

# Image Captioning, Multi Modal Image Quality Assessment & Captioning Model

**InHun's Multi Modal**

Jung Gayeon, Park Sangjun, Park Jayoung, Lee Inhun

The 10<sup>th</sup> TAVE Conference

DACON, 2023 Samsung AI Challenge : Image Quality Assessment

# Contents

**Introduction**

**Methodology**

**Data Preprocessing**

**Model Train**

**Results & Discussion**

**References**

# Introduction

주제 선정 배경



## 3회차 스터디 서기

[시작]

[종료]

스터디 목차

Chapter6 게이트가 추가된 RNN

Ch 6.1 RNN의 문제점 - 이유진

Ch6.2 기울기 소실과 LSTM - 전준석

Ch6.3 LSTM구현, Ch6.4 LSTM을 사용한 언어 모델 - 정가연

과제 한단

# “ Multi Modal ”

### 2023 Samsung AI Challenge : Image Quality Assessment

알고리즘 | 비전 | 언어 | 이미지 캡셔닝 | Custom Metric

₩ 상금 : 2,100만 원

🕒 2023.08.21 ~ 2023.10.02 09:59

+ Google Calendar

👤 394명 📅 마감



#### [ 주제 ]

화질 평가 및 Image Captioning

#### [ 설명 ]

카메라 영상 화질 정량 평가 및 자연어 정성 평가를 동시 생성하는 알고리즘 개발 대회

# Introduction

## Data 소개

<input type="checkbox"/>	img_name ▾	img_path ▾	mos ▾	comments ▾
1	41wy7upxzl	./train/41wy7upxzl.jpg	5.56923077	the pink and blue really compliment each other. like the dense color, blur.
2	ygujjq6xxt	./train/ygujjq6xxt.jpg	6.10317460	love rhubarb! great colors!
3	wk321130q0	./train/wk321130q0.jpg	5.54198473	i enjoy the textures and grungy feel to this. i also really like the deep rich red color.
4	w50dp2zjpg	./train/w50dp2zjpg.jpg	6.23484848	i like all the different colours in this pic, the brown, green, dark grey, light grey, cool image.
5	l7rqfxeuh0	./train/l7rqfxeuh0.jpg	5.19047619	i love these critters, just wish he was a little sharper, nice comp though.
6	iapcid06sr	./train/iapcid06sr.jpg	5.93846154	excellent use of light. great stuff.
7	twvec6pi41	./train/twvec6pi41.jpg	7.38931298	double trouble! what great detail and curious if you used flash on this.
8	h0wh5in2rd	./train/h0wh5in2rd.jpg	6.12307692	is this really meters from you? how lucky can you get. theyre so fuzzy!
9	j70kuiwf5y	./train/j70kuiwf5y.jpg	6.86614173	i generally dont like overprocessed images, but somehow it works here. very painterly, david hockneylike quality.
10	9qls7ros2m	./train/9qls7ros2m.jpg	5.75396825	awesome tones. great mood. beautiful stuff.

- img\_name : 이미지 파일명
- img\_path : 이미지 경로
- mos : 화질 평가 점수 (0~10, float)
- comments : 인지 화질 평가 내용을 캡처닝한 정보

