Comparative Analysis of Weight Distributions Across Layers

2021203098
연재혁

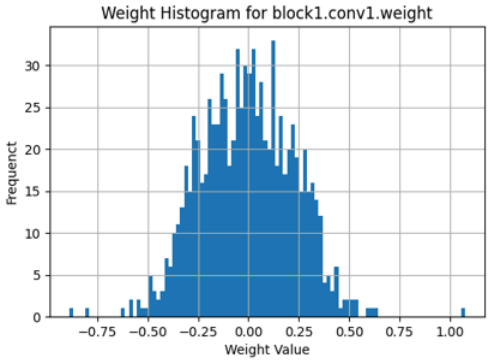
Weight Distribution Comparison Report

Dataset: CIFAR-10
실험 목적 : 테스트 정확도 차이에 따른 모델의 내부 최종 저장된 가중치 분포가 어떻게 차이가 나는지 분석하기 위함

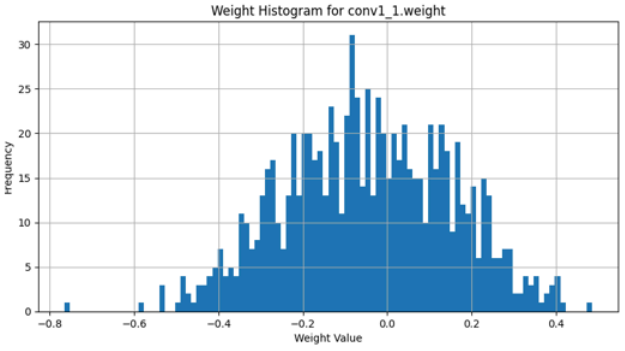
실험 설정 요약

모델 번호	모델 이름	특이사항	**Test Accuracy**
Model 1	Residual VGG-Inspired CNN	BatchNorm 포함, Crop+Flip 데이터 증강, Residual(Shortcut) 포함	**90%**
Model 2	VGG-Inspired CNN (Vanilla)	BatchNorm 없음, 일반적 학습 구조	**69%**

실험 결과 분석 - Conv Layer



Model1, Test Acc = 90%



Model2, Test Acc = 69%

Model1의 Weight Histogram을 보면, 나름 대칭적인 정규 분포 형태를 나타내며, 가중치 값은 -0.5~0.5 사이에 대부분 분포하는 것을 확인할 수 있다. 또한 중심이 0으로 잘 정렬되어 있다.

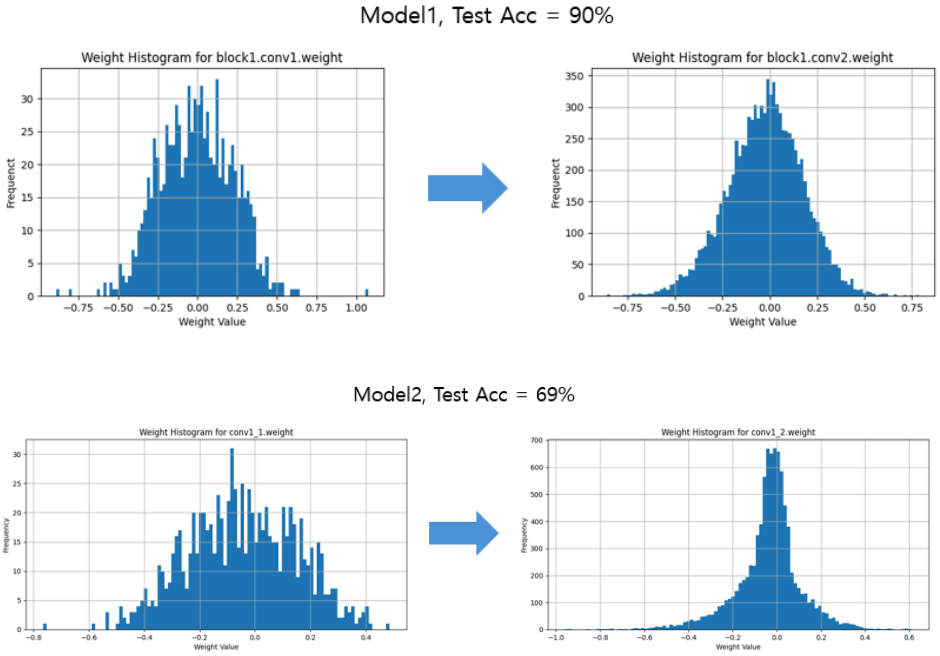
오른쪽 그림의 Model2는 전체적으로 좌측(-0.1 근처)에 치우쳐, 중심이 약간 음수로 편향된 모습을 보인다. 분포의 형태가 넓게 흩어져 있다.

