Data2Vec: A General Framework for self-supervised Learning in Speech, Vision and Language

Hate speech in pixels : Detection of Offensive Memes towards Automatic Moderation

성신여자대학교 미래융합기술공학과 이세영

2022. 01. 29

Contents

Data2vec

Hate Speech in Pixels

Data2Vec

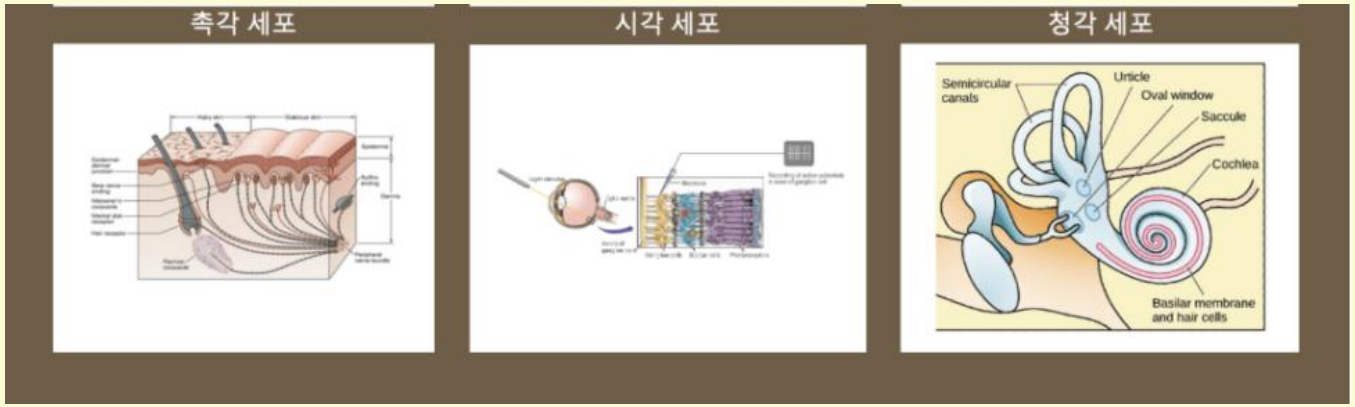
A General Framework for Self-supervised Learning in Speech, Vision and Language

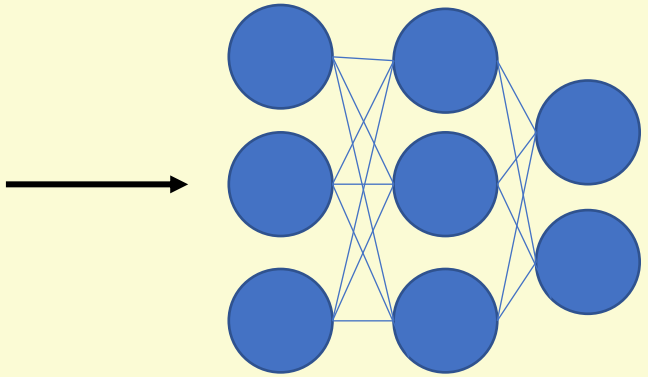
Introduction

- Modality 에 관계없이 일반적으로 잘 적용될 수 있는 Self-supervised learning 방식 제안
- Learning latent target representation: Teacher의 latent representation(Transformer 의 출력값)

  을 Student 가 예측.
- Masked prediction: masking 된 부분을 예측하는 방식으로 학습
- 실험을 진행한 Speech, vision, NLP 에 대해 SoTA 에 준하는 성능을 달성했다.

Introduction





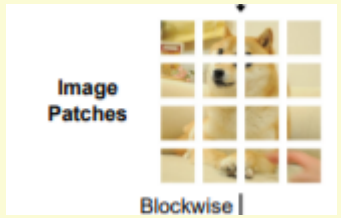
- Modality 에 관계없이 일반적으로 잘 적용될 수 있는 Self-supervised learning 방식 제안
- Learning latent target representation: Teacher의 latent representation(Transformer 의 출력값)

  을 Student 가 예측.
- Masked prediction: masking 된 부분을 예측하는 방식으로 학습
- 실험을 진행한 Speech, vision, NLP 에 대해 SoTA 에 준하는 성능을 달성했다.