**[ train Data ]**

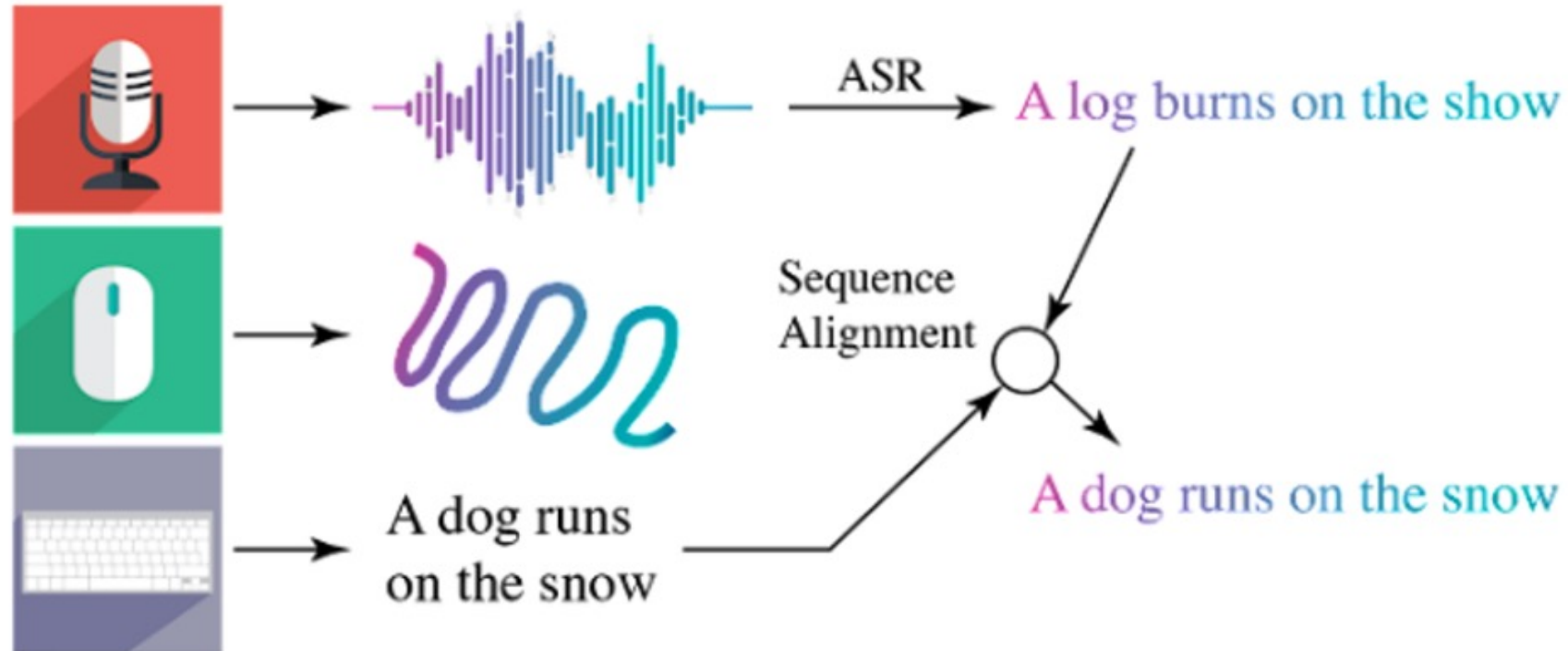
74,568 개

**[ Test Data ]**

13,012 개

# Introduction

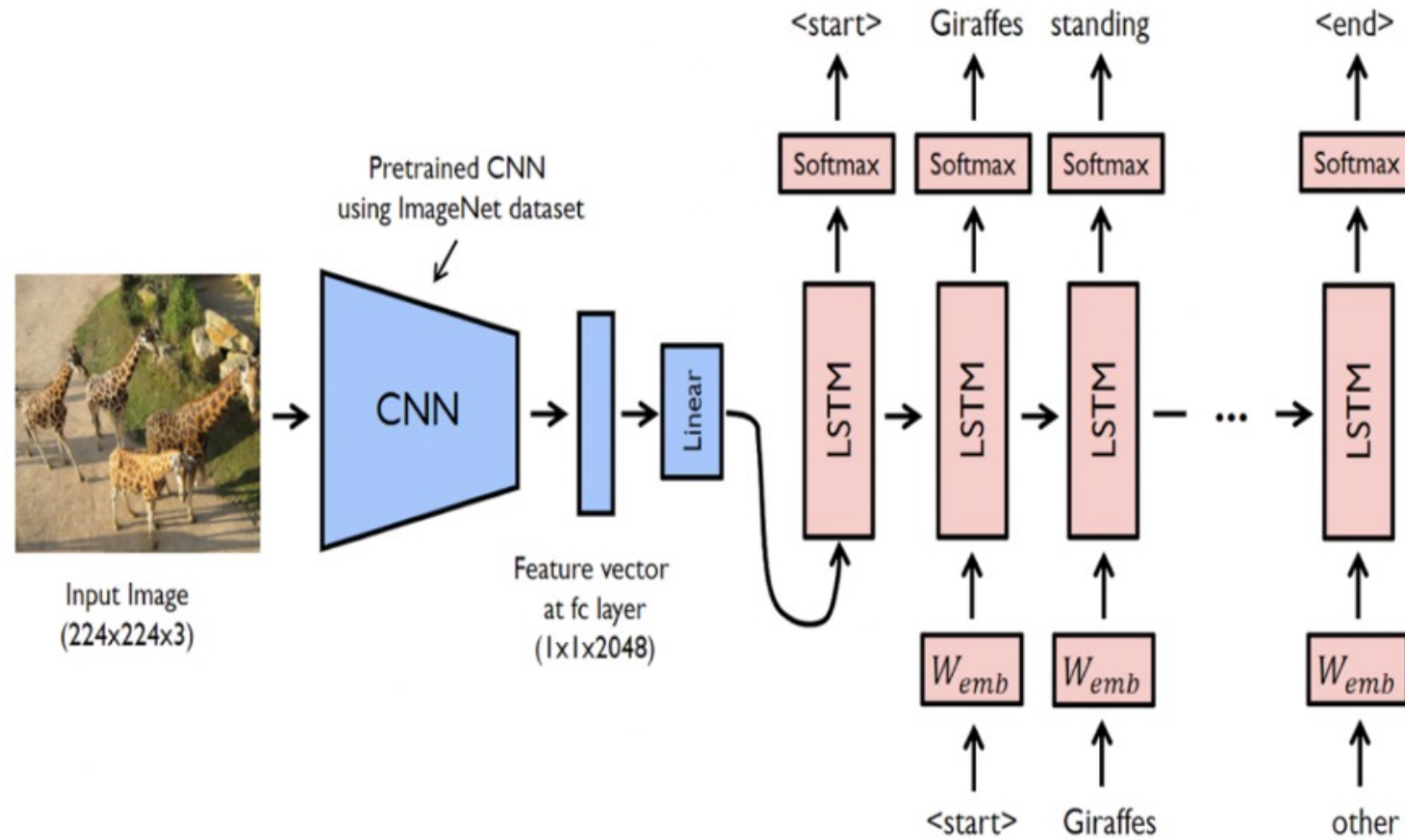
## Multi Modal



“ 서로 다른 특성을 갖는 Data Type 들을 함께 사용하는 학습법 ”