차이점 발생 원인 분석

- 뒤쪽 구조가 역전파**(Back Propagation)**을 통해 **Conv1**에 영향을 줌 : Conv1은 앞단이지만, gradient는 뒤에서부터 전달되므로, 뒤에 있는 BatchNorm, Residual

구조의 유무는 Conv1의 학습에 직접적인 영향을 미친다. **Model1**은 gradient 소실 없이 안정적으로 역전파되어, 가중치 분포가 고르게 유지되었다. 반면, **Model2**는 BatchNorm 없이 gradient가 특정 방향으로 편향되었거나, 일부 뉴런만 활성화되어 치우친 분포가 발생했다.

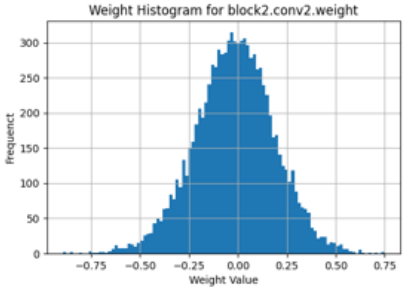
- 데이터 증강의 차이 : Model1은 Crop과 Flip 등의 증강을 통해 더 다양한 feature에 노출되었다. 이것은 다양한 입력을 처리하기 위한 weight분포가 균형있게 학습되게 한다.
  반면, **Model2**는 단조로운 데이터만 학습하여 일부 Weight만 유효하게 학습되었다. (편향된 weight 분포)



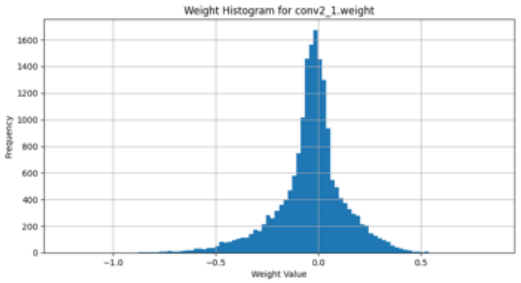
하지만 두 Model 모두 그 다음 Conv Layer에서는 첫 번째 Conv Layer보단 안정적인 분포를 보인다.

Conv1은 데이터에 가장 직접적으로 반응하는 Layer로, 초기에는 raw pixel 정보를 받아들이기 때문에, 노이즈나 입력 데이터 분포의 영향을 많이 받는다. 따라서 구조적 차이나 입력 분포에 따라 다소 불안정하거나 비대칭적인 가중치 분포가 형성될 수 있다.

이후 **Conv Layer**들은 전처리된 **feature map**을 입력받는다**.** Conv2 이후 계층은 Conv1에서 추출된 feature를 바탕으로 학습하며, 입력 자체가 더 정제되어 있다. 특히 Model1에서는 BatchNorm이 추가되어 입력 분포가 정규화되므로 이후 Conv 계층에서 가중치 학습이 훨씬 안정적으로 진행된다.



