전남대학교
CHONNAM NATIONAL UNIVERSITY


# 네트워크 침입 탐지 성능향상을 위한 비지도 학습 기술 연구

2021. 10. 16. (토)

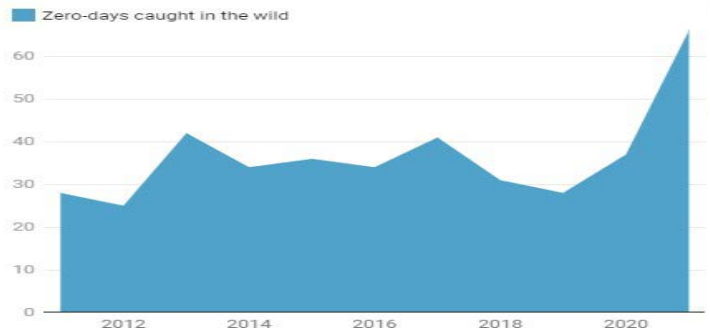
정보보안협동과정
석사과정 정한철

# 목차

# 연구 배경 및 목적

- ICT 산업의 발전으로 인해 네트워크 트래픽 증가
- 네트워크 사용량 증가에 따른 사이버 공격 시도 증가
- 제로 데이 공격, APT 공격 등 지능화된 신종 공격시도 증가
- 기존의 네트워크 침입 탐지 시스템으로는 효과적인 대응 한계 발생
- 딥러닝 알고리즘을 적용한 네트워크 침입 탐지 모델 연구 활발히 진행
- **비지도학습 모델인 RBM 모델을 이용하여 효과적인 네트워크 침입 탐지 모델 제안**


출처 : Cisco VNI 2017~2022 전 세계 IP 트래픽 전망



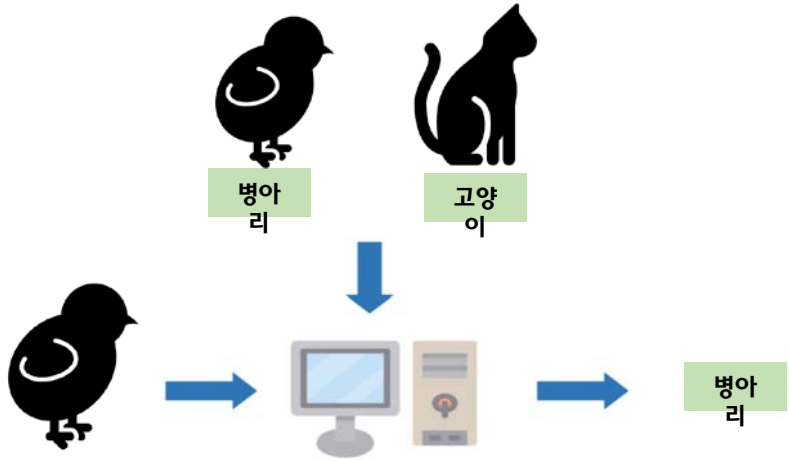

출처 : MIT Technology Review, 2021 has broken the record for zero-day hacking attacks

# 관 련 연 구

- 네트워크 침입 탐지 기술
  - 시그니처 기반 탐지 기법
    - 알려진 공격을 분석하여 패턴을 저장, 패턴과의 일치 여부 확인을 통해 탐지
    - 비교적 오탐률이 낮고 효율적인 탐지 가능, 알려진 공격 이 외에는 탐지 불가
  - 이상 탐지 기법
    - 정상데이터를 다양한 기법을 이용해 분석하여 패턴을 만들고 이를 기준으로 침입 탐지
    - 주성분 분석, 통계적 분석 등 다양한 기법을 이용하여 패턴 생성
    - 오탐률이 높고 정상과 비정상 구분을 위한 임계치 설정이 어려움
    - **인공지능 알고리즘을 활용한 연구** 활발히 진행

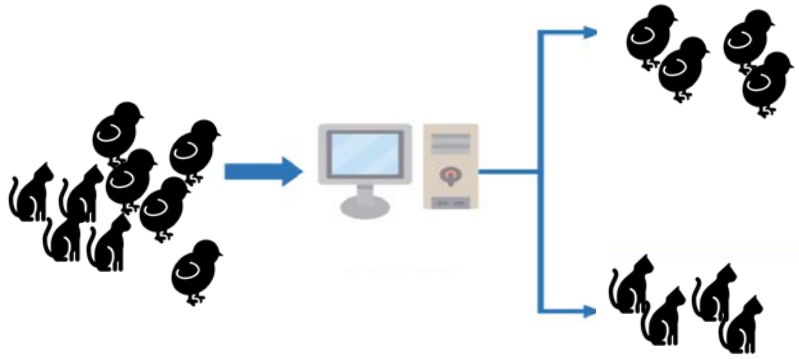
# 관 련 연 구

- **지도학습**을 이용한 연구
  - SVM을 이용한 침입탐지 알고리즘 연구 [1]
  - RNN 모델을 이용한 침입탐지 시스템 연구 [2]
  - SVM, KNN을 LSTM과 결합한 모델 제안 [3]
  - RBM과 SVM을 결합한 FCN 제안 [4]




- 지도학습 모델을 학습하기 위해서는 **레이블링된 데이터가 필요**
  - 네트워크 트래픽을 학습을 위해 레이블링하는 작업은 매우 어렵고 많은 자원이 소모된다.
  - 네트워크 환경의 특성 상 대부분이 정상 트래픽이기 때문에 데이터 불균형에 따른 문제 발생 가능