# Introduction

## Image Captioning



“ 주어진 Image에 대한 Caption을 예측하는 작업 ”

# Methodology

## ResNet + LSTM

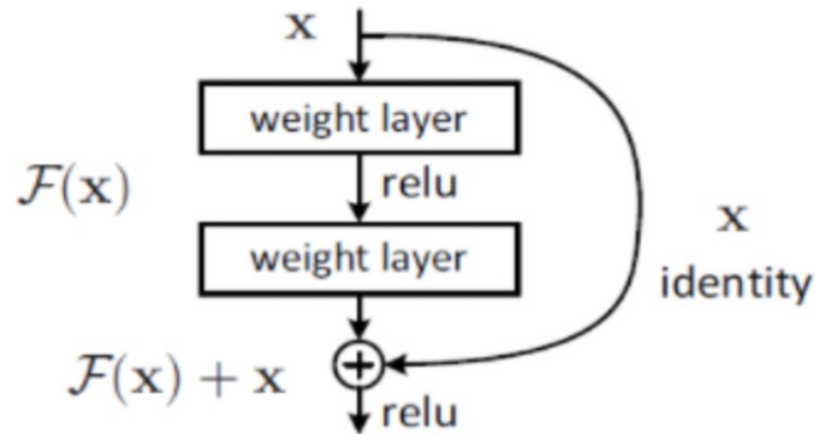
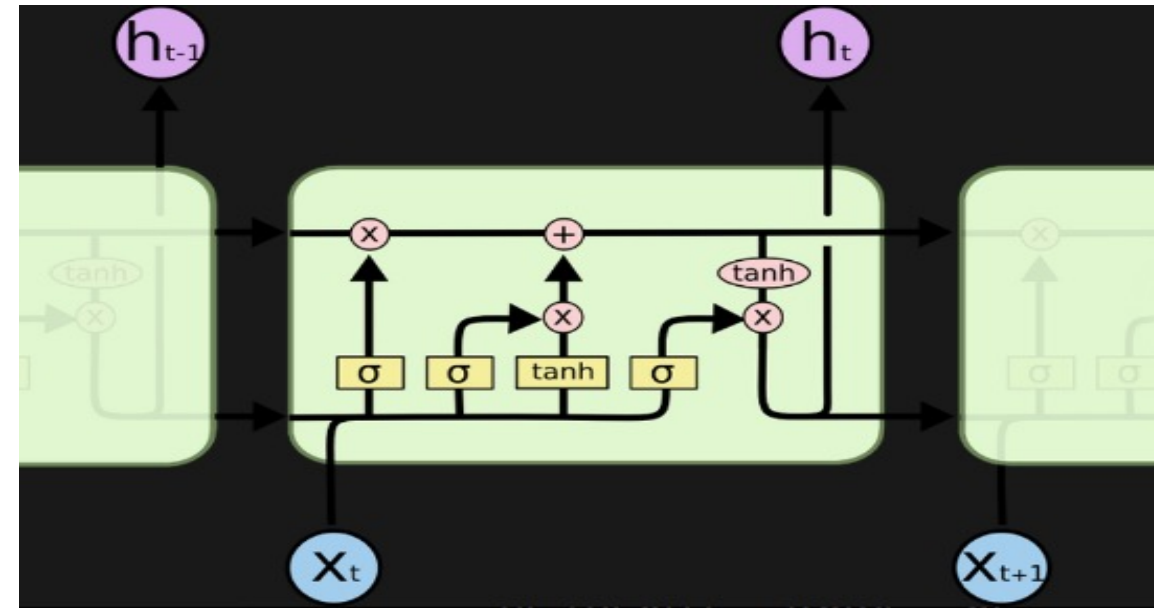