Model1, Test Acc = 90%



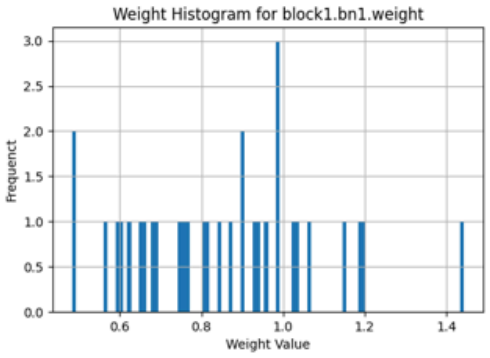
Model2, Test Acc = 69%

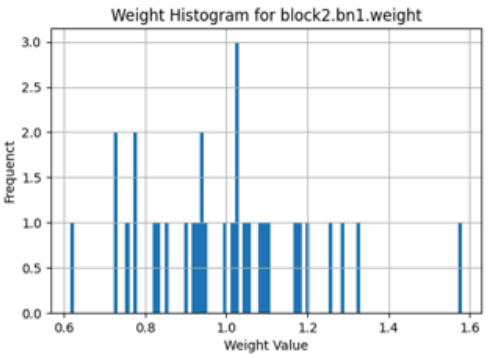
이후 대부분의 Conv Layer들의 가중치 분포는 위 그림처럼 나타났다. Model1에 대해서는, 정규분포 형태로, 중앙 근처에 균형있게 퍼져있는 형태이다. 반면, Model2는 정규 분포 형태이지만, 0 근처에 몰림 현상이 강함을 볼 수 있다.

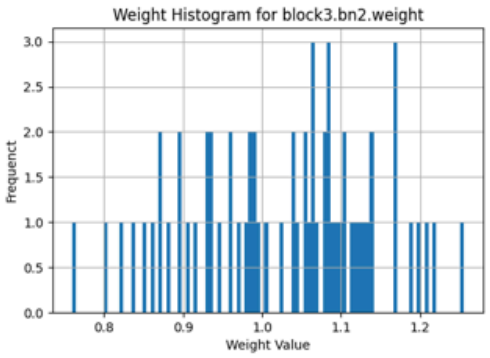
0에 가중치가 몰려 있다는 의미 : 이는 해당 Layer의 많은 Weight가 거의 사용되지 않거나, gradient가 제대로 흐르지 않았다는 것을 의미한다. 특히 BatchNorm이 없는 모델에서는 내부 활성값 분포가 편향되기 쉽다. 이로 인해 일부 weight는 활성화되지 않고 dead feature처럼 동작할 수 있다. 결과적으로, L2 norm이 작은 weight는 신경망에서 학습되지 못하거나, 미세한 값 조정으로만 남아 표현력 손실이 발생한다.

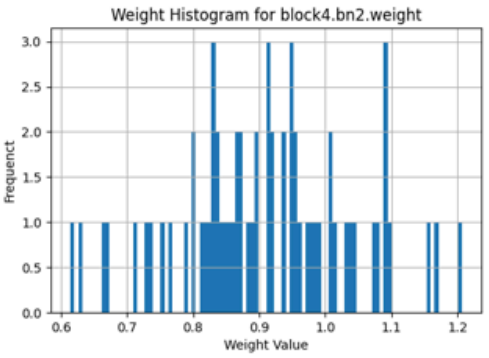
BatchNorm이 있는 경우 **weight**가 정규 분포에 가깝게 유지되는 이유 : BatchNorm은 입력을 평균 0, 표준편차 1로 정규화함으로써 gradient 흐름을 안정화하고, 이를 통해 각 필터가 더 균형 잡힌 방향으로 학습된다. 따라서 weight가 중심을 기준으로 양쪽 고르게 퍼지는 정규 분포 형태를 잘 유지할 수 있다.

