NeuralNet Advanced

20기 정규세션

19기 한진모

Contents

Unit 00

Introduction

Unit 00 | Introduction

20기 정규세션
TOBIG'S 19기 한진모

머신러닝의 목표

- Training Error의 최소화보다, Generalization Error의 최소화
- Generalization Error = (Training Error) + (Generalization Error와 Training Error의 차)
- Generalization Error의 최소화는 Training Error와 Generalization – Training Error의 차를 최소화하는 것

Optimization vs Regularization

- Training Error와 Generalization-Training Error의 차 최소화는 모델의 복잡도 면에서 서로 모순되는 목표임
- 전자를 최소화하는 과정을 Optimization, 후자를 최소화하는 과정을 Regularization이라고 함
- 두 과정 사이 황금균형을 찾아야 Generalization Error 최소화 가능

Unit 00 | Introduction

How to Enable Learning in "Deep" Layers?

- Layer를 깊게 쌓으면 Loss Function 부근의 Gradient가 최초의 층까지 제대로 전달되지 않을 수 있음
- Gradient Vanishment 혹은 Gradient Explosion으로 인해 학습이 제대로 되지 않을 수 있음
- 이를 방지하기 위해 Weight Initialization에 신경쓰거나, Batch Normalization 등의 기법을 활용

Unit 01

Optimization

Unit 01 | Optimization

20기 정규세션
.
TOBIG'S 19기 한진모

Deep Learning Optimization: Minimizing Training Error via Iterative Method

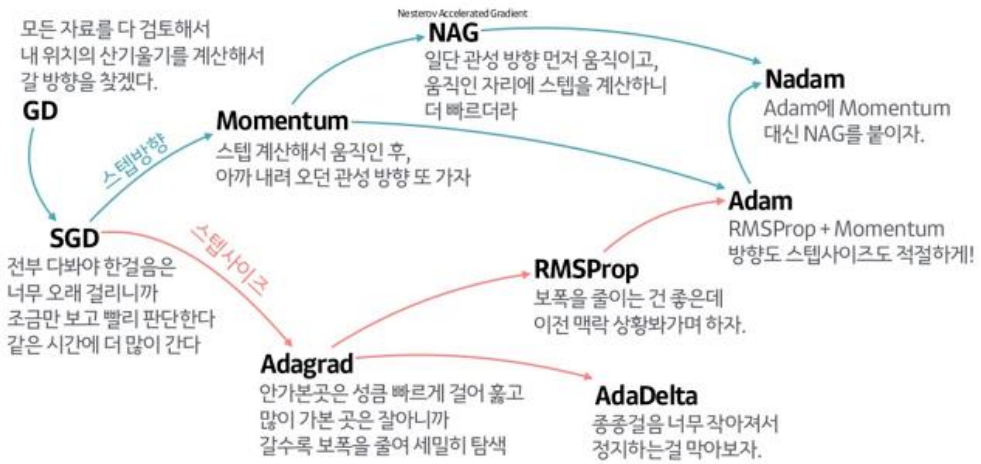
- 머신러닝은 데이터를 잘 설명하는 함수를 모델링하는 과정임
- 딥러닝은 함수를 모델링하는 한 가지 방법이며, 딥러닝의 Optimization은 최적의 가중치를 탐색하는 것
- 분석적 방법(미분해서 0이 되는 가중치 벡터를 바로 계산)은 역행렬의 연산비용이 비싸므로 비실용적
- 따라서 비교적 저렴한 Iterative Method(SGD)를 사용하여 '최적에 가까운' 가중치 벡터를 탐색함

