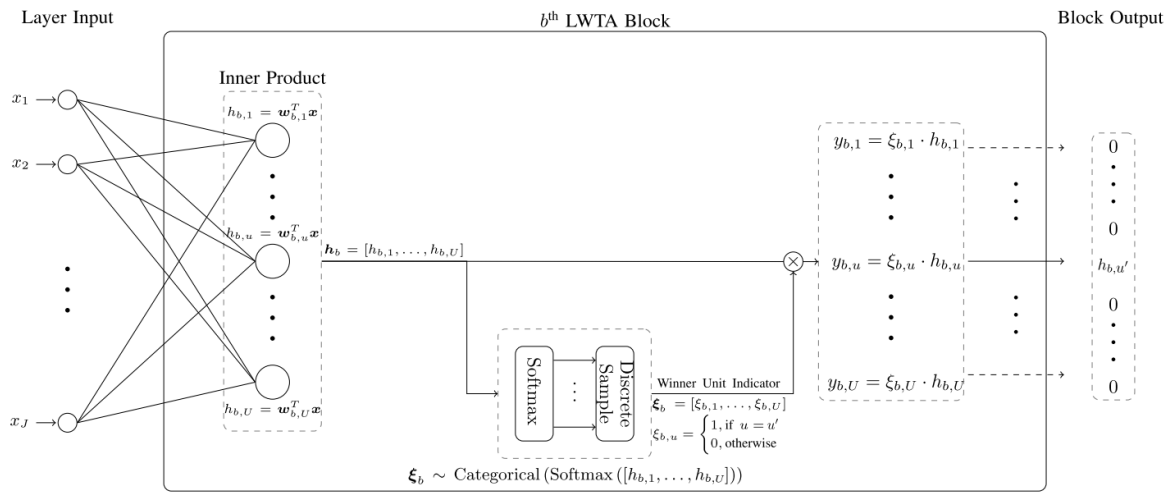
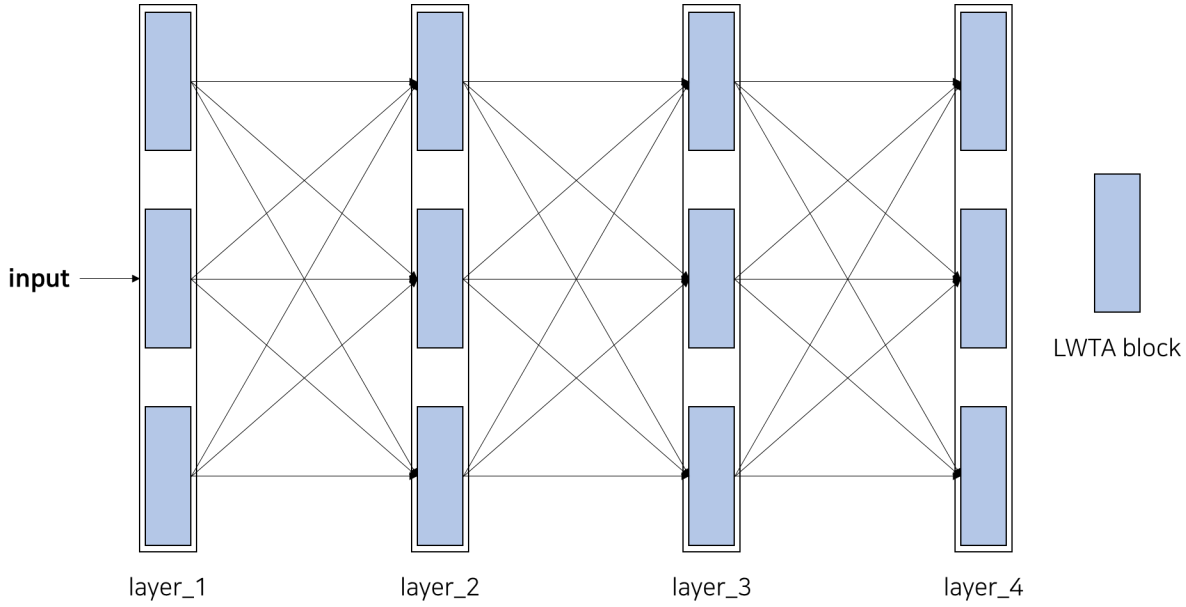
Stochastic Local Winner-Takes-All Networks Enable Profound Adversarial Robustness