실험 결과 분석 - BatchNorm





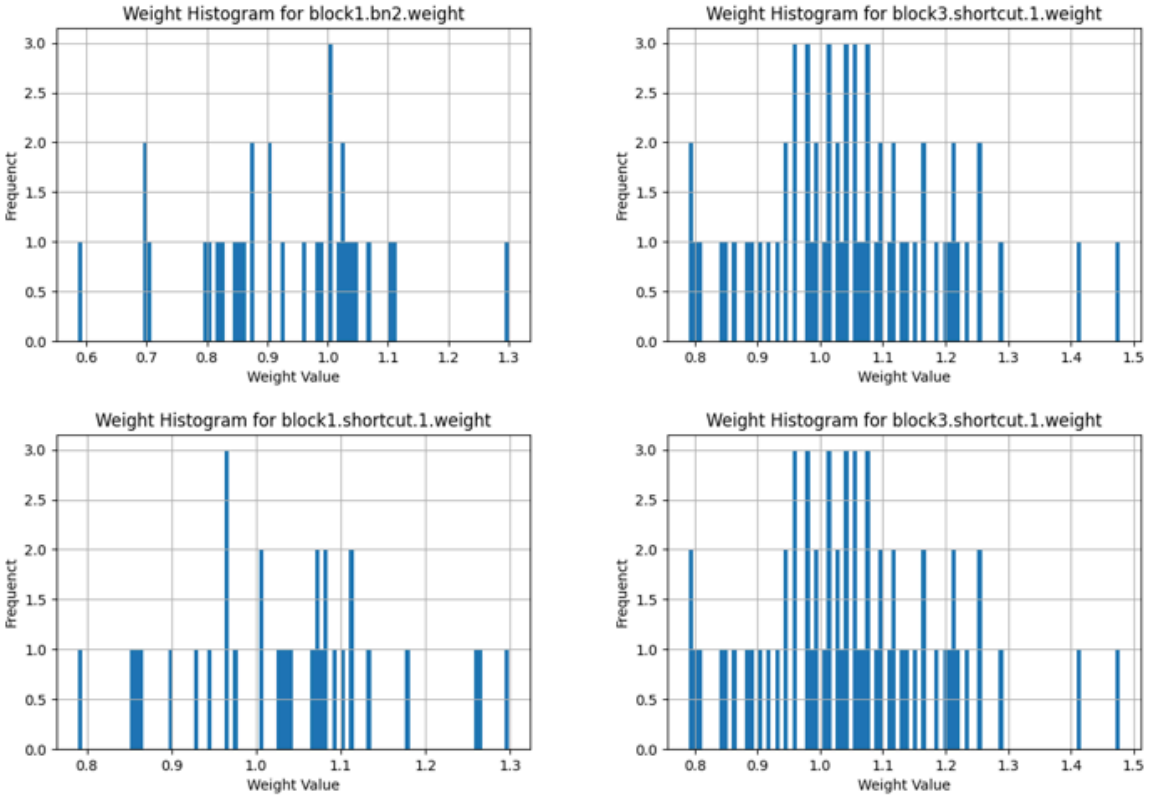




BatchNorm 레이어의 weight는 0.6~1.5 사이의 범위로 분포된다.

BatchNorm의 weight는 학습 가능한 스케일 파라미터 γ (gamma)로, 출력의 스케일을 조절하는 역할을 한다. 초기값은 일반적으로 1.0으로 설정되며, 학습을 통해 0.6~1.5 범위로 조절된다는 것은 출력의 범위가 각 채널마다 다르게 조정되고 있음을 의미한다.

