LID&Formulation

202111259 CSE 김수환

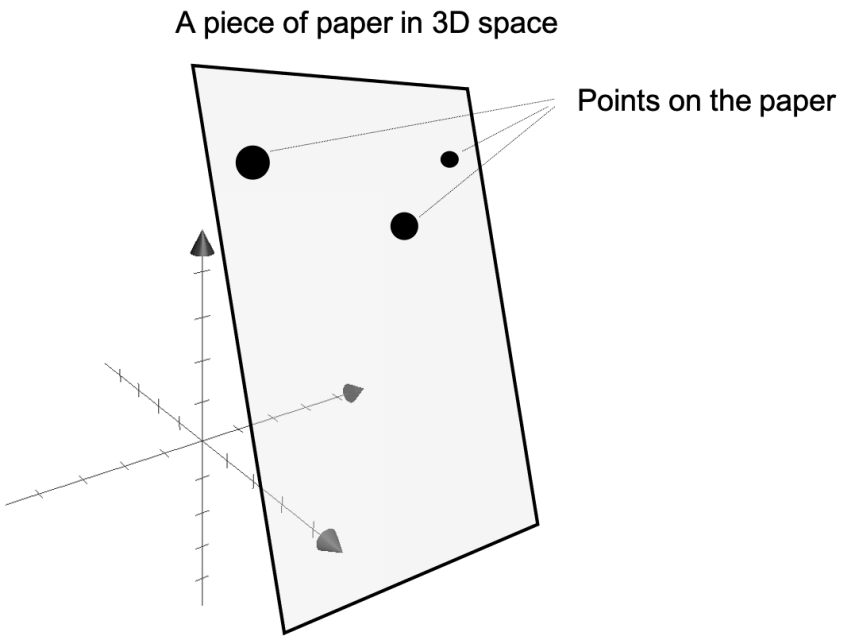
Contents

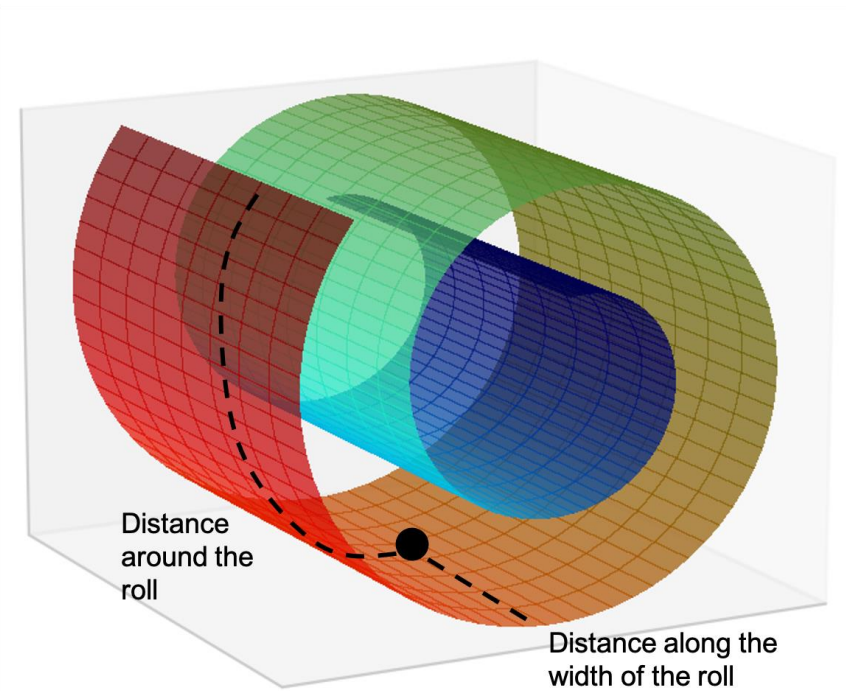
1. LID
2. LID at AT
3. Formulation

1. LID

Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality, ICLR 18'
Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