Figure 2. Residual learning: a building block.

### [ ResNet ]

**Residual Block** 구조가 쌓여 만들어진 모델.  
overfitting, gradient vanishing 문제를 해결하여  
성능을 향상시킨 모델



### [ LSTM ]

기존 RNN 모델에 cell-state가 추가된 구조  
**3개의 Gate**를 통해 메모리 값을 균일하게 유지하면서  
꼭 필요한 만큼의 정보를 기억하는 모델



# Methodology

## MobileNet, GoogLeNet

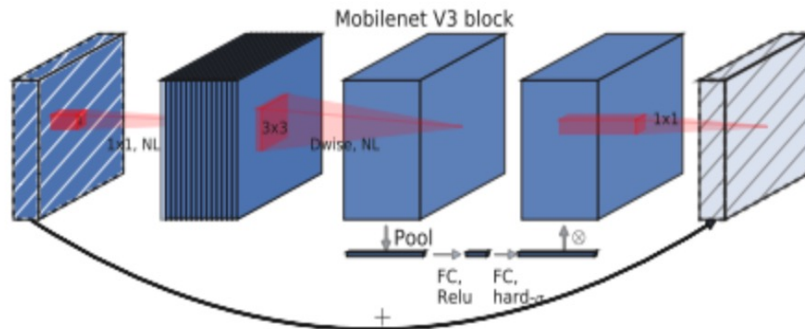


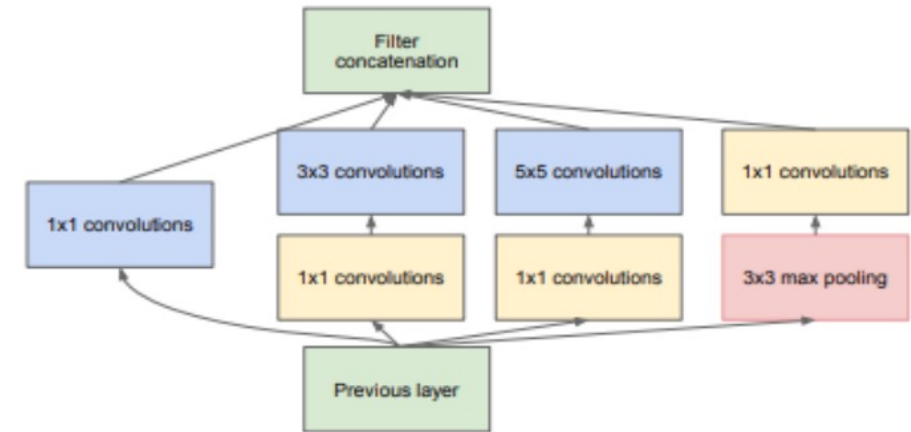
Figure 4. MobileNetV2 + Squeeze-and-Excite [20]. In contrast with [20] we apply the squeeze and excite in the residual layer. We use different nonlinearity depending on the layer, see section 5.2 for details.