# 관 련 연 구



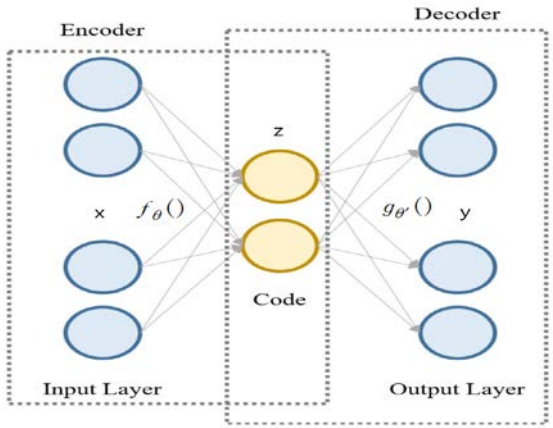
- **비지도학습**을 이용한 연구
  - **오토인코더**를 이용한 연구
    - 은닉층에 컨볼루션 레이러를 사용한 오토인코더 모델 제안 [5]
    - 오토인코더의 출력층에 sotfmax를 사용한 모델 제안 [6]
    - 노이즈를 추가한 오토인코더 모델 제안 [7]
    - 오토인코더의 재구성 손실 분포를 이용하여 임계치에 따른 성능 비교 연구 [8]
  - 오토인코더의 **재구성 손실 값의 분포를 이용해 임계치 설정**하여 탐지 진행
    - 일부 연구에서는 레이블링된 데이터를 사용하여 연구 진행

# 관 련 연 구

- 비지도학습을 이용한 연구
  - RBM을 이용한 연구
    - 사전학습에 RBM을 이용한 SVM,KNN 모델 제안[9]
    - RBM을 활용해 기존데이터의 노이즈와 이상치를 제거 후 차이 비교 연구[10]
    - RBM으로 구성된 DBN-SVM 모델 제안[11]
    - RBM을 이용해 특징 추출 및 축소한 데이터를  SVM을 이용하여 분류 [12]
  - RBM을 특징 추출에만 사용하거나 다른 모델의 일부로 사용
  - **RBM의 likelihood 함수 분포를 활용한 침입 탐지 모델 제안**

# 비지도학습 기반 침입 탐지 기법

- **오토인코더 기반 탐지 모델**




$$x \in R^d \quad z \in R^d \quad y \in R^d$$

$$z = f_\theta(x) = s(Wx + b)$$

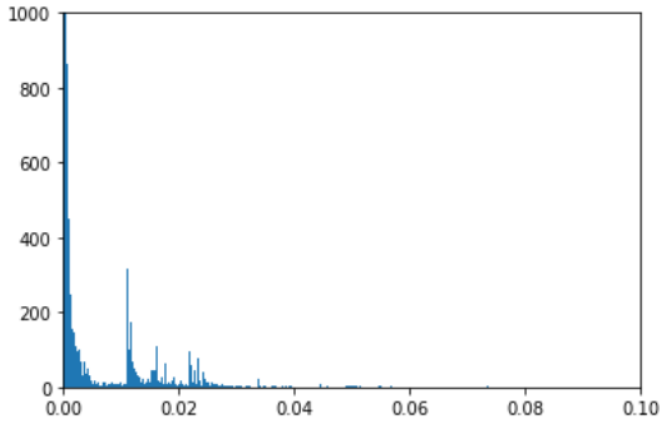
$$y = g_{\theta'}(z) = s(W'x + b')$$

$$\theta^*, \theta'^* = \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{n=1}^{n} L(x^{(i)}, y^{(i)})$$

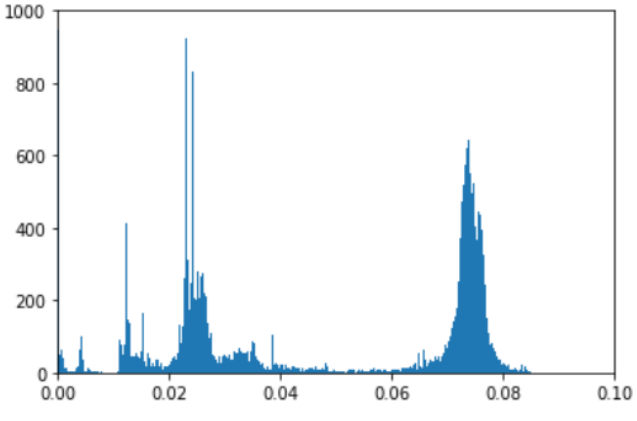
$$= \arg\min_{\theta,\theta'} \frac{1}{n} \sum_{n=1}^{n} L(x^{(i)}, g_{\theta'}(f_\theta(x^i)))$$

$$L(x', x) = \frac{1}{N} \sum_{i=1}^{N} (x' - x)^2$$

**정상 데이터의 Reconstruction Error 의 분포**
**공격 데이터의 Reconstruction Error를 비교 후**
**Threshod를 설정하여 탐지**



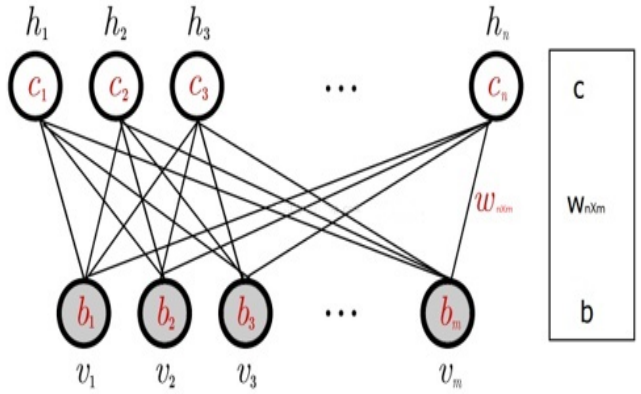

정상데이터 RE 분포




공격데이터 RE 분포

# 비지도학습 기반 침입 탐지 기법

- **RBM 기반 탐지 모델**
  - **RBM은 입력 데이터에 대한 확률분포를 학습할 수 있는 에너지 기반의 단층의 생성모델 신경망**




$$v_i, h_j \quad w_{ij} \quad a_i, b_j$$

$$\theta = w, b, c$$

$$p(v,h;\theta) = \frac{e^{E(v,h;\theta)}}{Z(\theta)} \qquad Z(\theta) = \sum_{u,g} e^{E(u,g;\theta)}$$