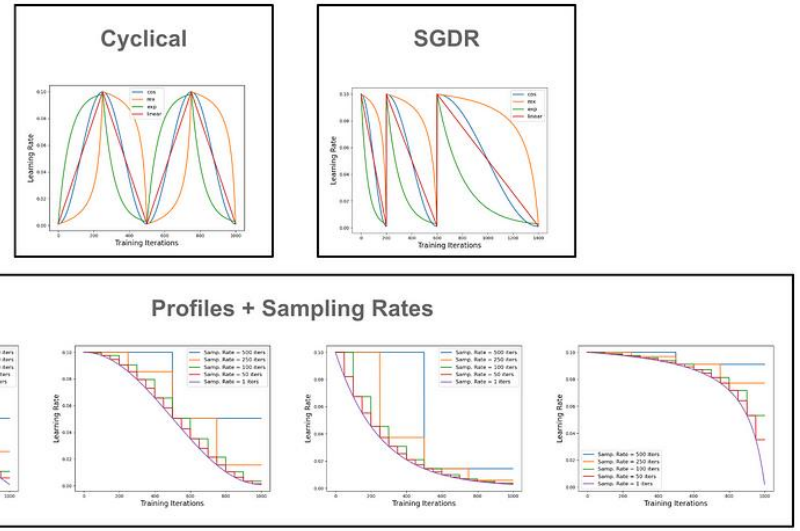
Unit 01 | Optimization

20기 정규세션
TOBIG'S 19기 한진모

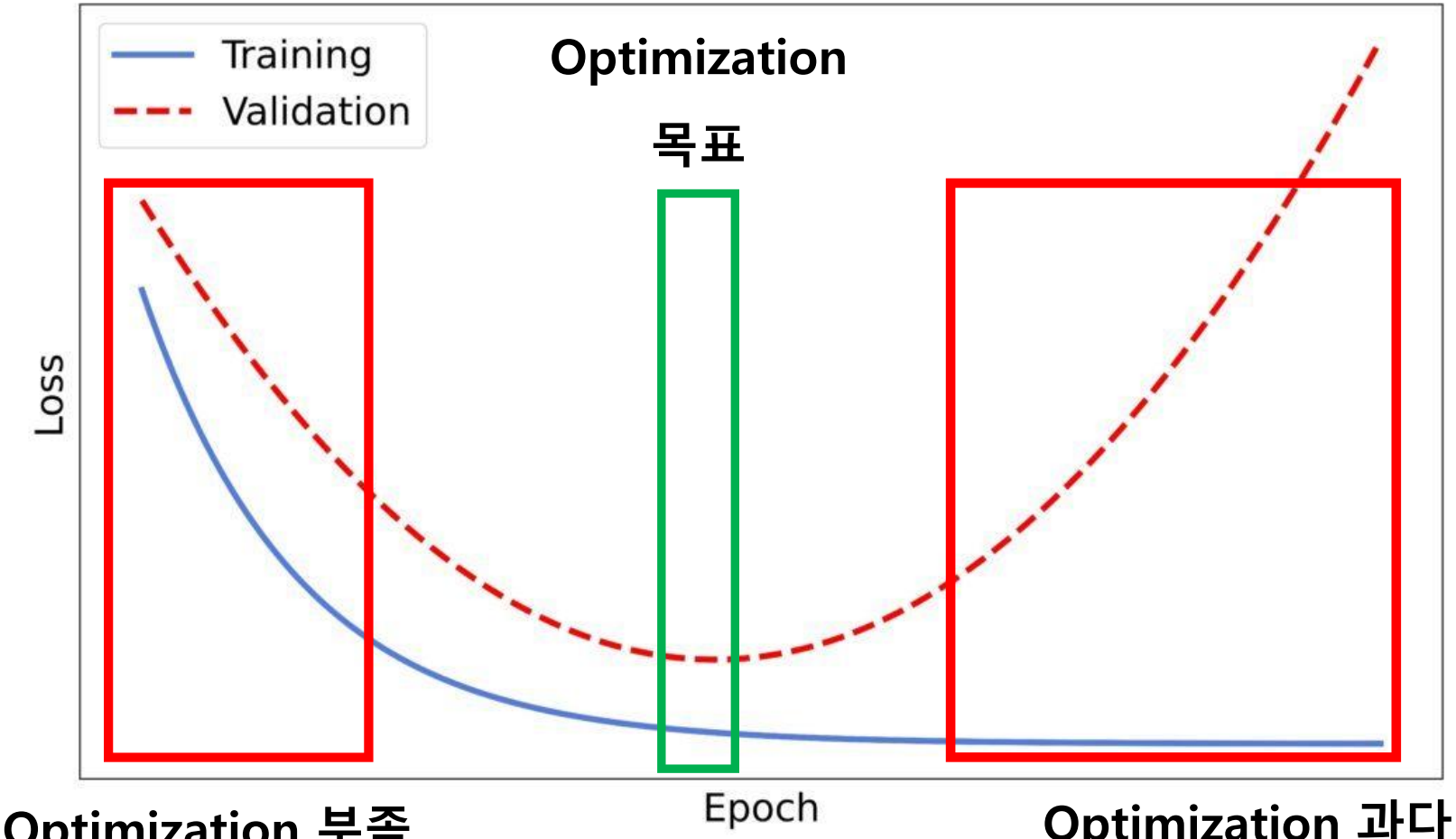
Adjustment of Optimizer

- SGD를 바탕으로 다양한 Variation이 존재(Momentum Approach, Adaptive Gradient, Adam...)
- 좋은 Learning Rate를 선택하는 것은 학습 속도를 높이는 데 가장 중요한 요인 중 하나
- 고정된 Learning Rate가 아닌, Learning Rate를 갈수록 줄이는 Learning Rate Scheduling도 대두함





Unit 01 | Optimization



Optimization 부족

: Underfitting

Optimization 과다

: Overfitting -> 규제 필요

20기 정규세션

TOBIG'S 19기 한진모

Unit 02

Regularization

Unit 02 | Regularization

20기 정규세션
TOBIG'S 19기 한진모

Regularization: Overfitting 해소를 통한 Generalization 성능 향상 도모

- 지나치게 복잡한 모델 혹은 과도한 Optimization은 모델이 Training Dataset의 noise까지 학습하도록 함
- 이는 Training 성능과 Generalization 성능의 괴리를 야기
- 따라서 모델이 Training Dataset에 대하여 지나치게 최적화되는 것을 규제할 필요가 있음

Several Regularization Approaches

- Regularization via Loss Function – Adding Penalty Term
- Regularization via Model Architecture - DropOut
- Regularization via Dataset – Data Augmentation

Unit 02 | Regularization

Regularization via Loss Function(Adding Penalty Term to Model Complexity)

- 기본적으로 딥러닝의 Training Phase는 Training Error를 최소화하는 것을 목적으로 함
- Regularization 적용 시 Augmented Error: (Training Error) + (Model Complexity) 를 최소화함

Several Augmented Errors

- L1 Norm Regularizer: Model Complexity의 척도로 가중치의 L1 Norm 채택 $Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y_i}) + \frac{\lambda}{2}|w|\}$

  Sparse Model에 비교적 용이

  Gradient-based Learning에 부적합(미분 불가능)