Related Work

Self-supervised learning in computer vision

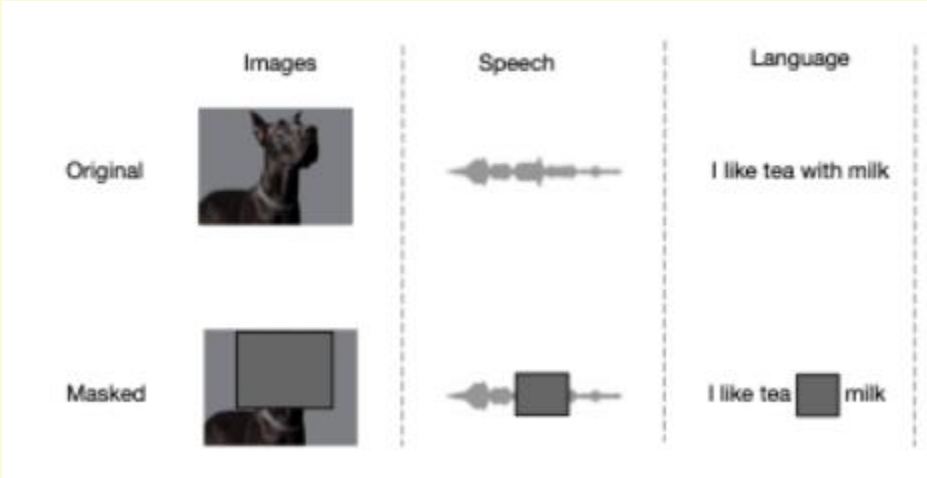
- 컴퓨터 비전 분야에서 레이블 없이 사전학습을 하는 방법은 활발하게 연구되고 있다.
  - 같은 이미지를 다르게 augmentation 한 것
  - online clustering
  - momentum encoder의 representation을 regression 하는 방식 (본 논문과 유사한 방식)
  - visual token 이나 입력 픽셀 자체를 masking 하고 예측하는 masked prediction



Related Work

Self-supervised learning in NLP

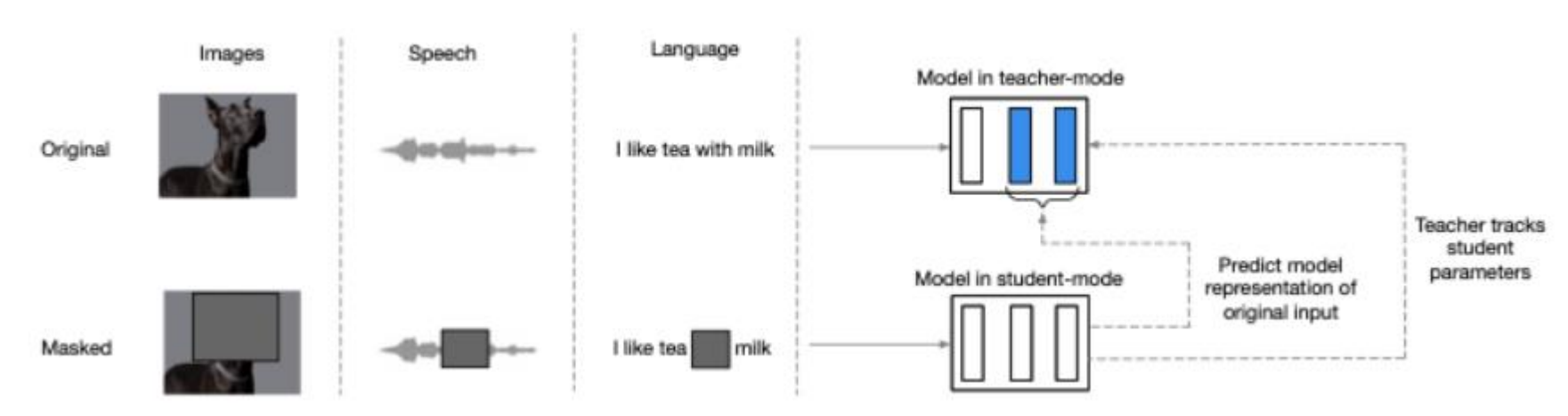
- pre-training 방법을 사용한 NLP 는 매우 좋은 성능을 보임.
- 대표적으로 BERT 의 masked language modeling



Self-supervised learning in Speech

- autoregressive (단방향), bi-directional(양방향) 모델링 모두 시도 하고 있음.
- 대표적으로 wav2vec 2.0, HuBERT 가 가장 유명한 방법.
- 음성의 discreate 한 단위를 예측하는 문제를 푼다.