$$E(v,h,;\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} v_i w_{ij} h_j + \sum_{i=1}^{I} b_i v_i + \sum_{j=1}^{J} c_j h_j$$

$$L(\theta;v) = \log p(v;\theta) = \log \sum_{h} p(v,h;\theta)$$

$$= \log \sum_{h} e^{E(v,h;\theta)} - \log \sum_{u,g} e^{E(u,g;\theta)}$$

**RBM을 학습하기 위해서 Log- Likelihood 함수를 최대화하는 기법을 사용하여 학습 진행**

# 비지도학습 기반 침입 탐지 기법

- 제안된 RBM 기반 탐지 모델
  - **Bernoulli-Bernoulli RBM(BBRBM)**

    $v \in 0,1^I$ and $h \in 0,1^J$

  - **Gassian-Bernoulli RBM[GBRBM]**

    $v \in R^I$ and $h \in 0,1^J$

  - **Kernel based RBM[KRBM]**
  - $\mathbf{v} = (v_1, v_2, \ldots v_n) \rightarrow \emptyset(\quad) \rightarrow \emptyset(v)$
  - $h = (h_1, h_2, \ldots h_m)^T$
  - $\Phi(W) = (\phi(w_1), \phi(w_2), \ldots \phi(w_m))$
  - $R^n \rightarrow R^f \quad k(v_i, v_j) = \quad \emptyset(v_i)^T \emptyset(v_j)$

$$p(v,h;\theta) = \frac{e^{E(v,h;\theta)}}{Z(\theta)} \qquad E(v,h,;\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} v_i w_{ij} h_j + \sum_{i=1}^{I} b_i v_i + \sum_{j=1}^{J} c_j h_j$$

$$Z(\theta) = \sum_{u,g} e^{E(u,g;\theta)}$$

$$p(v,h;\theta) = \frac{e^{E(v,h;\theta)}}{Z(\theta)} \qquad E(v,h,;\theta) = \sum_{i=1}^{I}\sum_{j=1}^{J} a v_i w_{ij} h_j - \sum_{i=1}^{I} \frac{a^2}{2}(v_i - b_i)^2 + \sum_{j=1}^{J} c_j h_j$$

$$Z(\theta) = \int_{u \in R^I} \ldots \int \sum_{g} e^{E(u,g;\theta)} du$$

$$p(\phi(\boldsymbol{v}), \boldsymbol{h};\theta) = \frac{e^{-\frac{1}{2}E(\phi(\boldsymbol{v}), \boldsymbol{h};\theta)}}{Z(\theta)}$$

$$E(\phi(v), h) = \frac{1}{\sigma^2}\phi(v)^T\phi(v) - \frac{2h^T}{\sigma^2}\boldsymbol{\Phi}(W)^T\phi(v) + h^T\Lambda^{-1}h = \frac{1}{\sigma^2}k(v,v) - \frac{2}{\sigma^2}\sum_{j=1}^{m} h_j k(w_j, v) + \sum_{j=1}^{m}\frac{h_j^2}{\lambda_j}$$

$$Z = \int \int e^{-E(\phi(\boldsymbol{v}), \boldsymbol{h})} d\boldsymbol{h} d\phi(\boldsymbol{v})$$

# 비지도학습 기반 침입 탐지 기법

- 제안된 RBM 기반 탐지 모델
  - **BBRBM Log-Likelihood**

$$L(\theta;v) = \log p(v;\theta) = \log \sum_h p(v,h;\theta)$$
$$= \log \sum_h e^{E(v,h;\theta)} - \log \sum_{u,g} e^{E(u,g;\theta)}$$