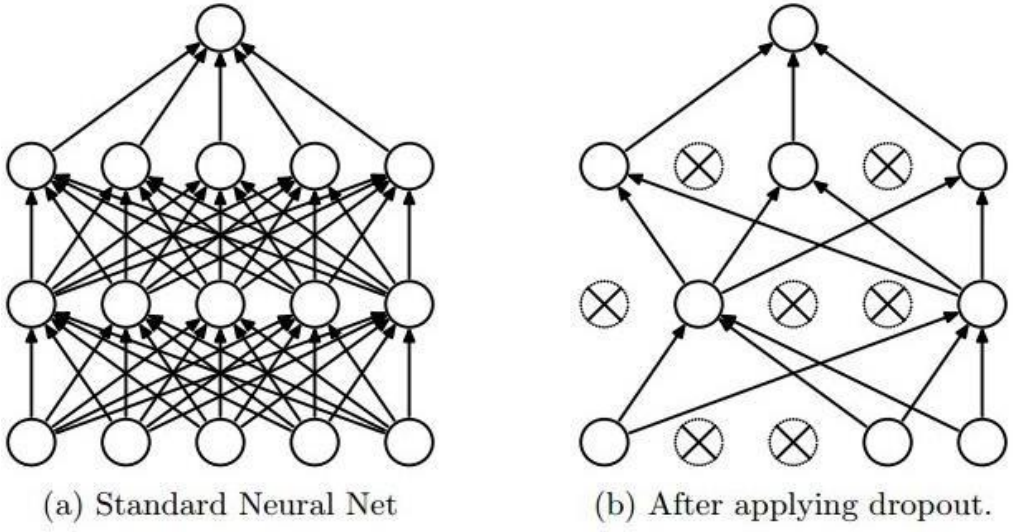
- L2 Norm Regularizer: Model Complexity의 척도로 가중치의 L2 Norm 채택 $Cost = \frac{1}{n}\sum_{i=1}^{n}\{L(y_i, \hat{y_i}) + \frac{\lambda}{2}|w|^2\}$

  L2 Norm은 미분이 가능함

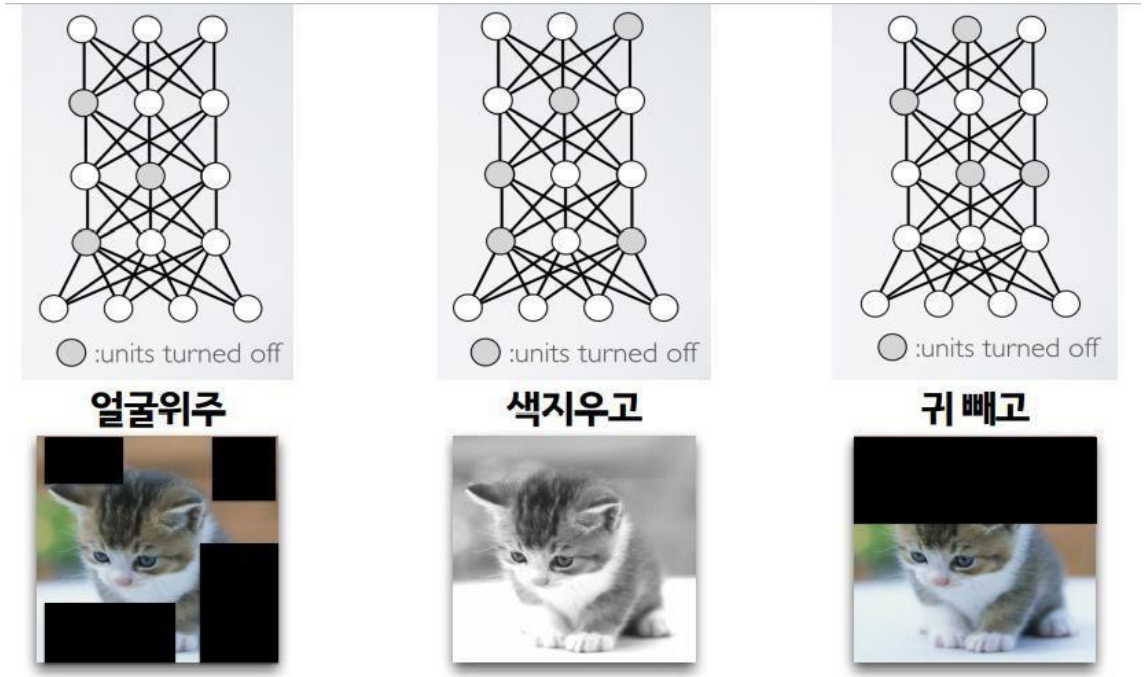
Unit 02 | Regularization

Regularization via Architecture(Dropout)