Q. Intrinsic Dimension이란?
A. 데이터 포인트 분포를 가장 잘 나타낼 수 있는 가장 작은 차원





→ 3차원으로 보이지만, data point의 특징은 2차원으로 충분히 나타낼 수 있음

https://mbernste.github.io/posts/intrinsic_dimensionality/

1. LID

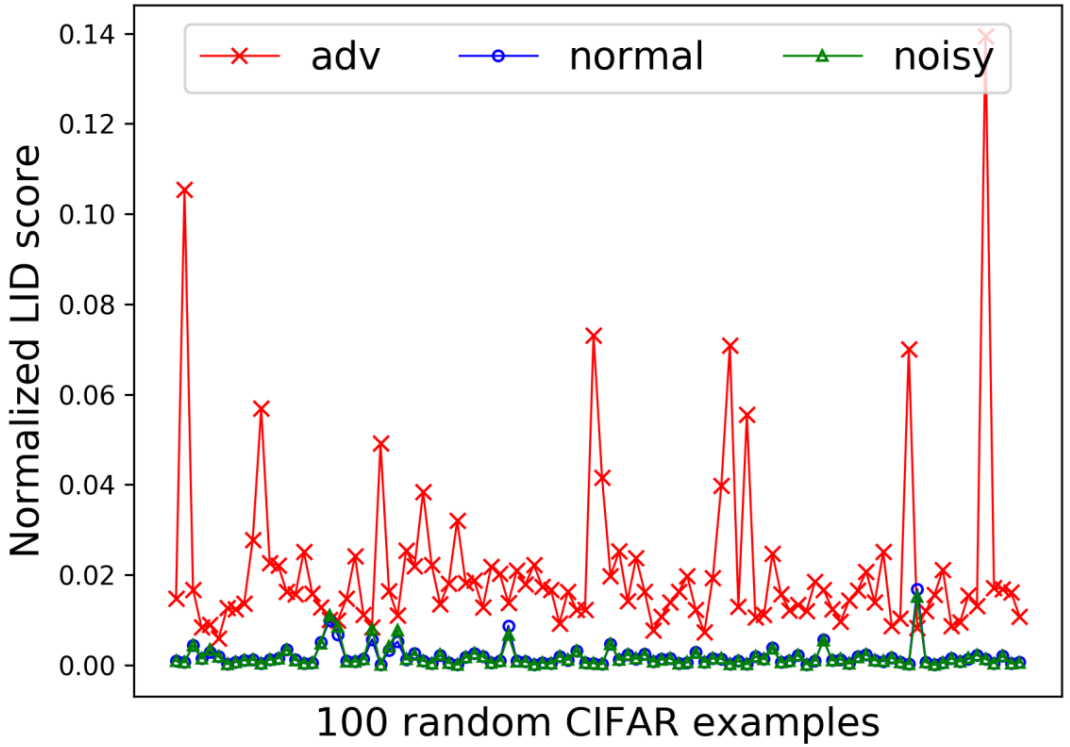
Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality, ICLR 18'

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)} \right)^{-1}.$$

$r_i(x)$: x와 i번째로 가까운 data point와의 거리

1. LID

Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality, ICLR 18'



→ Adversarial attack은 intrinsic dimensionality를 증가시키는 방향이다

1. LID

Characterizing Adversarial Subspaces Using Local Intrinsic Dimensionality, ICLR 18'

adversarial attack은
intrinsic dimensionality를 증가시킨다.

==

adversarial attack은
manifold의 수직 방향 성분을 포함한다.
(ID 늘어남)

==

데이터 포인트가 고차원일 수록
attack에 취약하다.