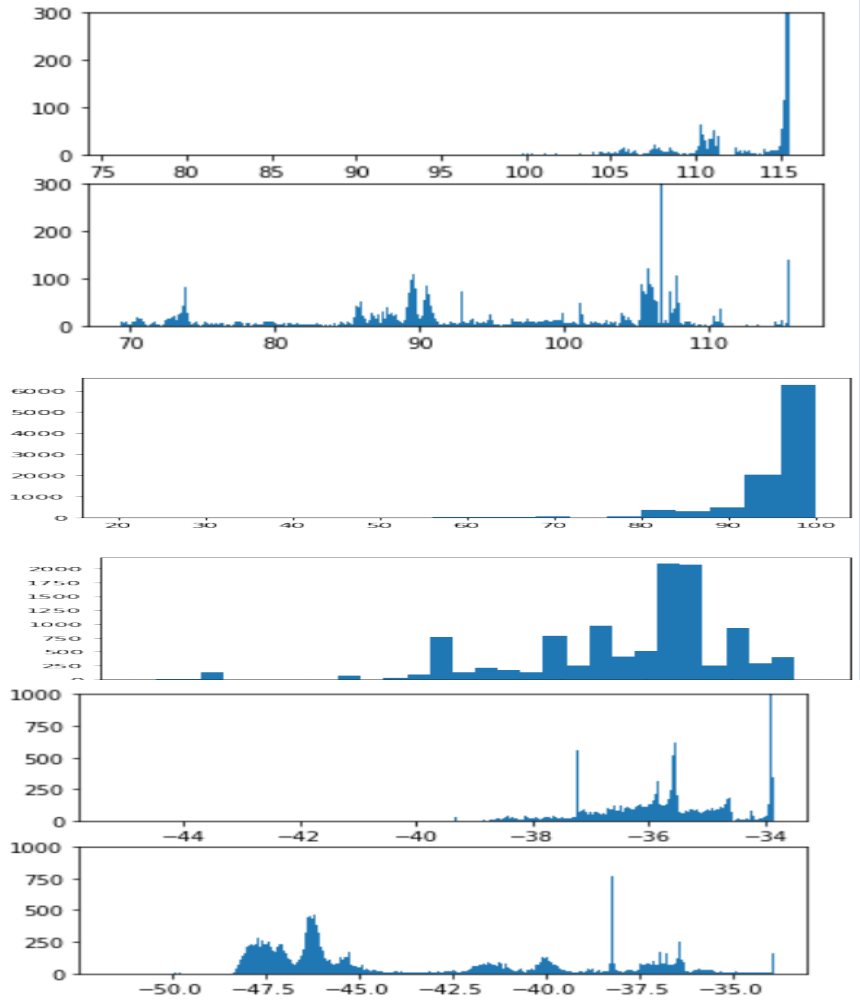
  - **GBRBM Log-Likelihood**

$$L(\theta;v) = \log p(v;\theta) = \log \sum_h p(v,h;\theta)$$
$$= \log \sum_h e^{E(v,h;\theta)} - \log \int_{u \in R^I} \cdots \int \sum_g e^{E(u,g;\theta)} du$$

  - **KRBM Log-Likelihood**

$$l = \ln p(\phi(v)|\theta) = \ln \frac{1}{Z} \int e^{-\frac{1}{2}E(\phi(v),h)} dh$$
$$= \ln \int e^{-\frac{1}{2}E(\phi(v),h)} dh - \ln \int \int e^{-\frac{1}{2}E(\phi(v),h)} dh d\phi(v)$$

Log- Likelihood distribution



# 실험 및 결과 분석

- 실험에 사용한 데이터 : NSL-KDD 데이터 셋
  - 98년 DARPA에서 수집된 데이터를 수집한 KDD Cup99 데이터셋의 중복 레코드를 제거하여 생성
  - 41개의 속성값과 40개의 공격 타입 존재

| | Total | Normal | Attack |
|---|---|---|---|
| KDDTrain+ | 125,973 | 67,343 | 58,630 |
| KDDTest+ | 22,544 | 9,711 | 12,833 |

| 데이터 전처리 과정 | |
|---|---|
| 데이터 병합 | 학습데이터 + 평가데이터 |
| 데이터 정규화 | minmax scaler, standard scaler |
| 원-핫 인코딩 | 범주형 데이터 --> 정수형 데이터 |
| 이진 분류 | normal, abnormal 분류 |
| 상관관계 분석 및 차원축소 | 상관 관계 분석 및 차원 축소 |
| 병합 데이터 분리 | 병합데이터 --> 학습, 평가 데이터 |
| 학습 데이터 내 공격 데이터 제거 | 학습데이터 --> 정상 데이터 |

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | tcp | ftp_data | SF | 491 | 0 | 0 | 0 | 0 | 0 | ... | 25 |
| 1 | 0 | udp | other | SF | 146 | 0 | 0 | 0 | 0 | 0 | ... | 1 |
| 2 | 0 | tcp | private | S0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 26 |
| 3 | 0 | tcp | http | SF | 232 | 8153 | 0 | 0 | 0 | 0 | ... | 255 |
| 4 | 0 | tcp | http | SF | 199 | 420 | 0 | 0 | 0 | 0 | ... | 255 |

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | tcp | ftp_data | SF | 3.558064e-07 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.098039 |
| 1 | 0.0 | udp | other | SF | 1.057999e-07 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.003922 |
| 2 | 0.0 | tcp | private | S0 | 0.000000e+00 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.101961 |
| 3 | 0.0 | tcp | http | SF | 1.681203e-07 | 6.223962e-06 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.000000 |
| 4 | 0.0 | tcp | http | SF | 1.442067e-07 | 3.206260e-07 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.000000 |

| | duration | protocol_type | service | flag | src_bytes | dst_bytes | land | wrong_fragment | urgent | hot | ... | dst_host_srv_count |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.0 | tcp | ftp_data | SF | 3.558064e-07 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.098039 |
| 1 | 0.0 | udp | other | SF | 1.057999e-07 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.003922 |
| 2 | 0.0 | tcp | private | S0 | 0.000000e+00 | 0.000000e+00 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.101961 |
| 3 | 0.0 | tcp | http | SF | 1.681203e-07 | 6.223962e-06 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.000000 |
| 4 | 0.0 | tcp | http | SF | 1.442067e-07 | 3.206260e-07 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 1.000000 |

| | protocol_type_icmp | protocol_type_tcp | protocol_type_udp | service_IRC | service_X11 | service_Z39_50 | service_aol | service_auth | service_bgp | service_courier | ... | flag_REJ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 3 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |
| 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | ... | 0 |