- Dropout: Training Phase마다 랜덤하게 일부 (0.1, 0.2 등의 비율로) 뉴런을 끈 채 학습시키는 기법
- Inference Phase에선 모든 뉴런을 다시 켜므로 본질적으로 Training Phase – Inference Phase간 차이 존재
- Dropout은 모델이 일부 정보에 집착하지 않고 융통성을 기를 수 있도록 함
- Dropout은 본질적으로 Ensemble로 볼 수 있음(같은 데이터셋에 대하여 서로 다른 모델을 만들어내 추론)


(a) Standard Neural Net

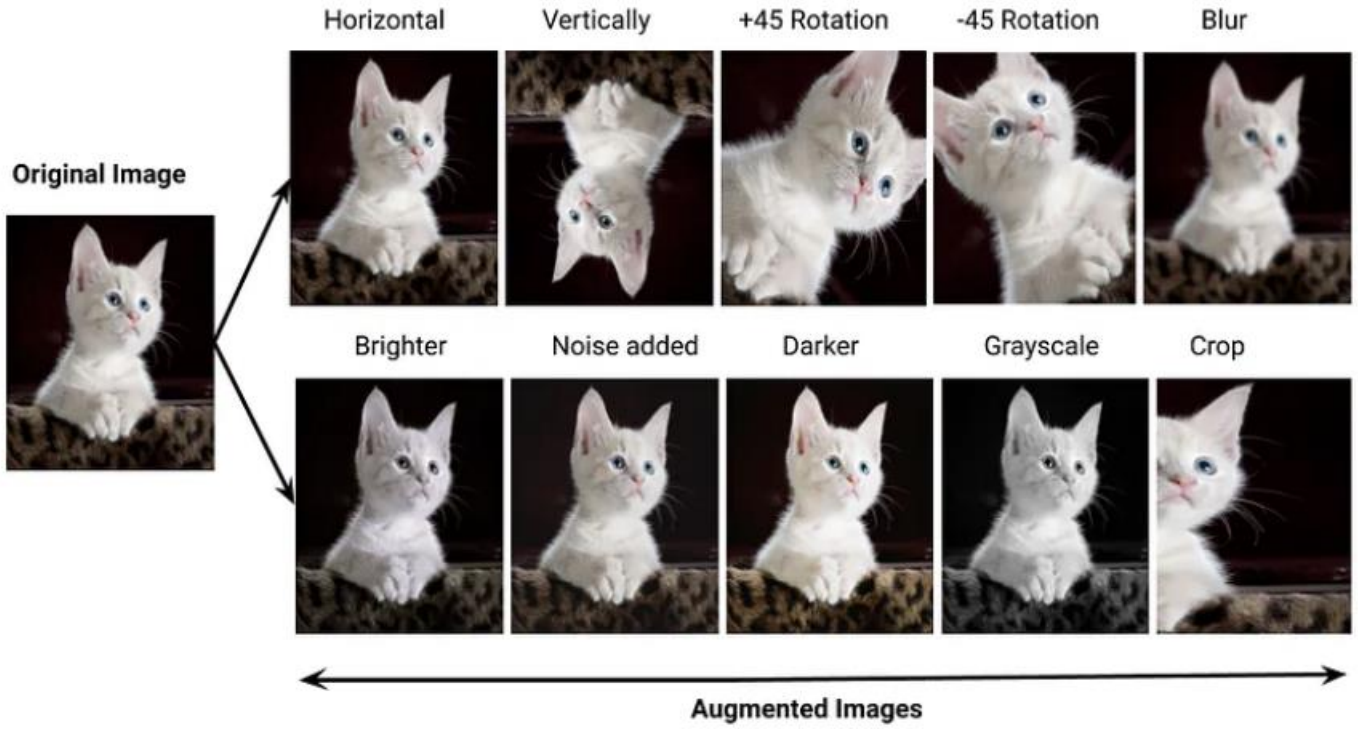
(b) After applying dropout.