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

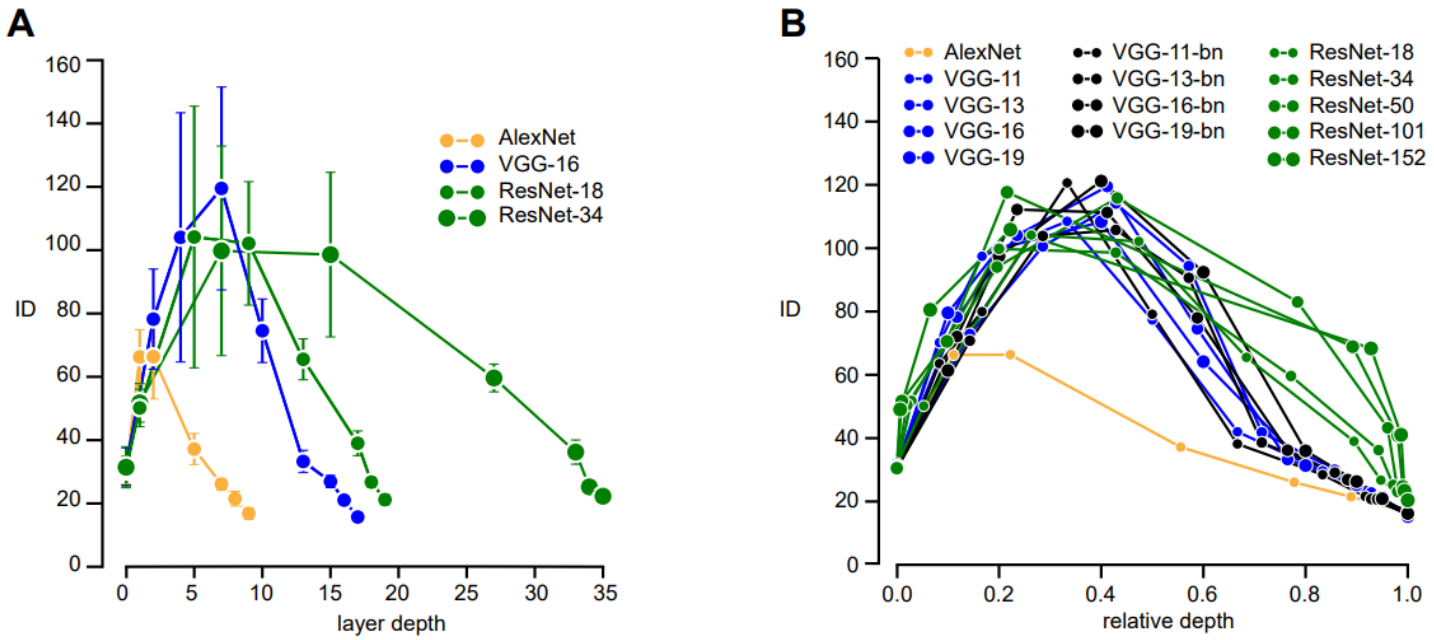
CNN모델이 inference를 진행하면서 만들어내는 feature의 ID는 어떻게 변하는가?

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

CNN모델이 inference를 진행하면서 만들어내는 feature의 ID는 어떻게 변하는가?

→ 초반에는 커지다가 작아짐



1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

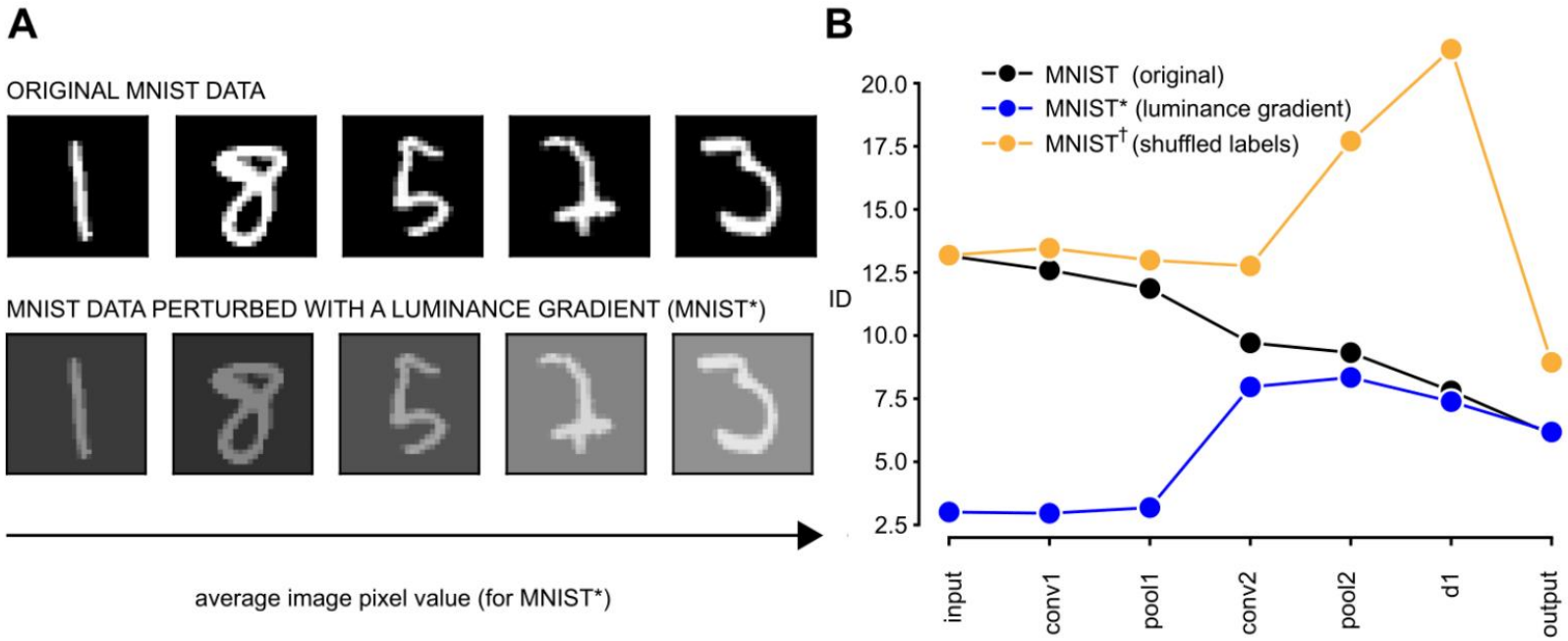
초반에 왜 ID가 확장되는가?