# 실험 및 결과 분석

- 성능평가 지표

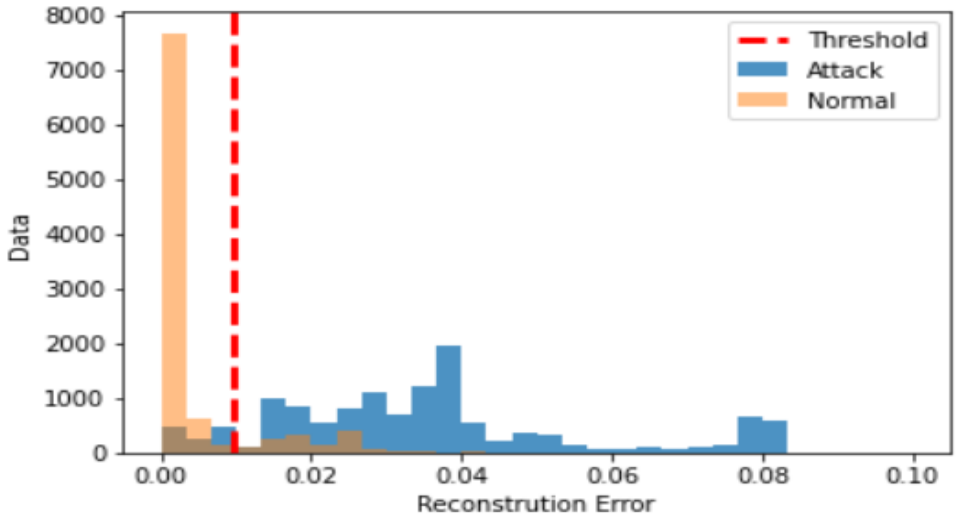
| Confusion Matrix | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | TP | FP |
| | Negative | FN | TN |

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad Recall = \frac{TP}{TP+FN} \qquad Precision = \frac{TP}{TP+FP}$$

$$F1\,score = 2\frac{Precision \times Recall}{Precision + Recall}$$

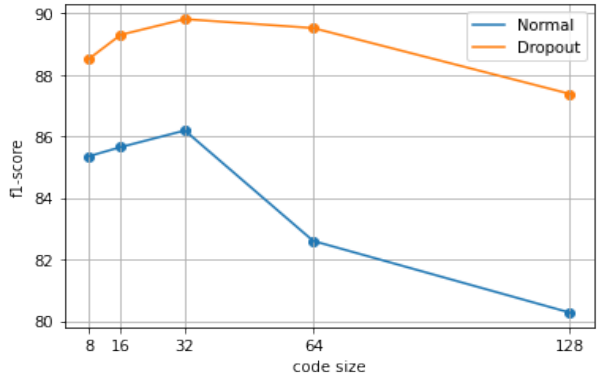
- 오토인코더 모델 침입 탐지 실험결과

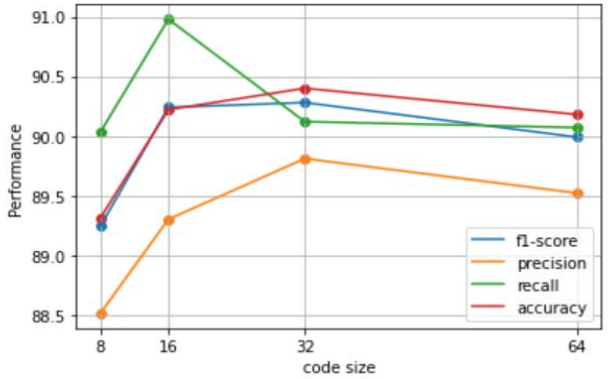
| | 모델 계층 구조 |
|---|---|
| 8 | input –512-256-128-64-32-16-**8**-16-32-64-128-256-512-output |
| 16 | input –512-256-128-64-32-**16**-32-64-128-256-512-output |
| 32 | input –512-256-128-64-**32**-64-128-256-512-output |
| 64 | input –512-256-128-**64**-128-256-512-output |

