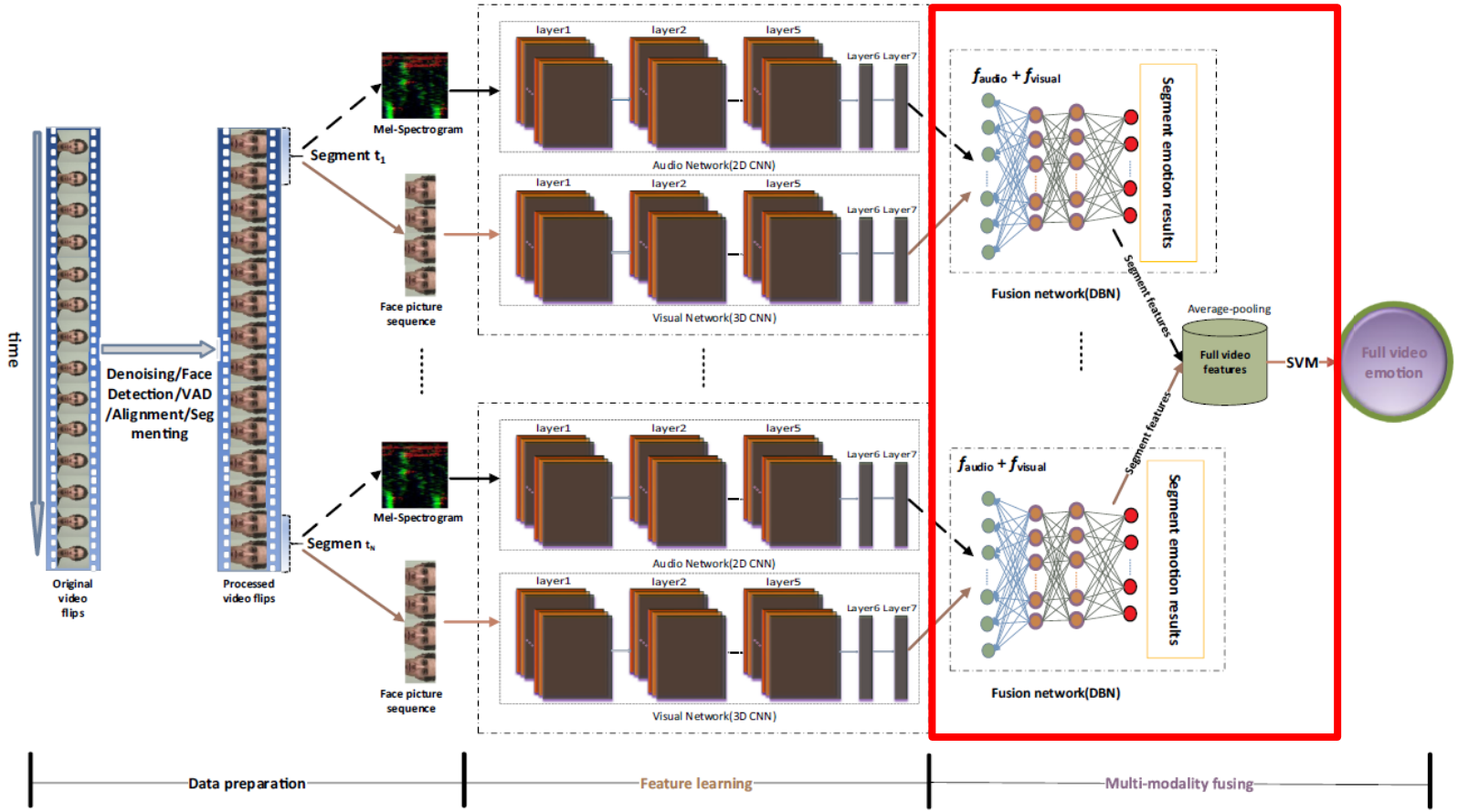
Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184-192.