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

초반에 왜 ID가 확장되는가?

→ label과 관계 없는 feature를 제거하는 효과



MNIST는 이미지에 noise가 없고, feature가 직접적으로 노출되어 있음

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

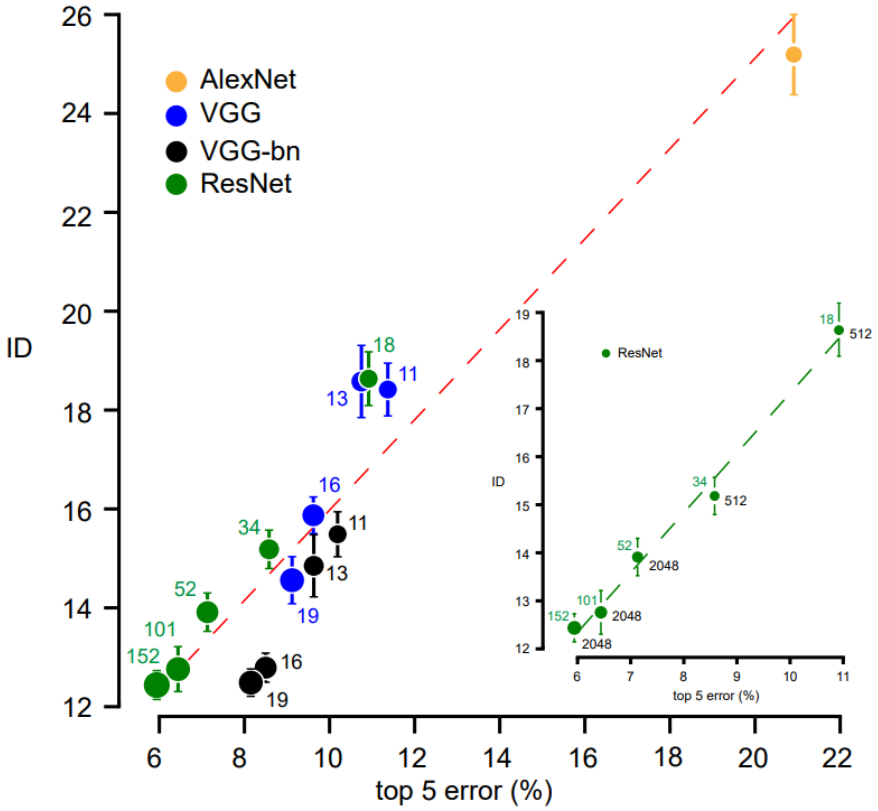
마지막 레이어의 ID는 무엇을 의미하는가?