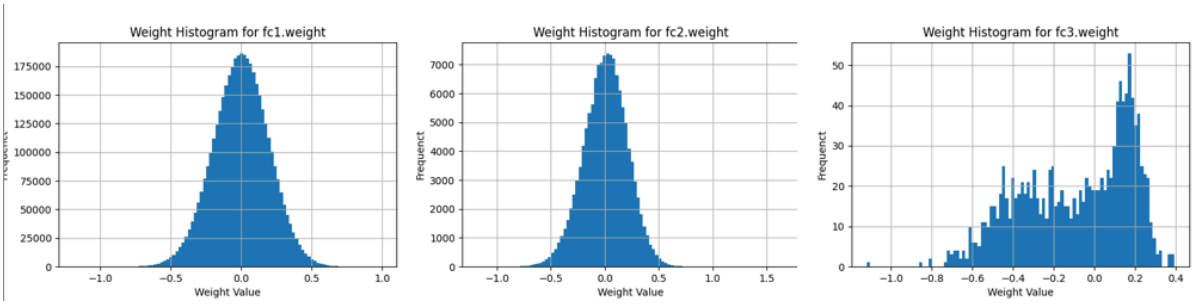
실험 결과 분석 - Residual Shortcut Layer



Shortcut(1x1 Conv) 가중치도 약 0.8 ~ 1.5에 펴진 분포를 보인다.

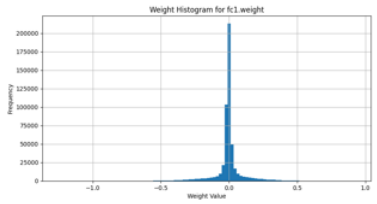
Shorcut Layer는 입력 feature의 shape를 맞추기 위한 identity mapping + 변형 역할을 한다. 보통 큰 변화를 주기보다는 feature를 그대로 유지하거나 가볍게 변형하는 목적이기 때문에, weight 분포가 1 근처에서 살짝만 변형된 상태로 유지된다. 또한 shortcut은 상대적으로 파라미터 수가 적기 때문에, 분포가 약간 불균일해 보여도 통계적으로는 정상이다.

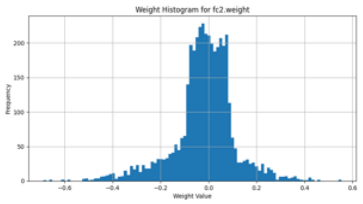
실험 결과 분석 - Fully Connected Layer(Model1, Model2)

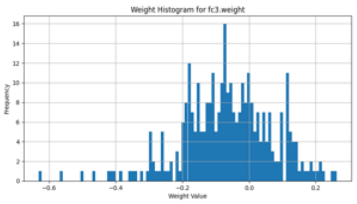


Layer	분포 형태	분석
`fc1`, `fc2`	이상적인 정규분포	안정된 gradient 흐름과 학습의 결과
`fc3`	-0.8 ~ 0.4 범위, 중심이 0.2 근처로 오른쪽 기울어진 형태	클래스 수가 적어 (10개) weight 수가 작고, 출력값의 확률 예측을 위해 bias된 분포가 형성됨

fc3은 Softmax 이전 마지막 선형 계층으로, class score(logit)를 출력한다. 보통 클래스별로 weight가 최적화되며, 크랠스 불균형 또는 초기값에 따라 편향된 분포가 나타날 수 있다. 또한 파라미터 수가 적기 때문에 히스토그램이 뾰족하게 나올 수 있다.







Layer	분포 형태	분석
`fc1`	거의 0 근처에 몰려 있음	학습되지 않음 or dead unit 현상
`fc2`	-0.1~0.1에 좁고 높게 몰림	gradient 흐름 약함, 활성화 미흡
`fc3`	일부 가중치가 튀는 현상	학습이 불균형하게 이루어짐, 일부 노드만 학습됨

BatchNorm이 없으면 중간 계층의 활성화가 gradient 흐름을 방해할 수 있으며, 이는 weight가 0근처에서 멈추거나 극단적인 방향으로 발산되는 원인이 된다. 특히 fc1의 분포가 0에 몰려 있다는 것은 해당 layer가 거의 아무 기능도 하지 않는 상태일 수 있음을 의미한다. 일부 튀는 값은 gradient 폭주나 불균형 학습의 결과로 볼 수 있다.