Related paper

Audio-visual emotion fusion (AVEF), 2019

❖ Feature vector concatenation 방식 소개 논문

- 각 데이터를 전처리한 후에 네트워크에 입력값을 사용하는 형태



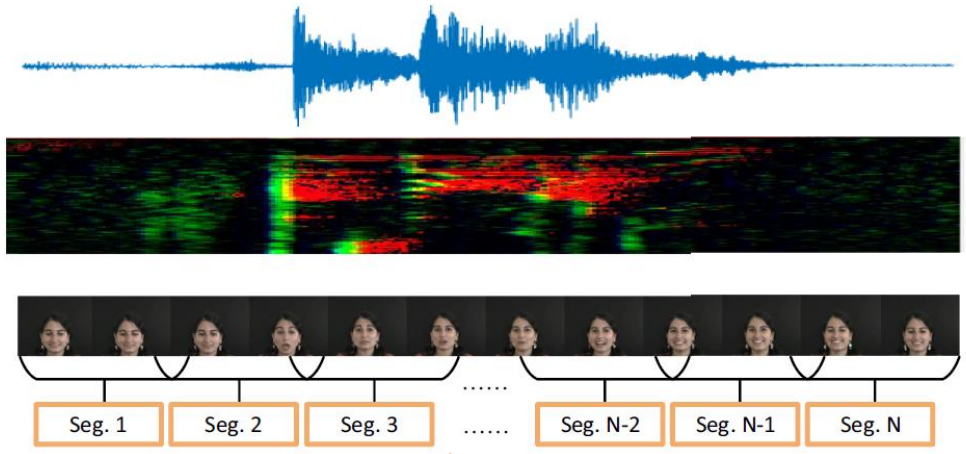
Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184-192.

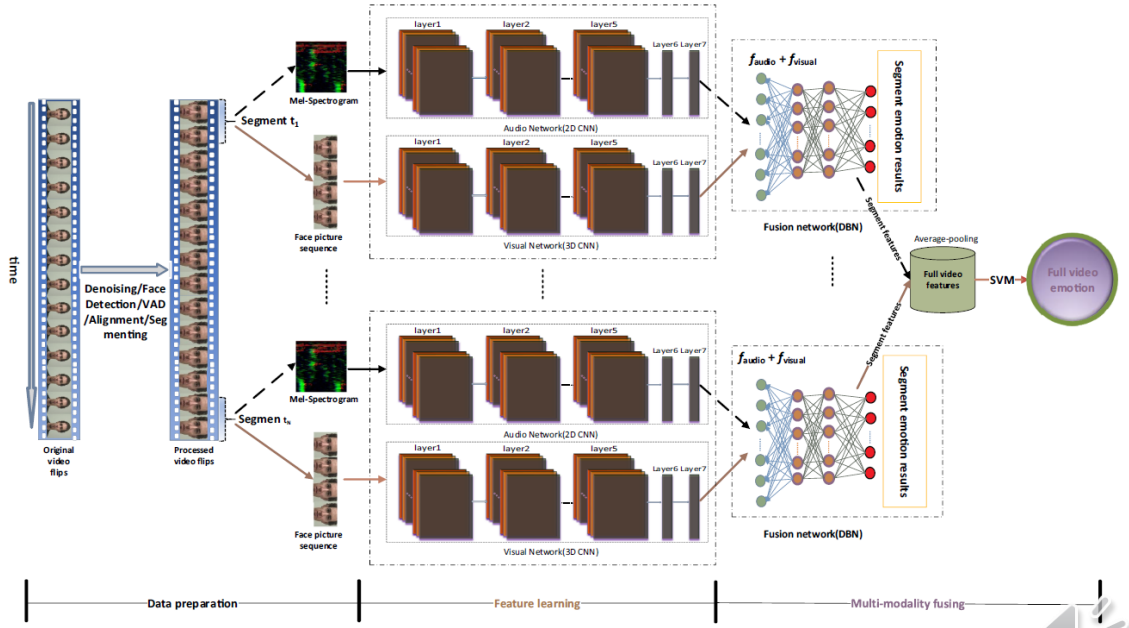
Related paper

Audio-visual emotion fusion (AVEF), 2019

❖ Feature vector concatenation 방식 소개 논문

- 주요 아이디어는 입력 데이터를 특정 segment별로 쪼개어 학습을 진행
- Segment 마다 추출된 오디오, 비디오 특징 벡터를 병합하고 난 후 전체 입력 데이터 수준에서 average-pooling 하는 방식





Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184-192.