Method



data2vec 의 전체적인 학습과정

1. **(Teacher mode) full input을 이용한 representation 생성**
2. **(Student mode) partial input을 이용하여 예측**

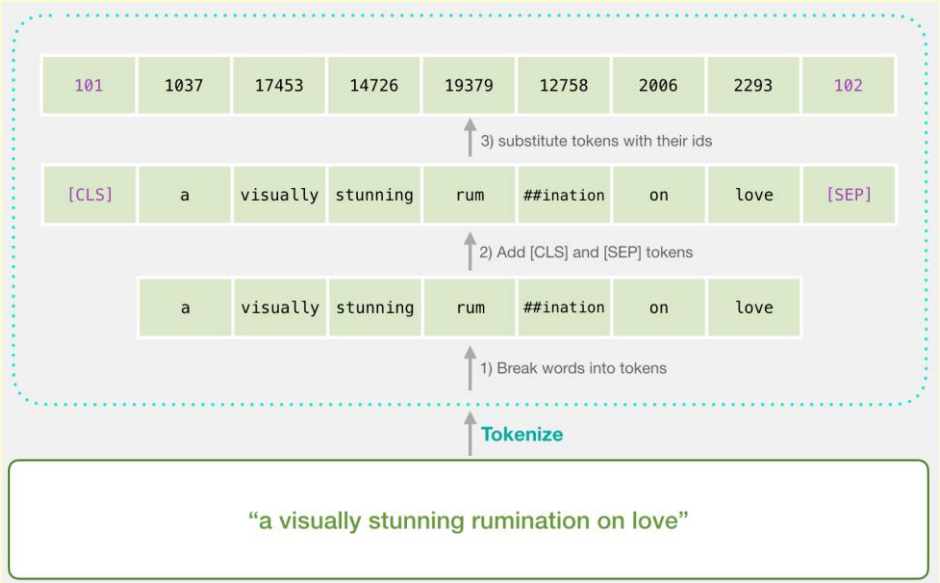
Method

Masking

Image

16*16 패치 영역의 픽셀을
하나의 토큰으로 인코딩

NLP(text)



sub word 단위로 토크나이징,
임베딩 벡터로 인코딩

Speech

multi-layer 1-D Convolution 연산을 이용해 16kHz waveform을 50Hz의 representation으로 매핑했다.

음성 데이터를 고정된 입력 사이즈로 매핑

Method

Teacher parameterization

Teacher's weight $\Delta \leftarrow \tau\Delta + (1 - \tau)\theta$

- τ 를 상수로 사용하지 않고, 학습 진행정도에 따라 스케쥴링함.
- 처음에는 랜덤으로 초기화.
- 첫 n번의 업데이트 동안 τ_0 에서 τ_e 까지는 선형적으로 증가, 남은 학습 동안은 상수로 유지
  - 학습 시작부에서는 파라미터가 랜덤 초기화되어있기 때문에, τ 를 작게하여 변화량을 더 크게 반영하고, 파라미터가 어느 정도 학습되고 난 후에는 τ를 크게하여 변화량을 적게 반영하기 위함임.