가중치 초기화 기법 비교 실험(Model2)

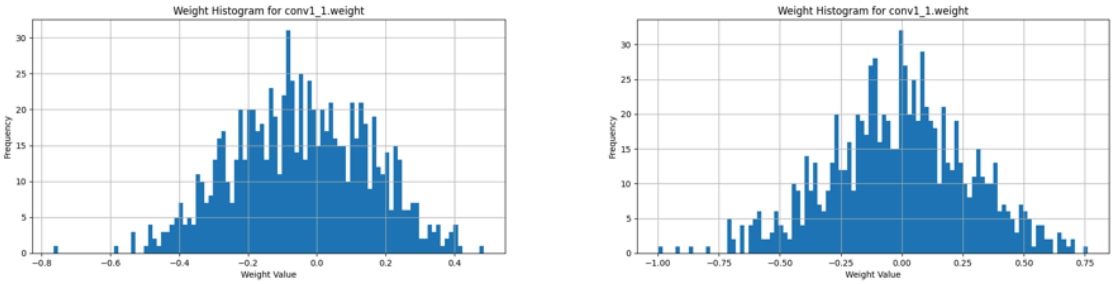
대상 모델 : 성능이 낮았던 VGG-Inspired CNN(No BatchNorm, No Residual)
목적 : weight initialization 방법에 따라 weight 분포 및 Test acc에 어떤 변화가 나타나는지 분석하기 위함
데이터셋 : CIFAR10
조건 : 초기화 기법 제외 모두 동일(모델 구조, 하이퍼파라미터 etc)

실험 횟수 : 각 초기화 방법 당 5회씩 진행
실험 결과 요약(performance)

초기화 기법	**Test Loss**	**Test Accuracy (%)**
He (Kaiming)	0.7715	**75.43**
Xavier (Glorot)	0.8594	74.43
Orthogonal	0.8263	74.99
Baseline (기본값**)**	1.8573	**68.94**

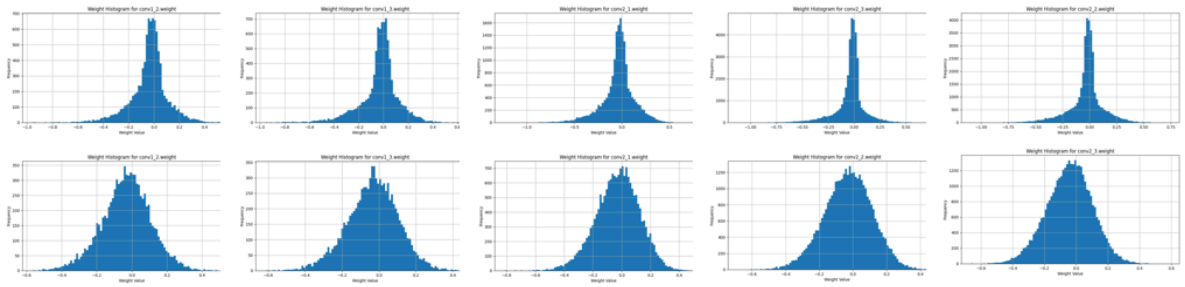
He Initialization(Kaming Normal)

가장 높은 성능 (75.43%)를 달성하였다. ReLU 활성화 함수에 최적화된 분산 스케일링이 적용되어, 각 계층에서의 gradient 소실 방지, 더 효과적인 깊은 네트워크 학습이 가능해졌다.