Related paper

Audio-visual emotion fusion (AVEF), 2019

❖ Feature vector concatenation 방식 소개 논문

- 총 6가지 감정에 대해 분류모델을 딥러닝 멀티모달 방식을 적용하여 진행
- 기존의 단일 모달 방식의 분류 성능보다 향상됨을 확인
- 인간의 감정 인식 문제에서는 멀티모달 데이터를 사용하여 모델링을 진행하는 것이 훨씬 효율적

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
anger	**91.13**	0.00	1.13	5.52	1.37	0.85
disgust	1.22	**79.45**	6.92	3.35	7.92	1.14
fear	1.32	1.01	**77.90**	1.62	13.5	4.65
happiness	7.65	2.38	1.54	**86.7**	1.27	0.46
sadness	2.84	3.64	12.55	3.64	**76.12**	1.21
surprise	2.35	6.32	7.99	1.34	1.40	**80.60**

[RML benchmark 데이터셋의 분류 성능]

Ma, Y., Hao, Y., Chen, M., Chen, J., Lu, P., & Košir, A. (2019). Audio-visual emotion fusion (AVEF): A deep efficient weighted approach. *Information Fusion*, *46*, 184-192.

Related paper

Tensor fusion network for multimodal sentiment analysis, 2017

❖ Feature vector concatenation 방식 소개 논문

- Multimodal sentiment 분석에서는 비디오, 목소리, 언어 등을 입력값으로 활용하고자 함
- 본 논문에서는 각 데이터 형태의 특징을 합치는 tensor fusion network를 제안함

Tensor Fusion Network for Multimodal Sentiment Analysis

Amir Zadeh[†], **Minghai Chen**[†]
Language Technologies Institute
Carnegie Mellon University
{abagherz,minghail}@cs.cmu.edu

Soujanya Poria
Temasek Laboratories,
NTU, Singapore
sporia@ntu.edu.sg

Erik Cambria
School of Computer Science and
Engineering, NTU, Singapore
cambria@ntu.edu.sg

Louis-Philippe Morency
Language Technologies Institute
Carnegie Mellon University
morency@cs.cmu.edu

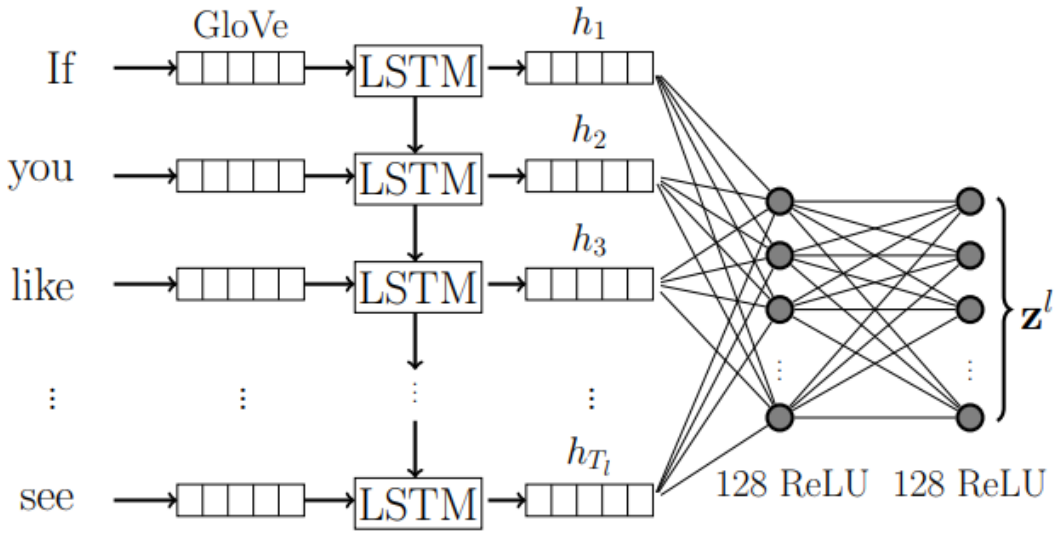
Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Related paper

Tensor fusion network for multimodal sentiment analysis, 2017

❖ Feature vector concatenation 방식 소개 논문

- 각 모달의 특징벡터는 2017년 기준의 우수한 pretrained 모델을 사용하여 특징 추출
- 언어 모델 → GloVe & LSTM, 비디오 모델 → FACET network, 오디오 모델 → COVAREP network



[언어 데이터 특징 추출 과정 도식화]

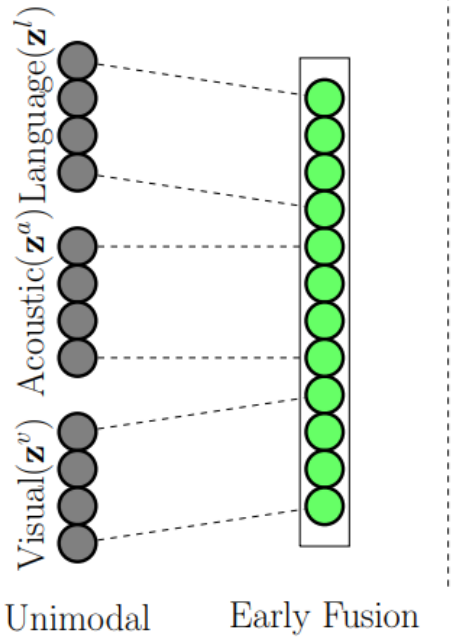
Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Related paper