20기 정규세션
TOBIG'S 19기 한진모

Regularization via Dataset(Data Augmentation)

- Data Augmentation: 주어진 Training Data에 변형을 주어 Dataset의 크기와 다양성을 키우는 기법
- 모델이 기존의 Training Dataset에만 특화되는 것을 방지하고, Robustness를 키우는 데 도움이 됨



Unit 03

Initialization

Unit 03 | Initialization

Weight Random Initialization

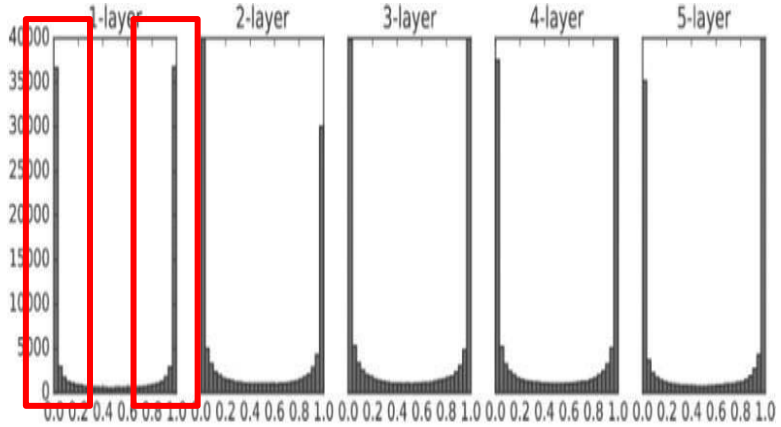
- 최적의 가중치를 찾는 데 Iterative Method를 적용하려면, Weight의 초깃값이 필요
- 초깃값이 모두 같으면 역전파 시 모든 가중치 값이 똑같이 갱신되므로 가중치 여러 개를 갖는 의미가 없음
- 따라서 All-one/zero Initialization보단 Random Initialization이 적절함
- 일반적으로 Zero-mean Gaussian Distribution 사용

Unit 03 | Initialization

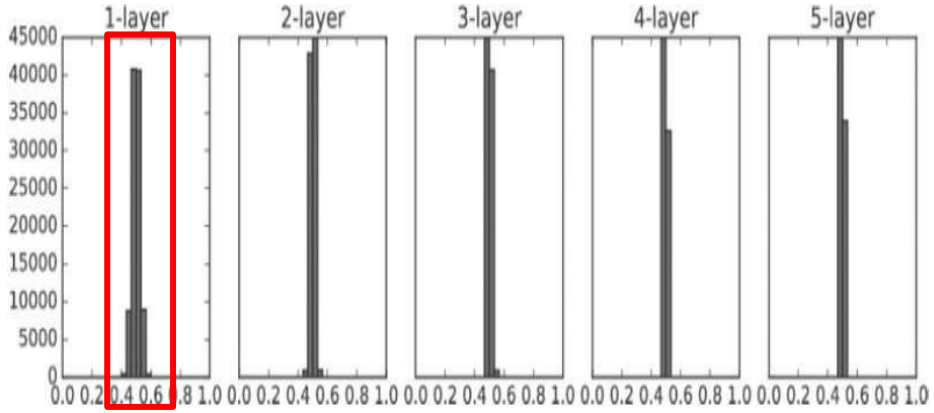
It's variance that weighs

- Gaussian의 Variance가 너무 크면 기울기가 소실되고, 너무 작으면 표현력이 제한되어 학습이 어려움
- 정보량을 보존하는 적절한 Variance의 선택이 Weight Initialization의 핵심