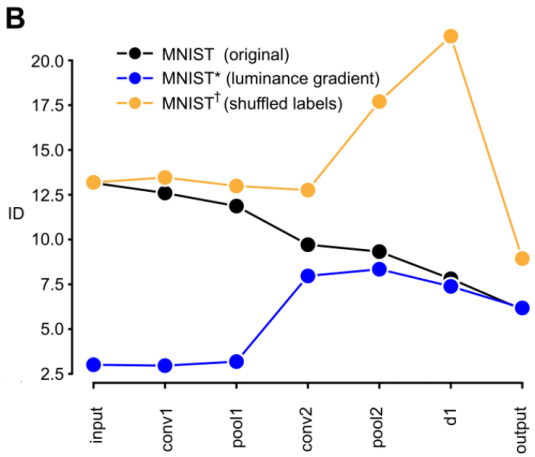
1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

마지막 레이어의 ID는 무엇을 의미하는가?

→ generalization의 지표





1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

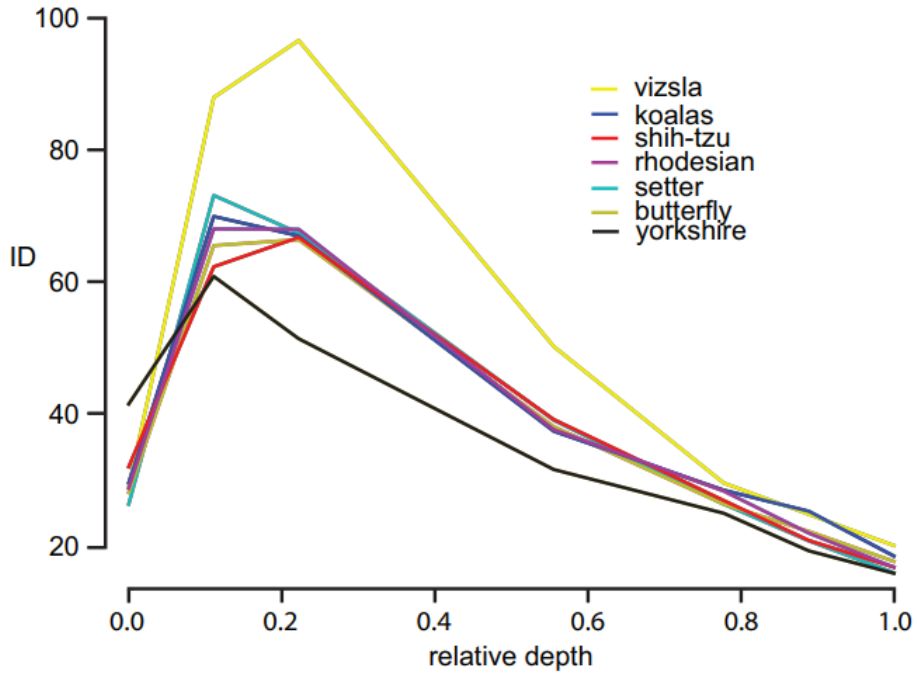
모든 class에서 동일한 현상이 관측되는가?

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

모든 class에서 동일한 현상이 관측되는가?