Method

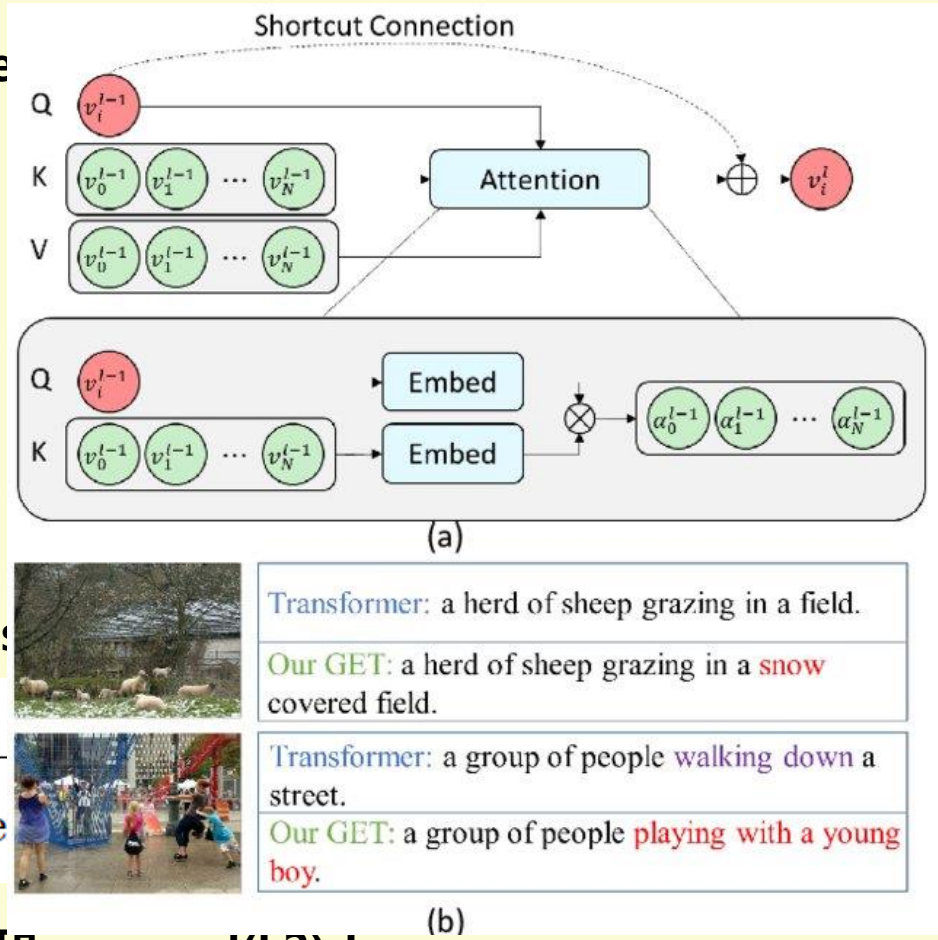
Training targets

- training target 은 L 개의 Transformer layer 로 구성된 Teache
  록 출력 값을 평균하여 만들어진다.

a_t^l : l번째 블록의 t time-step 의 출력값

1. normalization 을 수행해서 $\hat{a}_t^l$ 을 얻음.

2. L 개의 출력 값 중 K개의 출력을 평균한다. $y_t = \frac{1}{K}\sum_{l=L-K+1}^{L} \hat{a}_t^l$



Objective

- 주어진 타겟(y_t)과 Student의 출력값(f_t) 에 대해 Smooth L1 lo

$$\mathcal{L}(y_t, f_t(x)) = \begin{cases} \frac{1}{2}(y_t - f_t(x))^2/\beta & |y_t - \\ (|y_t - f_t(x)| - \frac{1}{2}\beta) & \text{othe} \end{cases}$$

- 주어진 타겟(y_t)과 Student의 출력값(f_t) 의 차이가 β 보다 작으면 squared(L2) loss
- 주어진 타겟(y_t)과 Student의 출력값(f_t) 의 차이가 β 보다 크면 absolute(L1) loss 사용
- outlier 에 덜 민감한 이점을 가짐. β 를 튜닝해야함.

Result

Computer vision

Table 1. Computer vision: top-1 validation accuracy on ImageNet-1K with ViT-B (86M parameters) and ViT-L (307M parameters) models. Our results are based on training for 800 epochs while as several other well-performing models were trained for 1,600 epochs (MAE, MaskFeat).

	ViT-B	ViT-L
MoCo v3 (Chen et al., 2021b)	83.2	84.1
DINO (Caron et al., 2021)	82.8	-
BEiT (Bao et al., 2021)	83.2	85.2
MAE (He et al., 2021)	83.6	85.9
SimMIM (Xie et al., 2021)	83.8	-
MaskFeat (Wei et al., 2021)	84.0	85.7
data2vec	84.2	86.2

ImageNet-1k 벤치마크를 이용해서 성능 비교.
다른 self-distillation 이나 masking 된 pixel 을 예측하는 방법으로 사전학습된 다른 모델들보다 좋은 성능을 보임.

Result
NLP

Table 3. Natural language processing: GLUE results on the development set for single-task fine-tuning of individual models. For MNLI we report accuracy on both the matched and unmatched dev sets, for MRPC and QQP, we report the unweighted average of accuracy and F1, for STS-B the unweighted average of Pearson and Spearman correlation, for CoLA we report Matthews correlation and for all other tasks we report accuracy. BERT Base results are from Wu et al. (2020) and our baseline is RoBERTa re-trained in a similar setup as BERT. We also report results with wav2vec 2.0 style masking of spans of four BPE tokens with no unmasked tokens or random targets.