<Sigmoid 출력값>



Mean = 0, Std = 1: 기울기 소실



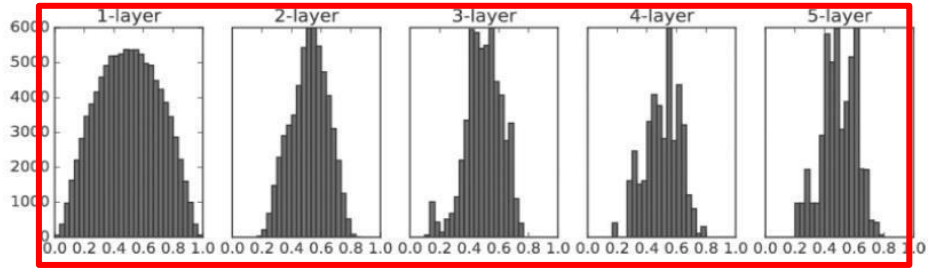
Mean = 0, Std = 0.01: 표현력 제한

Unit 03 | Initialization

20기 정규세션
TOBIG'S 19기 한진모

Xavier Initialization

- 주로 Sigmoid, Tanh와 함께 사용
- Xavier Normal Initialization과, Xavier Uniform Initialization이 있음



Xavier Normal Initialization

$$W \sim N(0, Var(W))$$

$$Var(W) = \sqrt{\frac{2}{n_{in} + n_{out}}}$$

Xavier Uniform Initialization

$$W \sim U(-\sqrt{\frac{6}{n_{in} + n_{out}}}, \ +\sqrt{\frac{6}{n_{in} + n_{out}}})$$

n_{in}:이전 layer(input)의 노드 수
n_{out}:다음 layer의 노드 수

Unit 04

Normalization

Unit 04 | Normalization

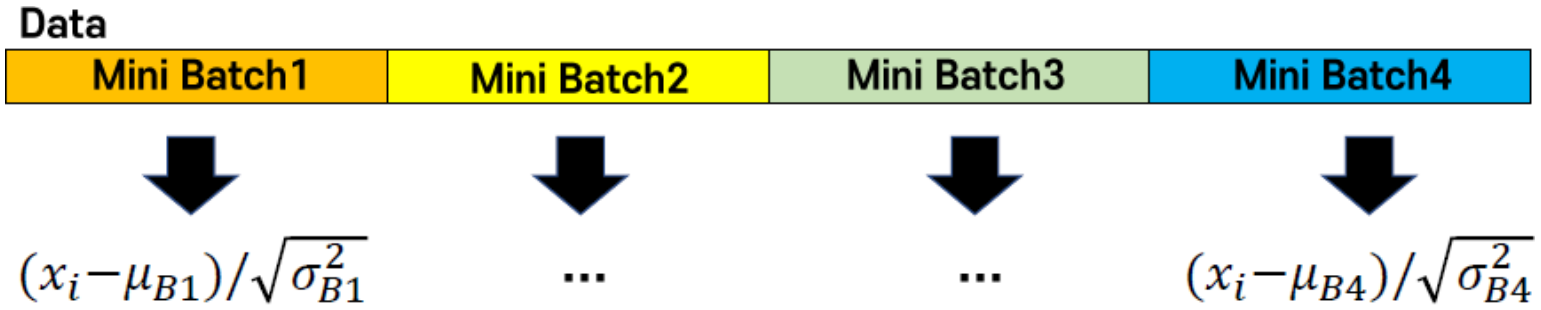
20기 정규세션
TOBIG'S 19기 한진모

Covariance Shift

- Internal Covariance Shift: Network의 각 Layer를 통과할 때마다 Input의 Distribution이 달라짐
- 이를 해결하기 위해 각 Layer를 통과할 때마다 Distribution을 일정하게 유지해줌(Normalization)

What is Batch Normalization?