→ yes



vizsla data에 irrelevant feature가 많았을 것이라 예상

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

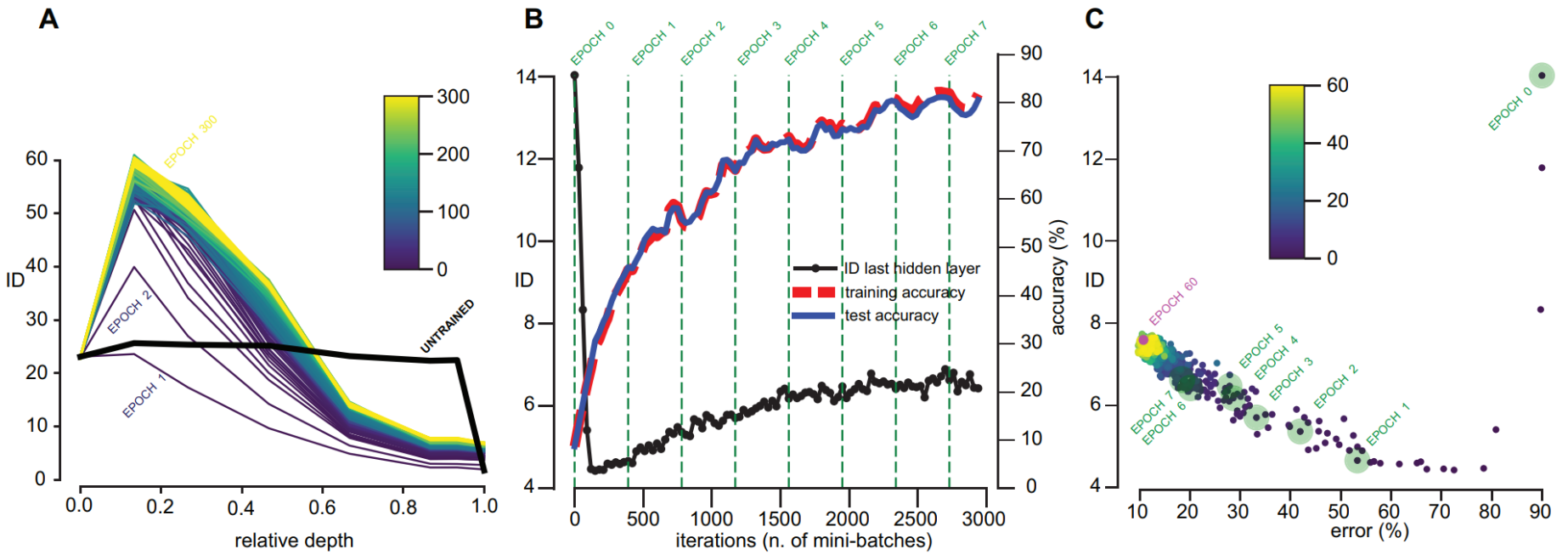
학습이 진행되면서 ID는 어떻게 변하는가?

1. LID

Intrinsic dimension of data representations in deep neural networks, NeurIPS 19'

학습이 진행되면서 ID는 어떻게 변하는가?

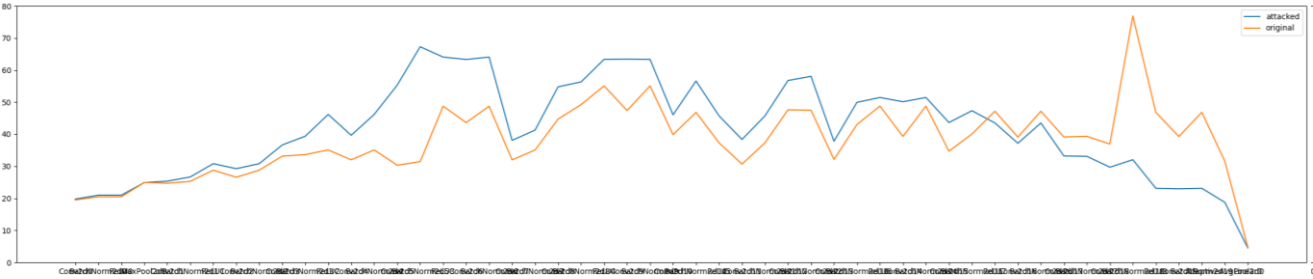
→ ID가 monotonic 하진 않음



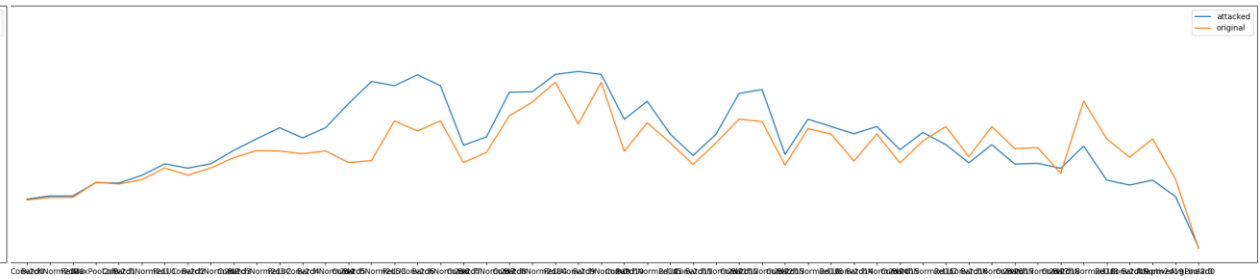
2. LID at AT

그러면, AT 모델은 어떤 특성이 나타나는가?