	MNLI	QNLI	RTE	MRPC	QQP	STS-B	CoLA	SST	Avg.
Base models									
BERT (Devlin et al., 2019)	84.0/84.4	89.0	61.0	86.3	89.1	89.5	57.3	93.0	80.7
Baseline (Liu et al., 2019)	84.1/83.9	90.4	69.3	89.0	89.3	88.9	56.8	92.3	82.5
data2vec	83.2/83.0	90.9	67.0	90.2	89.1	87.2	62.2	91.8	82.7
+ wav2vec 2.0 masking	82.8/83.4	91.1	69.9	90.0	89.0	87.7	60.3	92.4	82.9

GLUE 벤치마크를 이용해서 성능 비교.
BERT와 RoBERTa에 준하는 성능을 보임.

Result

Speech

	Unlabeled data	LM	Amount of labeled data				
			10m	1h	10h	100h	960h
Base models							
wav2vec 2.0 (Baevski et al., 2020b)	LS-960	4-gram	15.6	11.3	9.5	8.0	6.1
HuBERT (Hsu et al., 2021)	LS-960	4-gram	15.3	11.3	9.4	8.1	-
WavLM (Chen et al., 2021a)	LS-960	4-gram	-	10.8	9.2	7.7	-
data2vec	LS-960	4-gram	12.3	9.1	8.1	6.8	5.5

전자책 음성 데이터인 Librispeech 이용 (960시간의 음성 데이터)
error rate 비교
wav2vec 2.0, HuBERT 보다 우수한 성능을 보임

Conclusion

- Self-supervised 학습체제가 multi-modality 에서 효과적일 수 있음을 보여줌.
- 여러 Modality 로 학습하는 방법은 향후 시청각 음성인식과 같은 작업에서 활용할 수 있음.
- 향후 연구에서는 형식에 구애받지 않고 다른 양식의 데이터에 대해 공동으로 훈련하는 단일 마스킹 전략을 연구할 수 있다.

Hate Speech in Pixels

Introduction

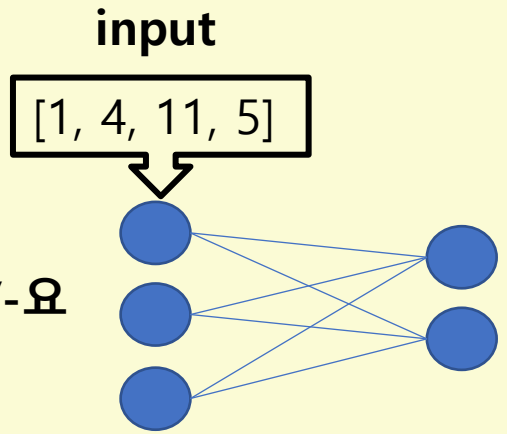
- SNS를 통한 혐오 메시지의 확산
- 혐오 발언(Hatespeech)의 경우 빠르게 발전하는 주제이고, 소셜 미디어의 트렌드에 민감
  - ➢ HS 탐지를 지속적이고 활발한 연구 주제로 만듦.
- 본 연구에서는 SNS를 통해 퍼지는 Meme 에서 시각적 + 언어적 혐오 발언 탐지

이미지

텍스트

일반적인 밈

Related Work



안녕하세요.
-> **안녕/-하-/-세-/-요**
-> 안/녕/하/세/요
-> 안녕/하다

- 혐오발언 탐지연구: 대부분 언어에 초점을 맞춰서 연구.
- 일반적으로 text를 임베딩해서 binary classifier에 공급
- Meme 에서 혐오발언을 탐지하는 연구는 없었음.
  - 언어와 시각에 대해 두 modal 에서 feature 을 추출하고 나중에 두 벡터를 통합.
  - multi-modal feature 를 classifier 에 공급