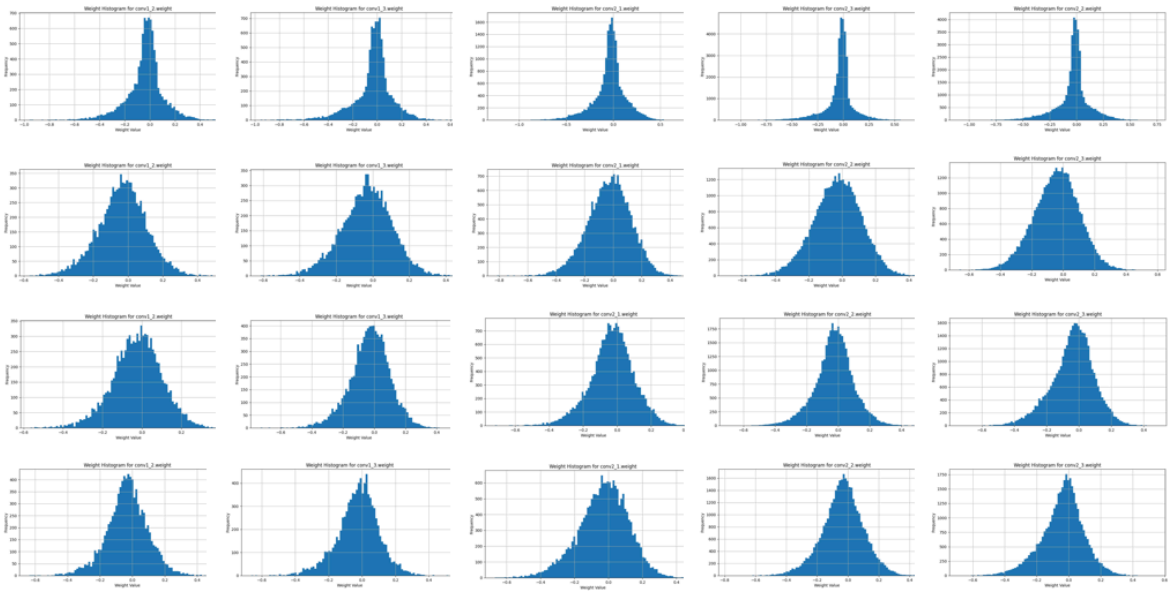
Original He-Initialization

가중치 분포 측면에서도 초기화 기법을 적용하지 않았을 때 첫 Conv Layer의 가중치 분포를 살펴보면, Original은 음수값(-0.1)에 편향되어 있는 반면, He 초기화 기법을 사용했을 때는, 가중치가 넓고 균형있게 퍼지며, 편향된 형태는 감소하였다.



이후 모든 합성곱(Conv) 계층에서도 초기화기법을 적용하지 않은 기존 모델 대비 훨씬 안정적인 정규 분포 형태의 가중치 분포를 확인할 수 있었다.

전체 가중치 분포 비교 이미지-Conv Layer



1. Original
2. He
3. Xavier
4. Orthogonal

안정적인 가중치 분포는 Gradient 흐름을 보장한다. 가중치 초기화가 불균형하거나 0에 몰려 있으면, 역전파 시 gradient가 지나치게 작아지는 gradient vanishing 또는 너무 커지는 gradient exploding 문제가 발생한다.

He(kaiming) 초기화는 ReLU 활성화 함수의 비대칭성(양수 출력) 특성을 고려하여 가중치의 분산을 조정한다.(각 Layer의 입력/출력 분산이 적절히 유지됨)

Xavier 초기화는 입력과 출력의 variance를 동일하게 유지시킨다.(학습 초기에 layer 간 값 폭주 방지)