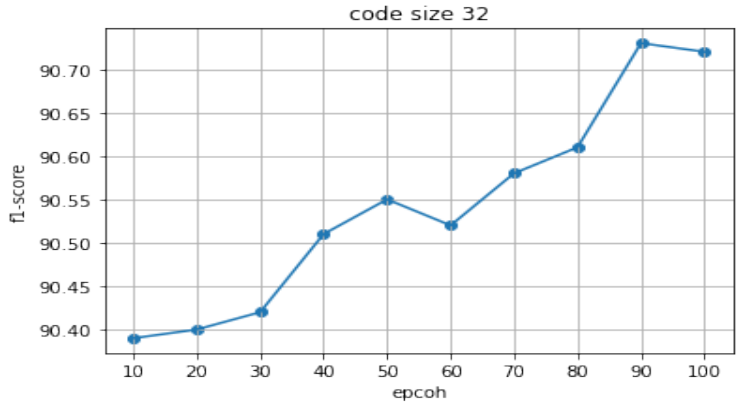


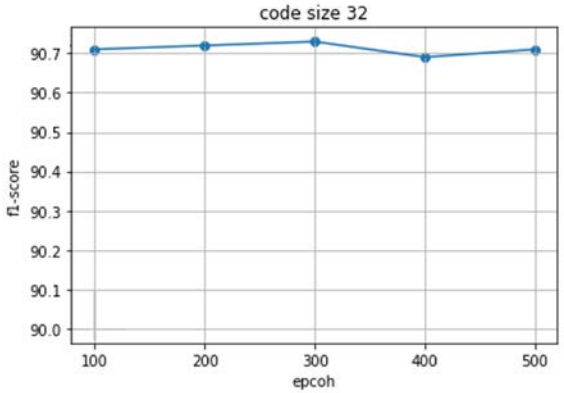
# 실험 및 결과 분석

- 오토인코더 모델 침입 탐지 실험결과



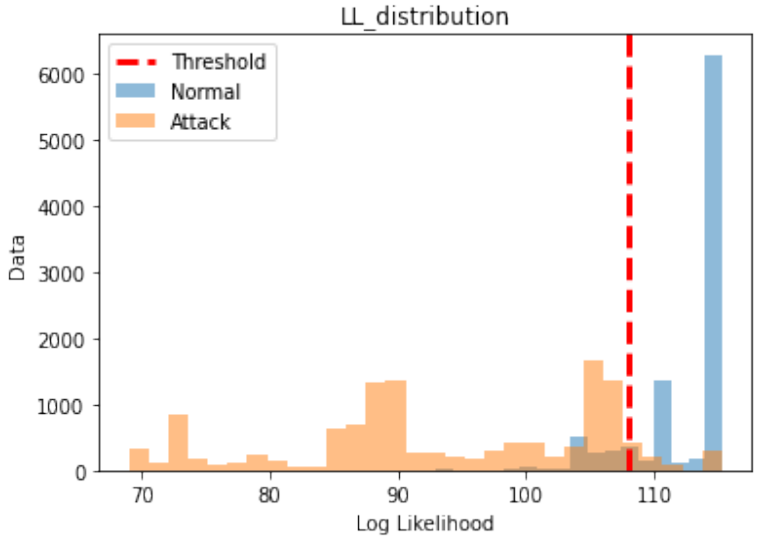



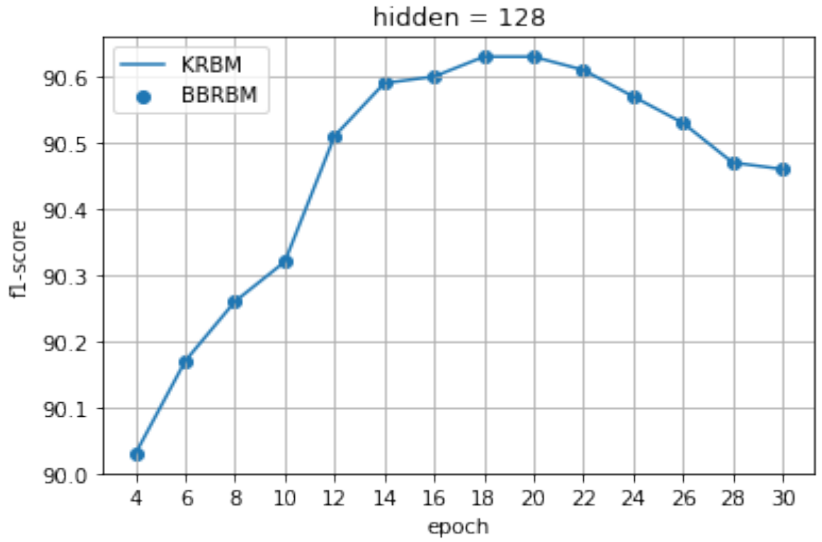



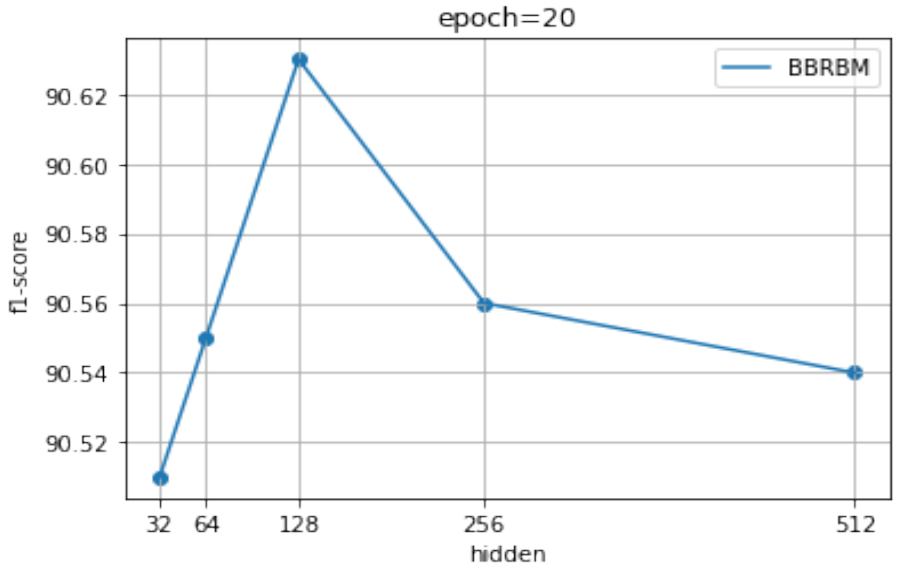



# 실험 및 결과 분석

- RBM 모델 침입 탐지 실험 결과
  - BBRBM 모델

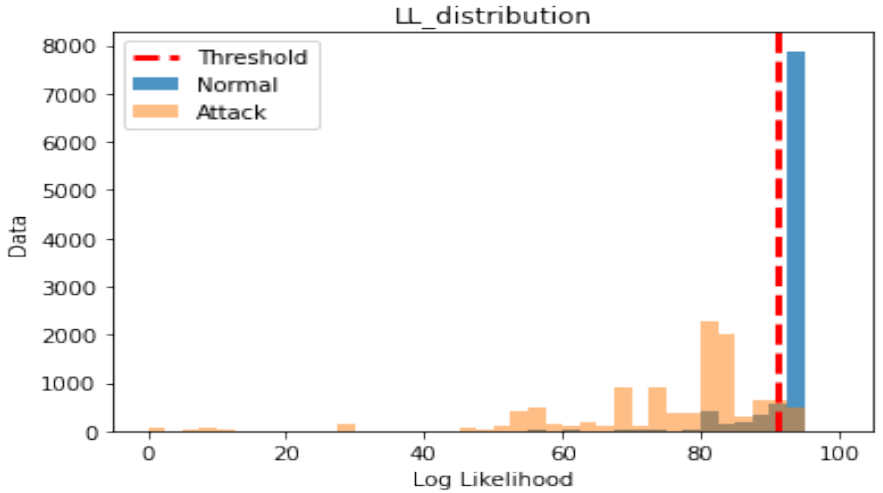


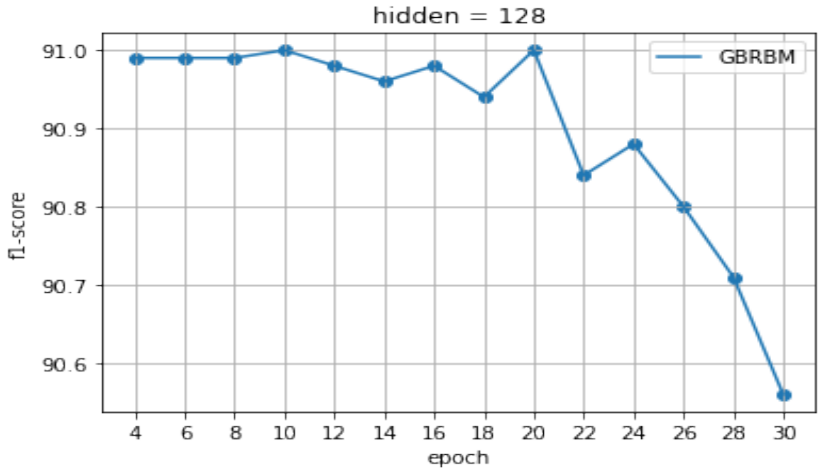



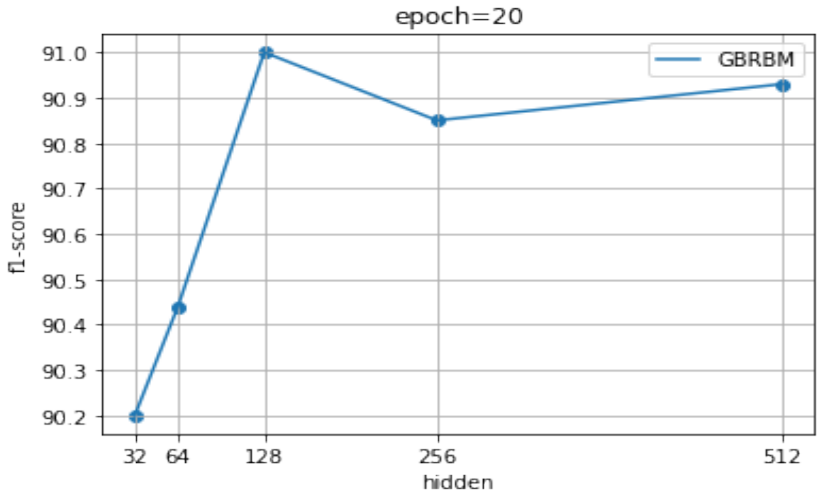