### [ MobileNet ]

**Depthwise Separable Convolution** 이용

연산량이 크게 늘지 않으면서 성능 개선

: 모바일과 같은 작은 규모에서도 사용 용이



### [ GoogLeNet ]

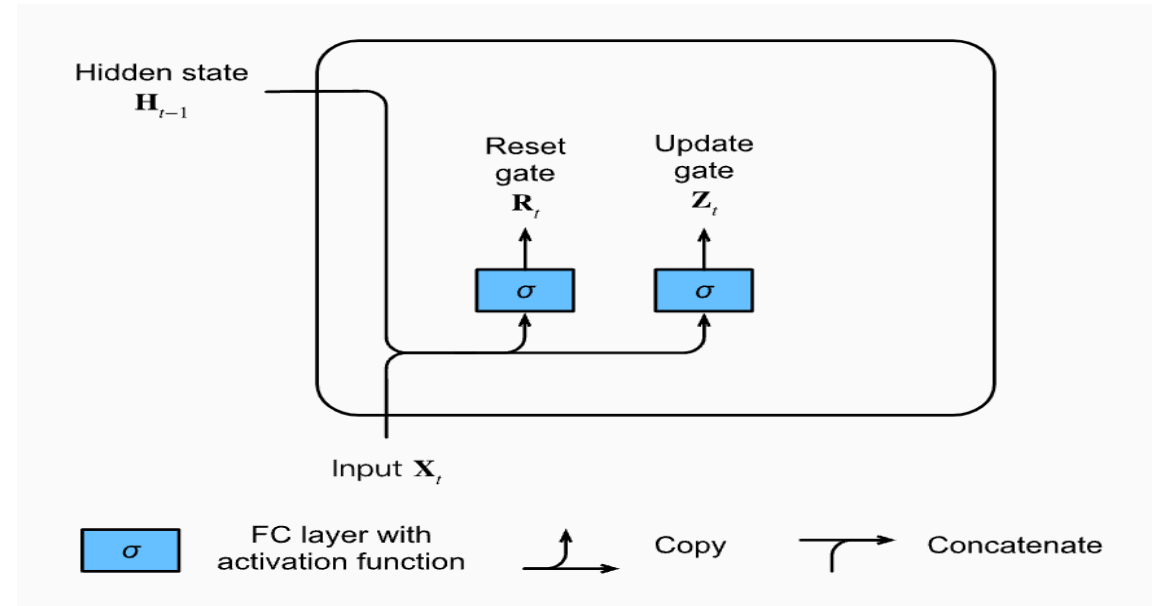
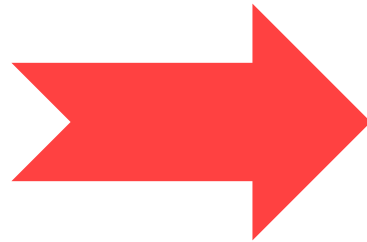
**Inception Module**(1x1 Convolution 활용)을 통한

연산량 감소 및 성능 개선

## Methodology

## GRU

[ LSTM ]



[ GRU ]