ST



AT

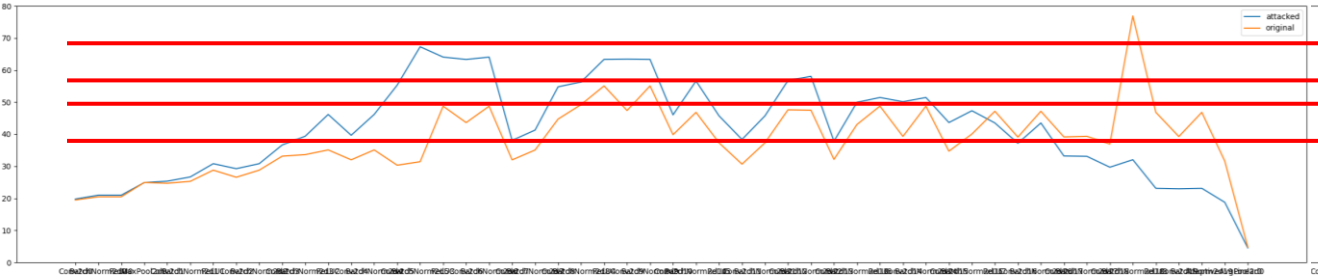


2. LID at AT

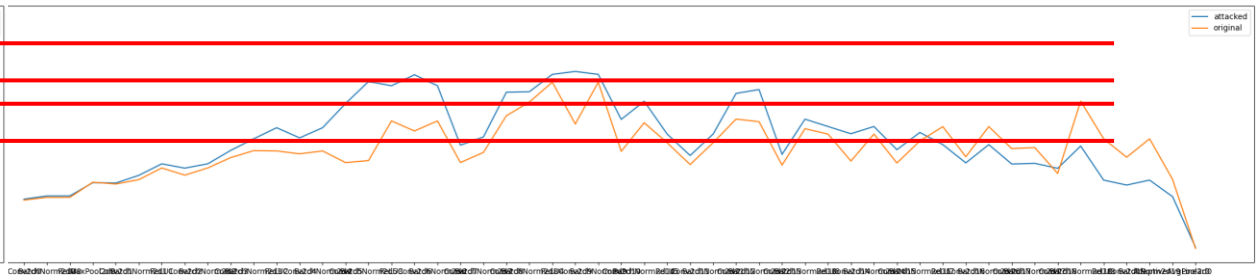
그러면, AT 모델은 어떤 특성이 나타나는가?