학회	Bayesian Deep Learning Workshop, NeurIPS 2021
# 연도	2021
# 이해도	70

LWTA block은 activation을 대체한다



→ 한 레이어에 여러개의 LWTA Block이 있다고 생각하면 됨



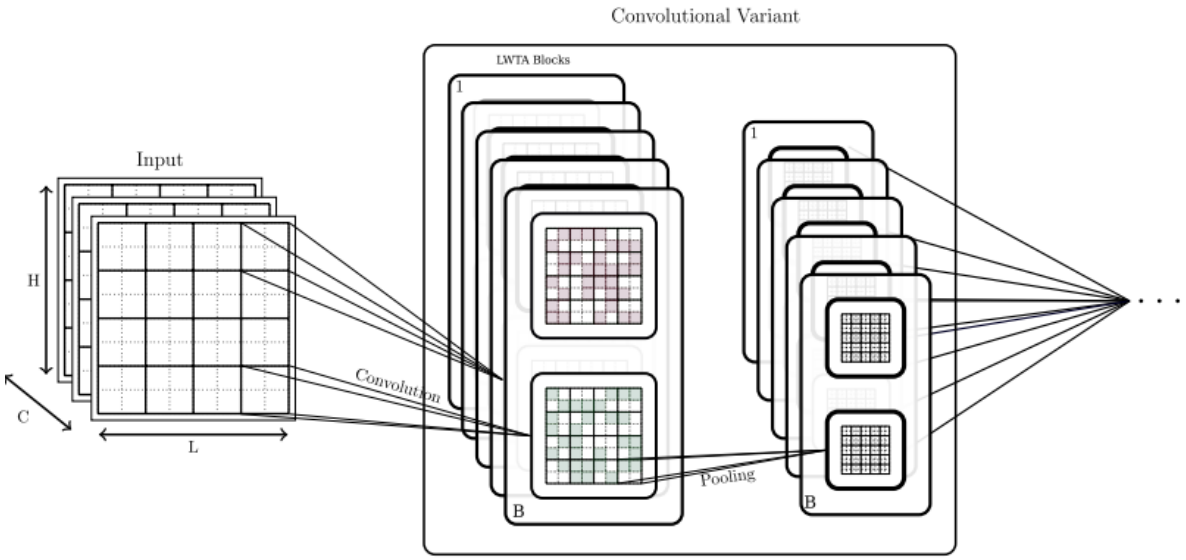
model input : $x \in R^N, N = U \times B$

block num : B

block input : $x \in R^U$

ξ shape : $\xi_i \in \{0, 1\}, \xi \in R^U$

convolution



(b)

channel의 묶음이 한 개의 LWTA block임.

즉, 128 채널 짜리 input이 있고, LWTA block이 16개 있다면, 한 LWTA block에 128/16=8 개의 채널끼리 경쟁함.

채널은 pixel 마다 경쟁함.

성능

Method	AutoAttack
TRADES(Zhang et al., 2019)	53.08
Early-Stop (Rice et al., 2020)	53.42
FAT (Zhang et al., 2020)	53.51
HE (Pang et al., 2020)	53.74
WAR (Wu et al., 2021)	54.73
Pre-training (Hendrycks et al., 2019)†	54.92
MART (Wang et al., 2020)†	56.29
HYDRA (Sehwag et al., 2020)†	57.14
RST (Carmon et al., 2019)†	59.53
Gowal et al. (2021)†	65.88
WAR (Wu et al., 2021)†	61.84
Ours (Stochastic-LWTA/PGD/WideResNet-34-1)	**74.71**
Ours (Stochastic-LWTA/PGD/WideResNet-34-5)	**81.22**
Ours (Stochastic-LWTA/PGD/WideResNet-34-10)	**82.60**

K-winners-take-all