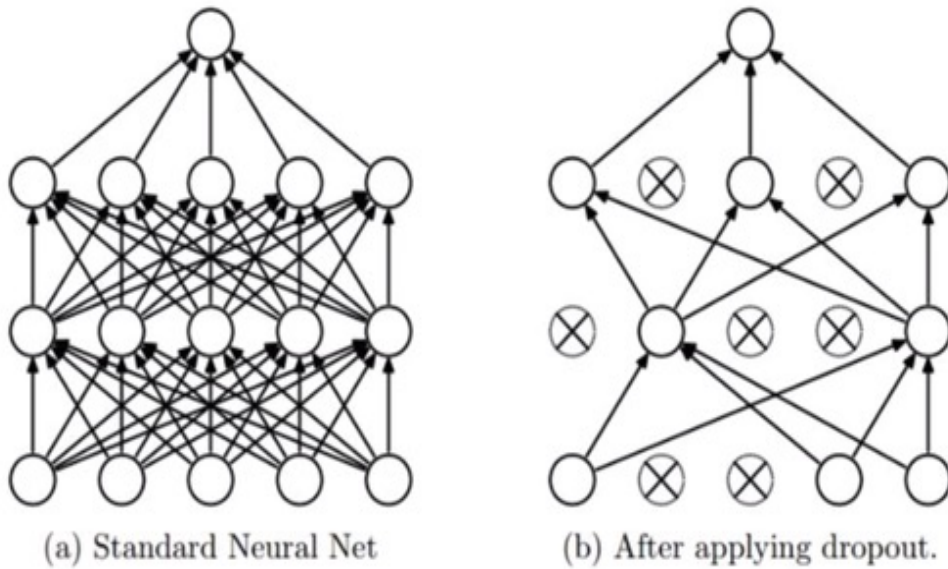
LSTM의 단점을 개선한 모델

두 개의 Gate로 효율적 계산

빠른 학습, 낮은 계산 복잡성 가짐

# Methodology

## Drop-out, Data parallelism

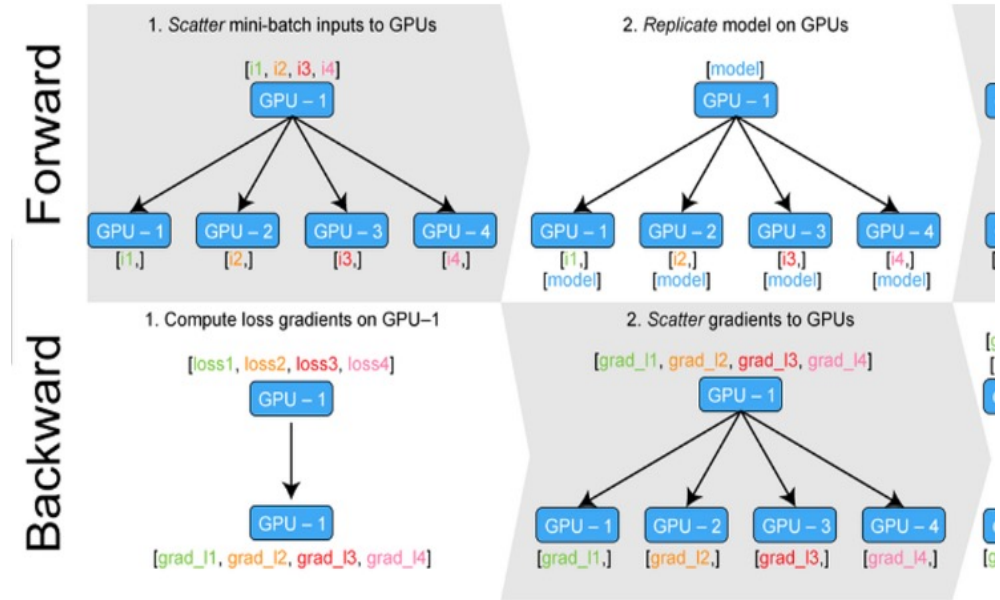


### [ Drop-out ]

Regularization의 일종

Layer에서 각각 독립적인 unit을 일정 비율에 맞춰

삭제하는 기법



### [ Data parallelism ]

다수의 GPU를 병렬적으로 활용하는 기법

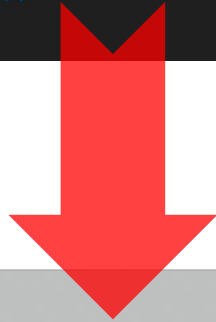
학습 속도 개선 효과

# Data Preprocessing

Color&Gray, Image Size

```
transform = transforms.Compose([
    transforms.Resize((CFG['IMG_SIZE'], CFG['IMG_SIZE'])),
    transforms.Grayscale(), # 흑백 변환 추가
    transforms.ToTensor()
])
```