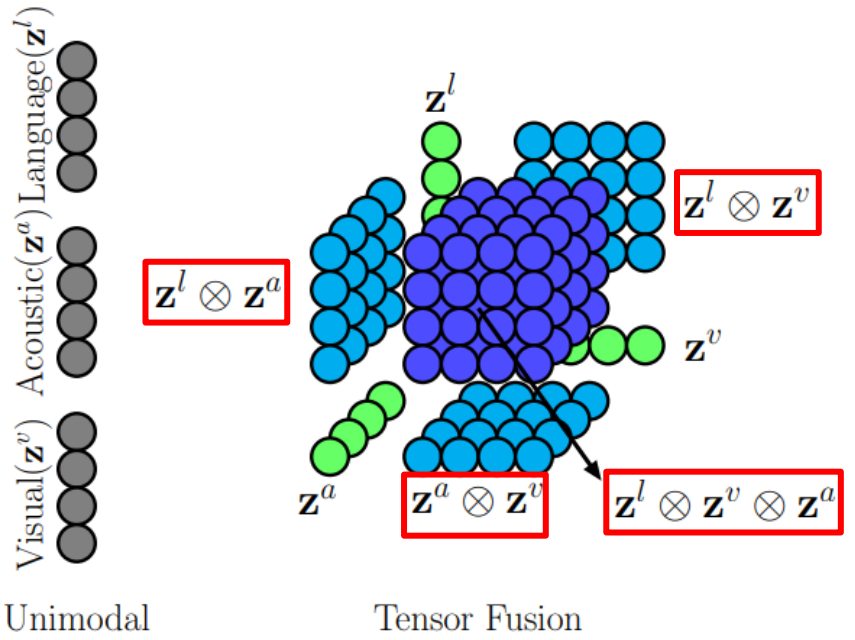
Tensor fusion network for multimodal sentiment analysis, 2017

❖ Feature vector concatenation 방식 소개 논문

- 각 모달 데이터 별로 추출된 특징벡터를 제안한 tensor fusion 방식으로 병합
- Bimodal의 특징들과 trimodal을 특징을 모두 잡아낼 수 있는 장점이 있음



[기존 단순 병합 방식]



$$\mathbf{z}^m = \begin{bmatrix} \mathbf{z}^l \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^v \\ 1 \end{bmatrix} \otimes \begin{bmatrix} \mathbf{z}^a \\ 1 \end{bmatrix}$$

[제안한 tensor fusion 방식]

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Related paper

Tensor fusion network for multimodal sentiment analysis, 2017

❖ Feature vector concatenation 방식 소개 논문

- 제안한 방법론은 감정 인식 문제에서 논문 작성 당시 SOTA 성능을 보임
- Feature vector의 병합 방식이 성능 향상에 도움을 준다는 것을 확인

Multimodal Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
Random	50.2	48.7	23.9	1.88	-
C-MKL	73.1	75.2	35.3	-	-
SAL-CNN	73.0	-	-	-	-
SVM-MD	71.6	72.3	32.0	1.10	0.53
RF	71.4	72.1	31.9	1.11	0.51
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
Human	85.7	87.5	53.9	0.71	0.82
Δ^{SOTA}	↑ 4.0	↑ 2.7	↑ 6.7	↓ 0.23	↑ 0.17

[제안한 방법론과 비교 방법론의 성능 차이]