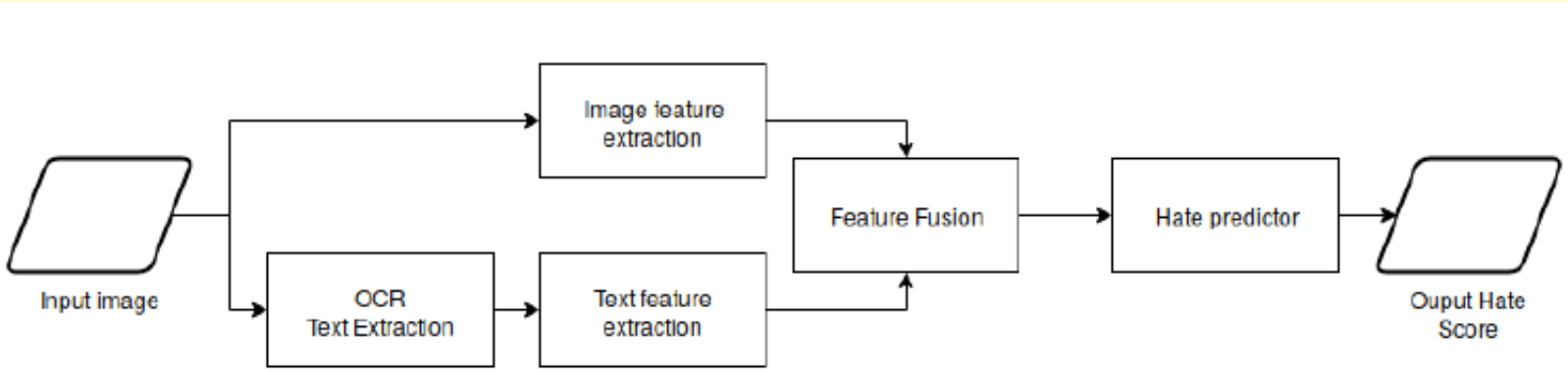
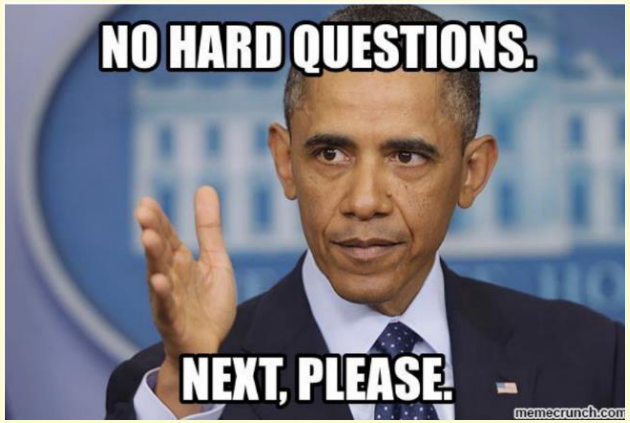


Figure 1: Block diagram of the system

Model

- **Input: Memes, Output: hate score**



4096차원

ImageNet에서 훈련된 VGG-16으로 인코딩

이미지 feature 추출

OCR로 텍스트 추출

pytesseract 사용

텍스트 feature 추출

BERT표현으로 인코딩

768차원

Feature vector 통합

4864차원 vector 생성

```
self.classifier = nn.Sequential(
    nn.Linear(TOTAL_FEATURES, hidden_size),
    nn.ReLU(),
    nn.Dropout(0.2),
    nn.Linear(hidden_size, hidden_size),
    nn.ReLU(),
    # nn.Dropout(0.5),
    nn.Linear(hidden_size, 1),
    # nn.Softmax()
)
```

Hate predictor

100개의 뉴런으로 구성된 2개의 hidden layer 가 있는 MLP.

마지막에는 Activation function 없는 단일 뉴런 추가

Hate Score

Dataset

- Hate/not hate 로 soft labeling 된 5020개 데이터셋 구축 (연구용으로만 공개)
- 혐오 밈은 Google image 에서 크롤링하여 얻음. (유대인, 무슬림 등 인종차별관련 밈)