- 활성화함수 통과 전 Minibatch 단위로 Normalize함으로써 Data의 분포를 일정하게 만드는 것
- 기울기 소실/팽창 방지, 자체적 Regularization 효과로 인해 dropout을 안 써도 된다는 등의 장점이 있음



Unit 04 | Normalization

Batch Normalization – Detail

- Normalization은 필연적으로 Data의 표현력을 저해함
- 이를 보정하기 위해 Normalization 후 'Scale and Shift'를 함으로써 표현력 회복
- 'Scale and Shift' Parameter는 Learnable Parameter

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

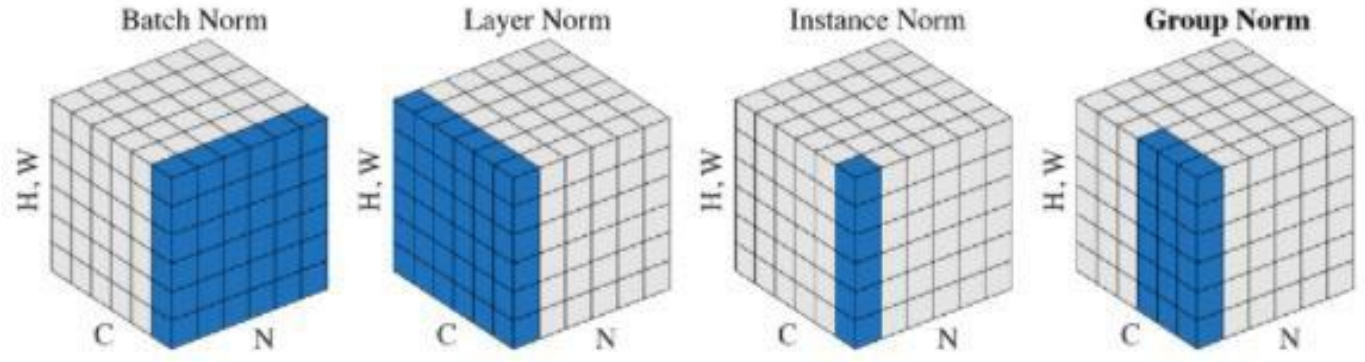
$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Unit 04 | Normalization

다양한 Normalization

- Weight Normalization(가중치에 대한 정규화)
- Layer Normalization(Feature 차원 정규화)
- Instance Normalization(채널 차원 정규화)
- Group Normalization(Layer Normalization + Instance Normalization)



Assignment

Assignment

이론

- 다음의 내용을 주제로 하는 보고서를 자유 형식으로 작성

  1) 다양한 Learning Rate Scheduler 중 적어도 한 가지를 소개 (A4 반 페이지 내외)

  2) 오늘 수업에서 소개되지 않은, Training Error와 Generalization Error 사이 간극을 줄이는 방안 (A4 반 페이지 내외)

실습

- 제공된 ipynb 파일의 주석 및 마크다운 지시를 통해 과제를 수행
- 적절한 Optimization, Regularization, Initialization, Normalization 전략을 각각 최소 1개씩 수행하여
- 15 에포크 이내에 제공된 데이터셋에서 Validation Accuracy 80%를 넘기는 것이 목표
- Ipynb 파일 마지막에 본인이 수행한 전략과 본인만의 분석이 담긴 나름의 '결론부' 작성
- 기본적으로 실습과제 성능 순으로 우수과제를 선별
- 이론 과제 및 실습 결론부가 인상적일 경우 가산점 부여

Reference

Learning Rate Scheduler 이미지:

https://towardsdatascience.com/the-best-learning-rate-schedules-6b7b9fb72565

Train vs Validation Loss 이미지:

https://www.baeldung.com/cs/loss-vs-epoch-graphs

L1/L2 Norm 공식:

https://light-tree.tistory.com/125

Dropout, Weight Initialization, Normalization 관련 자료:

TOBIG'S 19기 정규세션 자료 참고

Data Augmentation 이미지:

https://medium.com/@tagxdata/data-augmentation-for-computer-vision-9c9ed474291e

감사합니다