# 실험 및 결과 분석

- RBM 모델 침입 탐지 실험 결과
  - GBRBM 모델

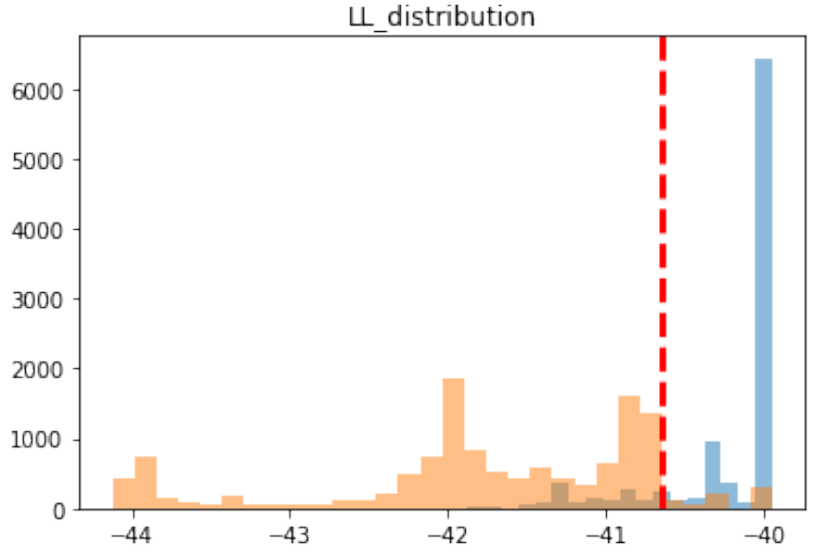


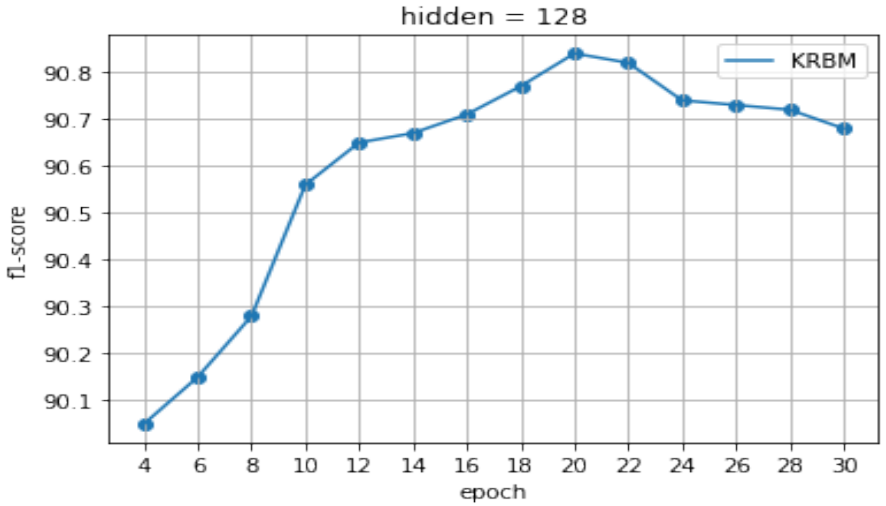



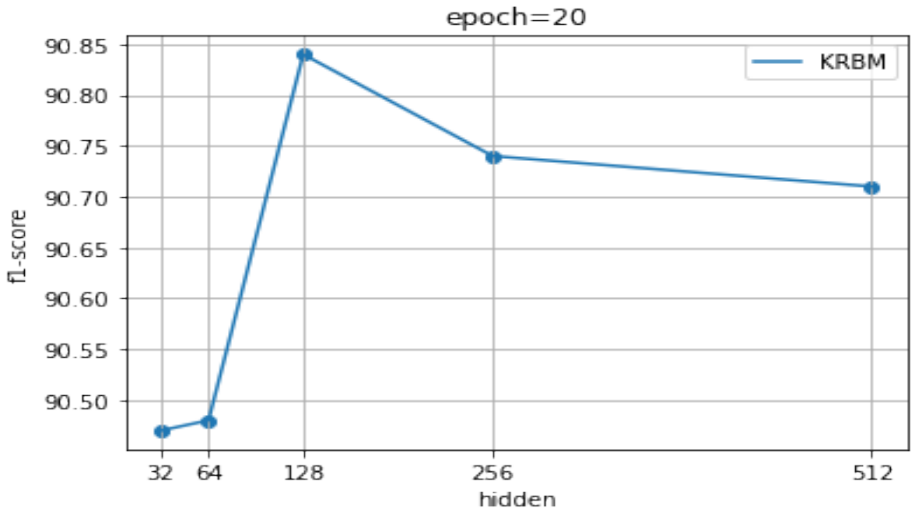



# 실험 및 결과 분석

- RBM 모델 침입 탐지 실험 결과
  - KRBM 모델

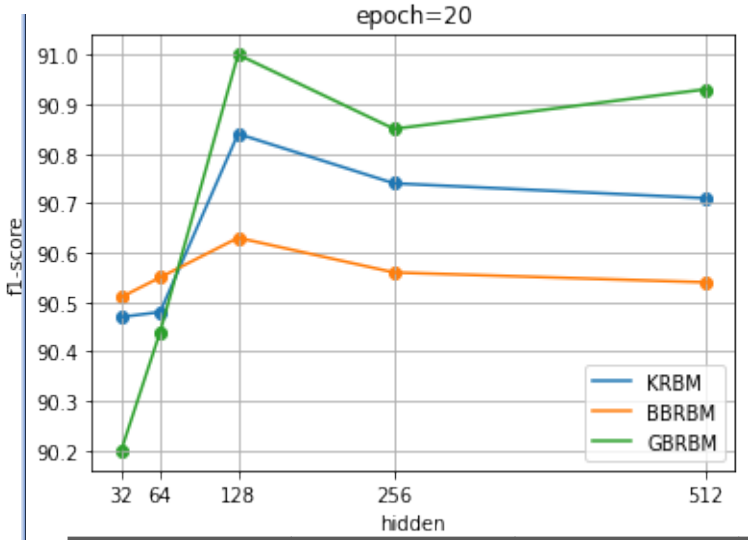


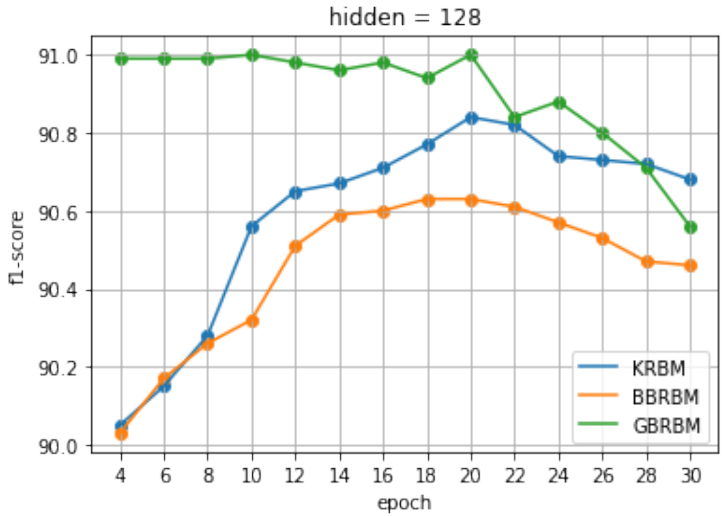



# 실험 및 결과 분석

- 실험정리




| model | 정확도 | 정밀도 | 재현율 | F1 score |
|---|---|---|---|---|
| KRBM | 92.38 | 87.62 | 94.31 | 90.84 |
| BBRBM | 91.83 | 87.61 | 93.85 | 90.63 |
| GBRBM | 92.51 | 87.81 | 94.45 | 91.00 |
| AE(32) | 92.23 | 89.93 | 91.57 | 90.73 |

| model | 소요시간 |
|---|---|
| KRBM | 229.89초 |
| BBRBM | 91.18초 |
| GBRBM | 94.92초 |
| AE(32) | 331.17초 |

# 결론 및 향후 연구 방향

- 결론
  - 실험에 사용한 모델 중 GBRBM을 활용한 모델이 가장 좋은 성능을 보임
  - 오토인코더 모델의 탐지 시간이 RBM 모델에 비해 더 많이 소요
  - RBM 모델은 네트워크 침입 탐지에 효과적인 모델임이 확인
- 향후 연구 방향
  - 다양한 데이터 셋을 활용하여 모델 검증
  - 제안된 RBM 모델 향상을 위한 추가 실험 진행

# Q&A