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Related paper

Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021

❖ Feature vector concatenation 방식 소개 논문

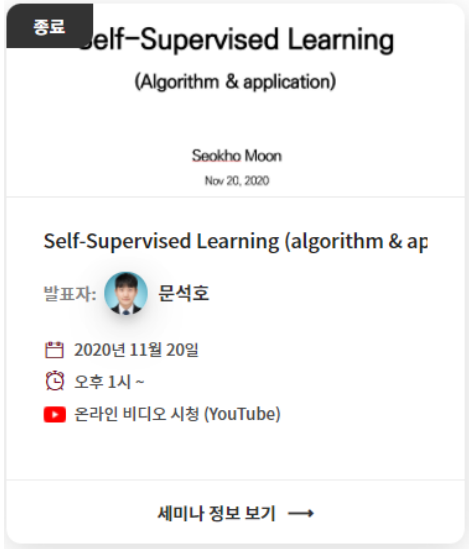
- 본 논문은 label에 없는 대규모 데이터셋에 대해서 최적의 multimodal feature를 추출하는 것을 목표로 함
- 이전 세미나에서 다뤘던 self-supervised learning 개념을 도입하여 supervision이 없는 상태에서 CNN이 아닌 transformer 기반의 multimodal 학습 방식을 제안함

VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

Hassan Akbari* [1,2], Liangzhe Yuan[1], Rui Qian* [1,3], Wei-Hong Chuang[1], Shih-Fu Chang[2], Yin Cui[1], Boqing Gong[1]

[1]Google [2]Columbia University [3]Cornell University

{lzyuan,whchuang,yincui,bgong}@google.com {ha2436,sc250}@columbia.edu {rq49}@cornell.edu



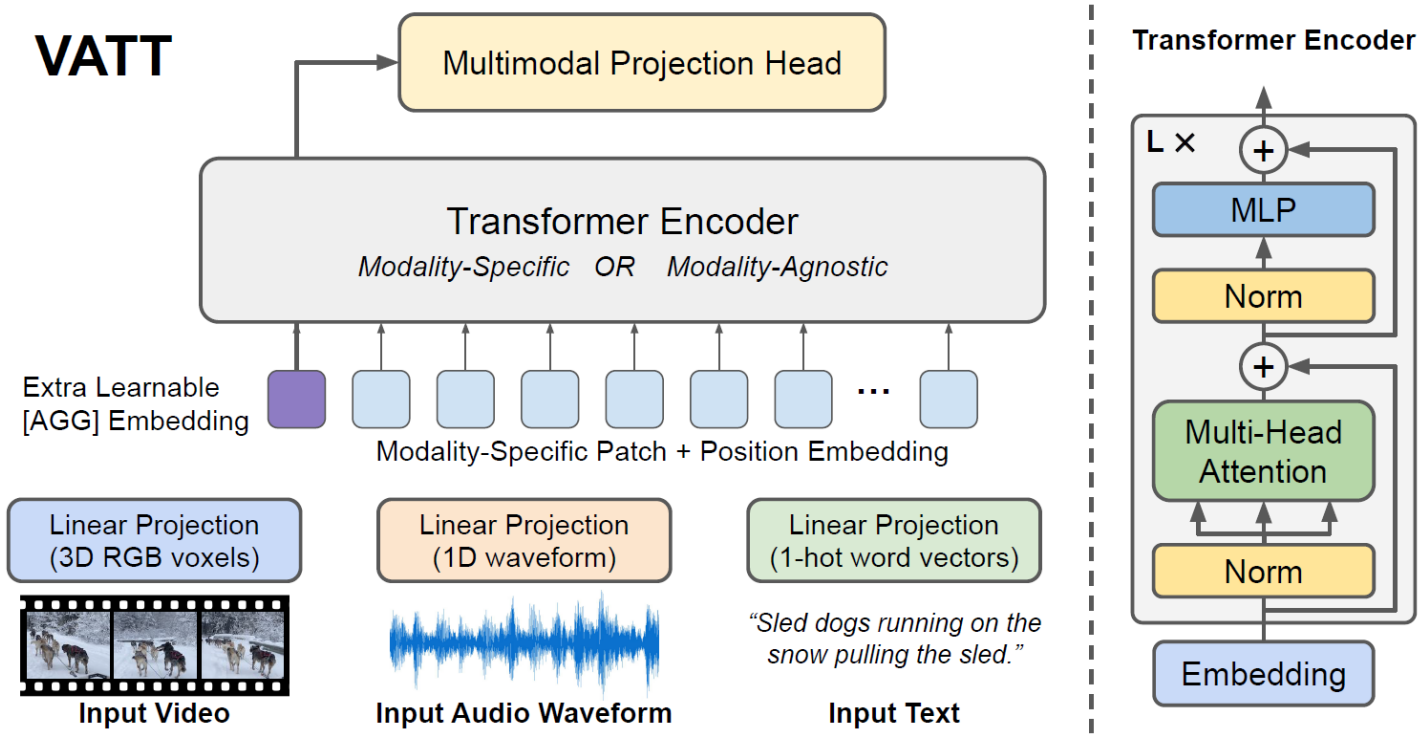
Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*.