→ ID가 전반적으로 작음

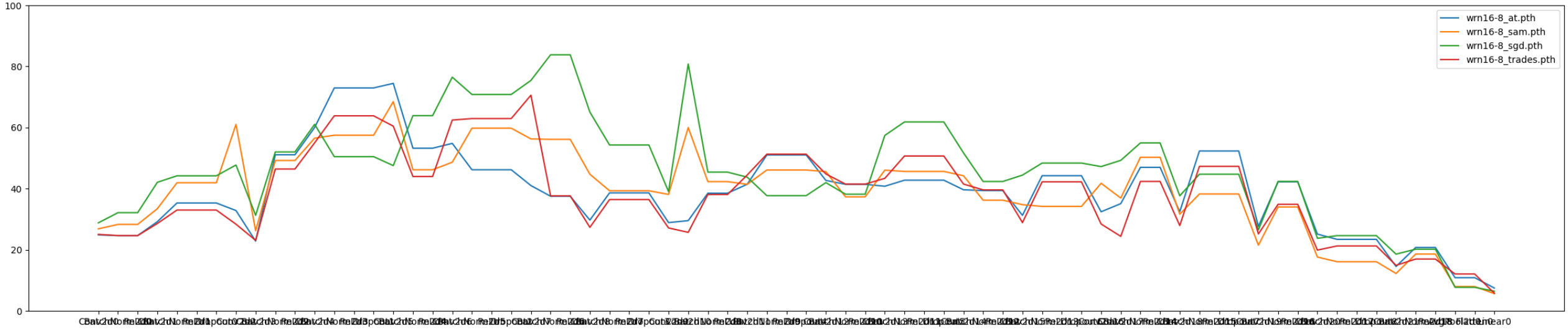
ST

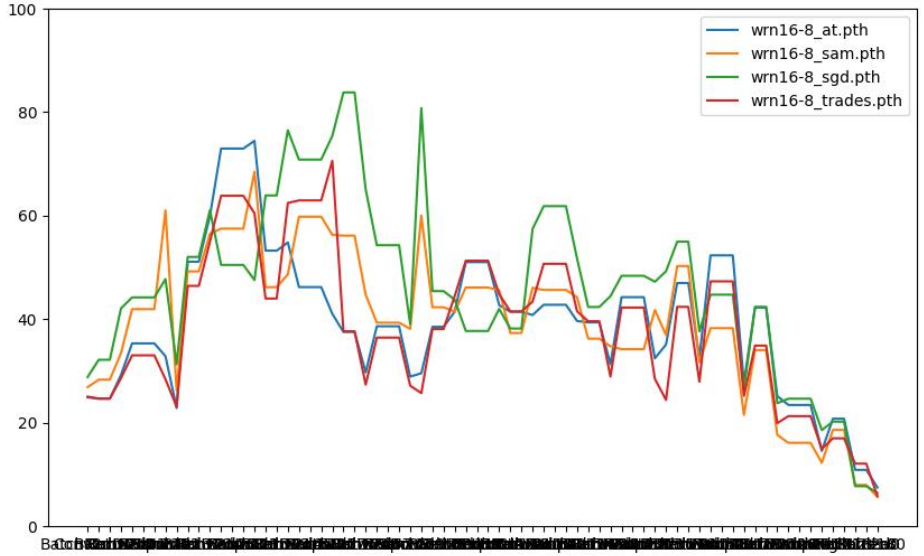


AT



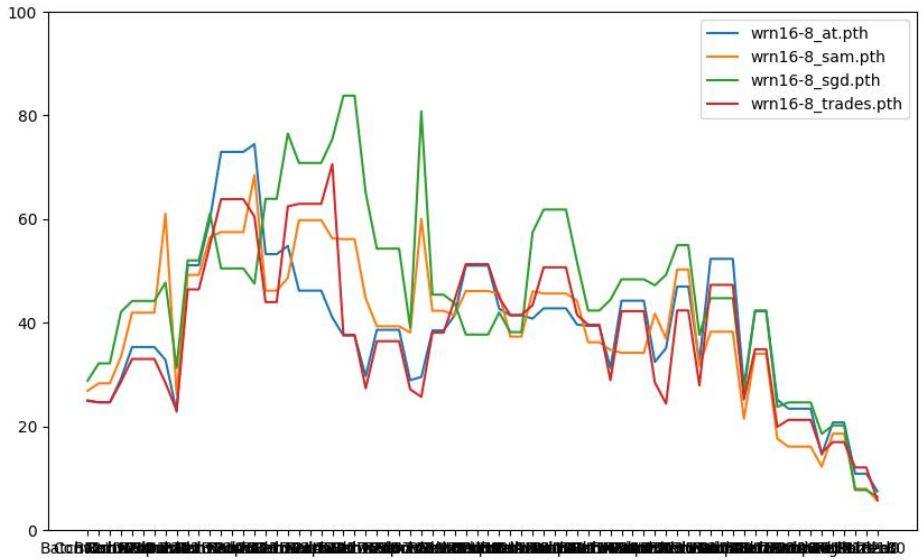
2. LID at AT





model	clean acc.	robust acc.
sgd	96.67%	8.79%
~~adam~~	~~77.0%~~	~~6.83%~~
sam	95.2%	19.17%
at	69.6%	60.32%
trades	65.18%	59.79%

2. LID at AT



model	clean acc.	robust acc.
sgd	96.67%	8.79%
~~adam~~	~~77.0%~~	~~6.83%~~
sam	95.2%	19.17%
at	69.6%	60.32%
trades	65.18%	59.79%

Model	Sum	Avg	Var	Std
SGD	3320.92	46.12	287.15	16.95
SAM	2848.67	39.56	196.71	14.03
AT	2832.19	39.34	193.32	13.90
TRADES	2766.10	38.42	209.12	14.46

3. Idea

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k} \sum_{i=1}^{k} \log \frac{r_i(x)}{r_k(x)} \right)^{-1}.$$

- Idea 1 : LID를 낮추는 방향으로 모델을 학습

ex) 가장 영향을 많이 끼치는 layer 분석 후 해당 layer만 LID 최적화 or 모든 layer의 LID 최적화

- Idea 2: LID로 지금까지 나온 AT 기법 분석