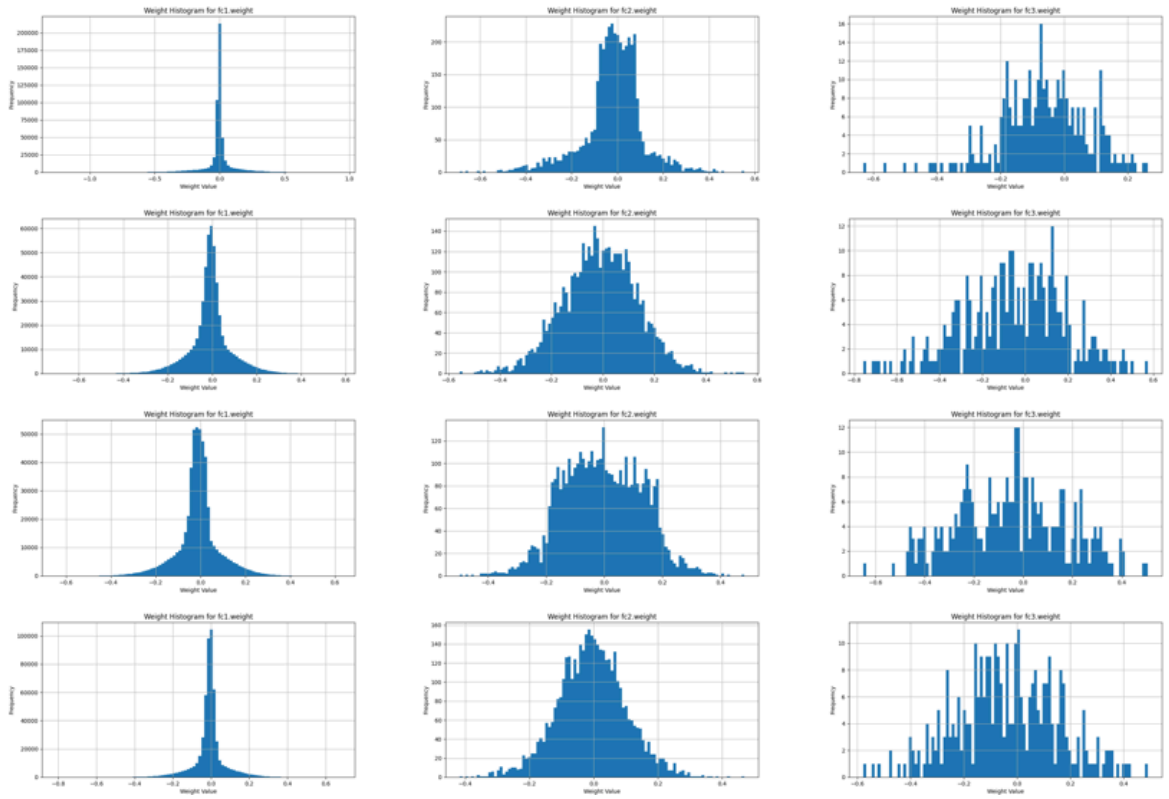
즉, 가중치 초기화를 하면 학습이 초반부터 안정적으로 진행되며 빠른 수렴, 성능향상을 기대할 수 있다.

기본 모델에서 일부 레이어의 가중치가 0 근처에 치우쳐 있으면, ReLU 활성화 후 출력이 0으로 고정되는 dead neuron이 발생한다. 초기화로 weight 분포가 넓고 대칭적으로 퍼지면, 뉴런이 다양한 범위에서 활성화되어 feature 표현력이 강화된다.

즉, 더 많은 뉴런이 학습에 참여하게 되어 feature 다양성을 확보하게 되고, 정확도 향상을 기대할 수 있게된다.

잘 초기화된 가중치는 학습 초기에 서로 다른 방향의 feature detector를 만들 가능성이 높다. 반면 잘못 초기화된 경우, 많은 필터가 유사하거나 무의미한 feature를 학습하여 표현력이 감소된다. 안정적인 정규분표 형태는 다양한 scale과 방향성을 가진 필터를 생성하며 풍부한 표현력을 만들어낸다.
즉, 안정적인 가중치 분포는 gradient 소실/폭주를 방지하고, dead neuron을 줄이며, 학습 초기 feature 다양성을 확보하여 최적화 과정이 원활하게 진행되도록 한다. 이로 인해 모델 성능이 크게 향상된다.

가중치 분포 비교 이미지 - FC



1. Original
2. He
3. Xavier
4. Orthogonal

마찬가지로, FC Layer에서도 가중치 분포가 안정된 모습을 볼 수 있다.
FC Layer는 Conv Layer에서 추출된 feature map을 입력으로 받는다. Conv Layer 가중치가 초기화에 의해 안정화되면 Conv Layer의 출력값(activation)도 정규화에 가까워진다. 입력이 안정되면 FC Layer에 들어오는 값의 분산이 균형을 이루게 되어, 가중치 업데이터가 급격하게 치우치치 않게된다.