**! Fail !**



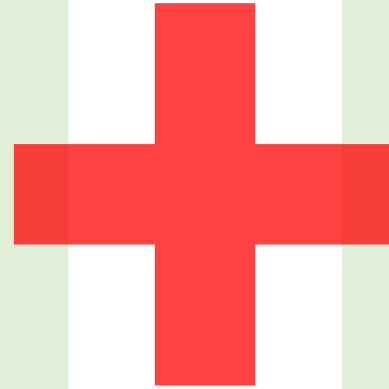
comments
done done done done done done done done done
done done done done done done done done done
done done done done done done done done done
done done done done done done done done done
more done done done done done done done done
done done done done done done done done done
done done done done done done done done done

```
CFG = {
    'IMG_SIZE': 128, # 224
    'EPOCHS': 5,
    'LR': 0.01,
    'BATCH_SIZE' : 32, #64
    'SEED': 41
}
```

**224 → 128**

**MobileNet**

**GoogLeNet**



**GRU**

**CNN**

**RNN**

Learning rate : 0.01

Epochs : 10

Batch Size : 32

64 → 32

## 모델 점수 비교

	Public Score	Private Score
ResNet + LSTM (224)	0.200708494	0.190623514
MoblieNet + GRU + Dropout	0.200708494	0.190623514
MoblieNet + GRU + Dropout + Parallel	0.200708494	0.190623514
GoogLeNet + GRU + Dropout + Parallel	0.199509431	0.190258485

## 모델 학습 시간 비교

	1 epoch 당 평균 소요 시간	전체 학습 시간
ResNet + LSTM (224)	17분	3시간
MoblieNet + GRU + Dropout	13분 50초	2시간 30분
MoblieNet + GRU + Dropout + Parallel	13분 40초	2시간 30분
GoogLeNet + GRU + Dropout + Parallel	14분 20초	2시간 35분

정확한 값이 아닌 대략적인 수치임을 알려드립니다.

# Result & Discussion

## 한계 및 개선방향

### [ 한계점 ]

#### 1. GPU 자원의 한계

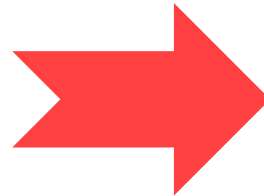
: 다른 SoTA 모델 다뤄보지 못함

: 추가 실험 진행 불가

#### 2. 모델 성능 < 모델 경량화

: 성능보다 경량화에 비중을 맞춘 진행

(모델 선택, 이미지 크기 등)



### [ 개선방향 ]

#### 1. 파라미터 조정


: 각 모델에 따라 최적의 파라미터 값 서치

#### 2. 모델 이분화

: 화질 평가와 Captioning 모델 분리



의의



TAVE  
Deep Learning Project

인 혼 이 네

**Multi Modal**

- CO (최종)Googlenet+GRU\_128\_dropout.ipynb 👤
- CO (최종)Mobile+GRU\_128\_dropout\_parallel.ip... 👤
- CO (최종)Mobile+GRU\_128\_dropout.ipynb 👤
- CO (최종)Mobile+GRU\_224\_dropout\_흑백.ipynb 👤
- CO (최종)Mobile+GRU\_224\_dropout.ipynb 👤

1. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).
2. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014). Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632.
3. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
4. ResNet 구조 이해 및 구현, <https://wjunsea.tistory.com/99>
5. RNN과 LSTM을 이해해보자!, [https://ratsgo.github.io/natural\\_language\\_processing/2017/03/09/rnnlstm/](https://ratsgo.github.io/natural_language_processing/2017/03/09/rnnlstm/)
6. PyTorch Multi-GPU 제대로 학습하기, <https://medium.com/daangn/pytorch-multi-gpu-학습-제대로-하기-27270617936b>
7. Multi-modal Learning, <https://velog.io/@ysw2946/9.-Multi-modal-Learning>
8. Image-Quality-assessment, <https://github.com/kjae0/image-quality-assessment>
9. MobileNet – V3 논문 리뷰, [https://velog.io/@pre\\_f\\_86/MobileNet-V3-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0](https://velog.io/@pre_f_86/MobileNet-V3-%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0)
10. Gated Recurrent Units (GRU), [https://d2l.ai/chapter\\_recurrent-modern/gru.html](https://d2l.ai/chapter_recurrent-modern/gru.html)