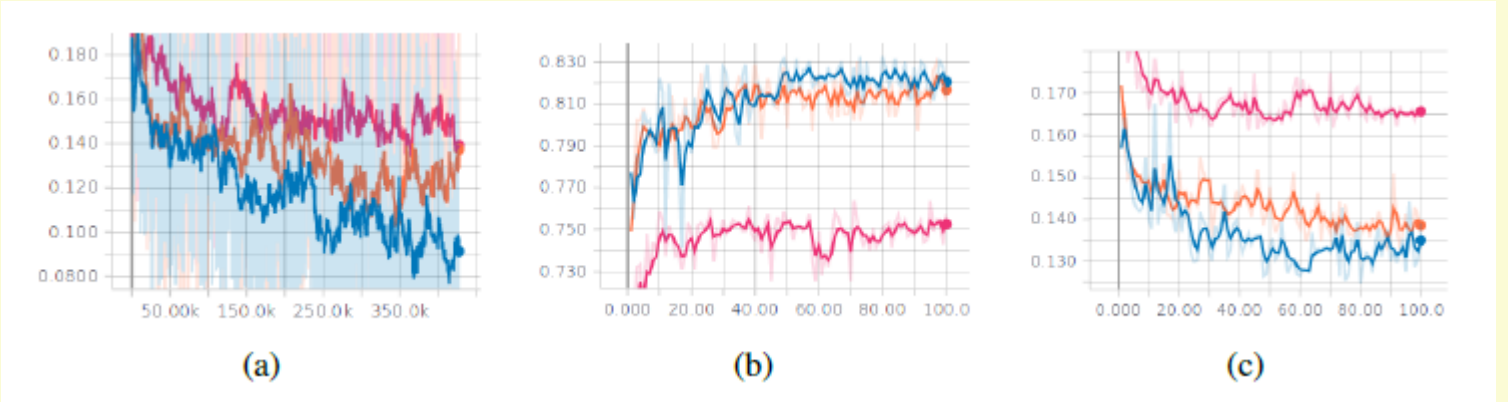
  1695개 수집: racist meme (643 memes), jew meme (551 memes), and muslim meme (501 Memes).
- 일반 밈은 Kaggle-reddit 에서 얻음. 3325개 수집
- train 4266 memes / validation 754 memes 로 분리

구글 이미지 크롤링 코드: GitHub - hardikvasa/google-images-download: Python Script to download hundreds of images from 'Google Images'. It is a ready-to-run code!
일반 밈 kaggle: Reddit Memes Dataset | Kaggle

Experiments

Multimodal(our) VS NLP VS Vision

- 네트워크는 MSE(평균제곱오차)로 훈련, 평가는 Accuracy(정확도)로 평가



NLP(빨강)
Vision(주황)
Multimodal(파랑)

Train loss　　validation loss　　Test loss

- **Test loss 를 기준으로 Multimodal 이 최상의 성능을 보이고, NLP 가 최하의 성능을 보임.**

Table 1: Accuracy results for the three configurations

Model	Max. Accuracy	Smth. Max. Accuracy
Multimodal	0.833	0.823
Image	0.830	0.804
Text	0.761	0.750

- **Accuracy 를 비교했을 때도 Multimodal이 최상의 성능을 보이고, NLP 가 최하의 성능을 보임.**

Experiments

- **NLP 모델보다 vision 모델이 더 나은 성능을 보인 이유**

  ① vision 이 더 많은 차원수로 더 많은 정보를 담고 있기 때문일 것 (vision 4096 dim + NLP 768 dim)

  ② dataset에 visual bias 가 있을 수 있음. 일반 밈은 modern meme 이 많고, 혐오 밈은 classic meme이 많기 때문.

  ③ 마지막으로, 밈은 이미지가 왜곡된 경우가 많아 OCR 인식 품질과 언어 인코딩에 영향을 주었을 수 있음.

(a) Modern meme (b) Classic Meme (c) Distorted Meme

Figure 3

Conclusion

- 제안된 방법은 간단한 네트워크 구성으로 자동화 되기에 충분하다.
- 하지만, SNS 의 일부 밈은 필터링 할 수 있지만, 여전히 많은 밈에 대한 휴리스틱 필터링이 필요하다.
- 또한, 이 시스템은 hate memes detection 에도 사용될 수 있지만, 창조에도 사용 가능하다.
- 본 연구에서는 밈에서는 언어적 표현보다 시각적 표현이 더 중요하다는 것을 보여주었다.

Code

Facebook ai blog (Meta AI)
https://ai.facebook.com/research/

Data2VecCode
https://github.com/pytorch/fairseq/tree/main/examples/data2vec

Hate speech in pixels
https://github.com/imatge-upc/hate-speech-detection

AWS 강연 (2.24 목 9시~12시, 14시~17시)
https://aws.amazon.com/ko/events/aws-innovate/machine-learning/#agenda

Thank you