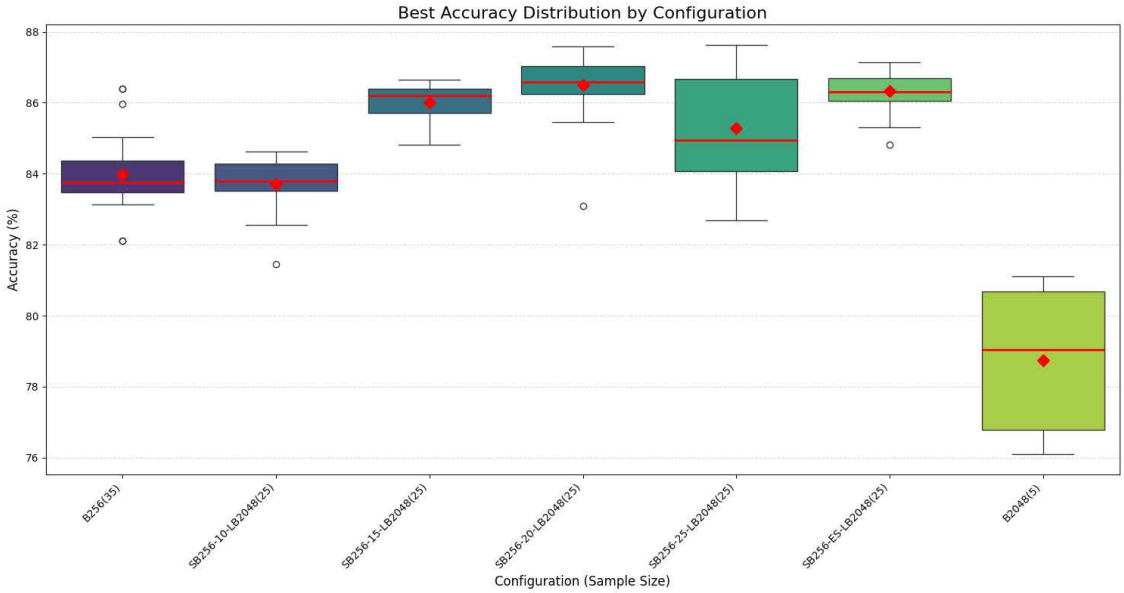
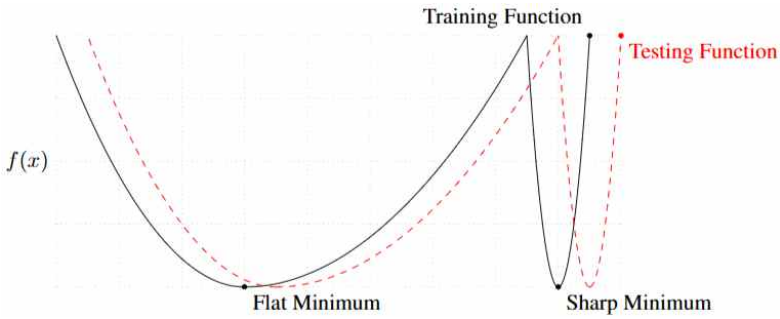


위 이미지는 가장 학습이 뛰어났던 SB(Small batch, 256)에서 일반화 능력이 떨어졌던 LB(Large batch, 2048)로 warm starting방식으로 학습한 그래프이다.
가장 왼쪽의 그래프는 배치사이즈 256으로 학습한 SB방식이고, 가장 오른쪽은 배치사이즈 2048으로 학습한 LB방식이다. 모두 early stopping, patience 10으로 하여 loss가 가장 낮았을 때를 기준으로 Acc를 모은 정보들이다.
SB256-10-LB2048의 의미는 SB인 256에서 epoch 10을 기점으로 LB인 2048로 학습방식을 전환한 후 early stopping을 적용하였다는 의미이다.
오른쪽에서 2번째 그래프는 SB인 256에서 early stopping을 사용하여 가장 loss가 작았던 모델을 기점으로 LB인 2048로 학습 방식을 전환한 후 early stopping을 적용하였다는 의미이다.
()괄호 안 숫자는 샘플의 개수를 의미

논문 *ON LARGE-BATCH TRAINING FOR DEEP LEARNING: GENERALIZATION GAP AND SHARP MINIMA*에서 언급된 말을 인용하여 위와 같은 방식을 해석하면 다음과 같다:
이는 처음 학습할 시 SB의 노이즈들이 flat minima로 이끌도록 하였고, 이후 LB의 일반적인 결론(정답)에 가까운 loss를 통해 학습을 하여 flat minima에서 보다 일반화된 결론에 가까운 파라메터로 수정할 수 있었고, 결론적으로 단순 SB와 LB로만 학습한 모델에 비하여 SB에서 LB로 warm starting하는 방식으로 학습한 모델의 일반화 성능이 좋았다.



최상단 그래프 해석 및 가설: **배치사이즈 256에서 가장 낮은 loss를 기록했을 때의 평균 Epoch은 19.4**이다. 위 그래프에서 가장 좋은 성능을 보이는, **epoch 20을 기점으로 SB에서 LB로 이어서 학습한 모델의 일반화 성능이 가장 좋은데**, 이는 **SB 학습에서 관찰된 순간적인**

loss의 최저 지점이, 반드시 LB가 안정적으로 수렴할 수 있는 넓은 Flat Minima 영역을 의미하지는 않는다는 것을 시사한다. **오히려 epoch 20 시점이, 모델이 탐색을 마치고 Flat Minima의 더 안정적인 구간으로 진입한 상태일 가능성이 높으며, 이것이 더 나은 일반화 성능으로 이어진 것으로 분석**된다.

하지만 모델의 가장 일반화된 모델을 찾는 과정에서 시간상의 제약이 있는 상황이라면 SB에서 최저 손실 모델을 기점으로 LB로 학습하는 것이 시간면에서 효율적일 것으로 보인다.

결론을 통해서 할 수 있는 질문:

1. SB 학습 도중 LB가 안정적으로 수렴할 수 있는 Flat Minima인 영역을 구분하는 방법은 무엇이 있을까?
2. SB에서 LB로 점진적으로 배치 사이즈를 늘려나가는 방식과, SB에서 LB로 한번에 전환하는 방식(현재 실험한 방식) 중 어느 것이 효과적일까?
3. 이와 같이 SB에서 LB로 warm starting하는 방식이 다른 데이터셋이나 더 복잡한 모델 아키텍처(ResNet, Transformer 등)에서도 동일하게 나타날까?
4. SB에서 LB로 전환할 때, 옵티마이저의 상태를 그대로 이어가는 것이 좋을까, 아니면 새롭게 초기화 하는 것이 더 효과적일까?(Adam의 경우 momentum이 있기에)