Related paper

Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021

❖ Feature vector concatenation 방식 소개 논문

- 각 모달의 데이터는 토큰화한 후 linear projection을 통해 나온 값을 transformer enconder의 입력값으로 사용함



[VATT transformer 구조]

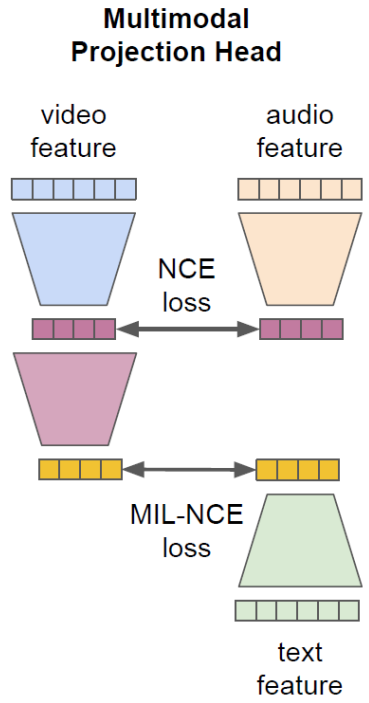
Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*.

Related paper

Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021

❖ Feature vector concatenation 방식 소개 논문

- Transformer에서 추출된 video, audio, text 특징 벡터를 contrastive learning 기반으로 학습시킴
- Video-audio pair에서는 NCE-loss를 사용하고, video-text pair에서는 MIL-NCE를 사용함



[self-supervised learning 구조]

$$\mathrm{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) = - \log \left(\frac{\exp(\boldsymbol{z}_{v,va}^{\top} \boldsymbol{z}_{a,va}/\tau)}{\sum_{i=1}^{B} \exp(\boldsymbol{z}^{i}{}_{v,va}^{\top} \boldsymbol{z}^{i}{}_{a,va}/\tau)} \right), \quad (4)$$

$$\mathrm{MIL\text{-}NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = - \log \left(\frac{\sum_{\boldsymbol{z}_{t,vt} \in \mathcal{P}(\boldsymbol{z}_{v,vt})} \exp(\boldsymbol{z}_{v,vt}^{\top} \boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt} \in \mathcal{P}(\boldsymbol{z}_{v,vt}) \cup \mathcal{N}(\boldsymbol{z}_{v,vt})} \exp(\boldsymbol{z}_{v,vt}^{\top} \boldsymbol{z}_{t,vt}/\tau)} \right), \quad (5)$$

$$\mathcal{L} = \mathrm{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda \mathrm{MIL\text{-}NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}), \quad (6)$$

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*.

Related paper

Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021

❖ Feature vector concatenation 방식 소개 논문

- 본 논문에서 제안한 방식으로 120만개의 각기 다른 unlabeled 비디오를 학습 데이터로 활용하여 학습 진행
- 이렇게 학습된 모델을 self-supervised learning 검증 방식 중 하나인 fine-tune한 모델의 성능을 확인하여 우수성 입증
- Kinetics-600에서 아래와 같은 성능을 확인

METHOD	TOP-1	TOP-5
I3D-R50+Cell [99]	79.8	94.4
LGD-3D-101 [75]	81.5	95.6
SlowFast [34]	81.8	95.1
X3D-XL [33]	81.9	95.5
TimeSFormer-HR [10]	82.4	96.0
MoViNet-A5 [51]	82.7	95.7
VATT-Base	80.5	95.5
VATT-Medium	82.4	96.1
VATT-Large	**83.6**	**96.6**
VATT-MA-Medium	80.8	95.5

Table 3. Results for video action recognition on Kinetics-600.

Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. (2021). Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*.

Conclusion

❖ Comments

- 이번 세미나를 통해 multimodal learning의 개념과 최근 연구 흐름을 살펴볼 수 있었음
- 인간 행동 인식이나 감정 인식 등의 문제에서는 단일 모달보다 멀티 모달의 데이터를 활용한 모델 구조가 더 우수함을 확인하였고, 이에 따른 다양한 연구가 진행되고 있음
- 특히, 최근에는 대규모 데이터의 labeling 비용 문제를 해결하기 위한 self-supervised learning 방식이 각광받고 있으며 멀티 모달 구조에서도 가능함을 확인
- 다양한 도메인이나 멀티 모달 데이터가 꼭 활용될 수 있는 분야로 확장 가능성