딥러닝 기반 암호화 트래픽 분류 데이터셋의 노이즈 분류 체계 및 정제 효과 분석

장윤성[1], 유경민[1], 남승우[1], 최양서[2], 김명섭[1†]

고려대학교[1], 한국전자통신연구원[2]

{brave1094, rudals2710, nam131119}@korea.ac.kr, yschoi92@etri.re.kr, tmskim@korea.ac.kr†

Noise Taxonomy and Refinement Effect Analysis for Deep Learning-based Encrypted Network Traffic Classification Datasets

Jang Yoon Seong[1], Yu Gyeong Min[1], Nam Seung Woo[1],

Choi Yang Seo[2], Kim Myung Sup[1†]

Korea Univ.[1], ETRI(Electronics and Telecommunications Research Institute)[2]

요 약

딥러닝 기반 네트워크 트래픽 분류(NTC) 연구는 공개 데이터셋을 활용하여 다양하게 진행되고 있으나, 수집·라벨링 과정에서 혼입된 노이즈를 육안으로 식별하기 어렵다는 도메인 특수성으로 인해 노이즈 혼입 여부를 충분히 검토하지 않은 채 데이터셋을 활용해 왔다. 본 논문은 딥러닝 학습 관점에서 NTC 데이터셋 노이즈를 정의하고 분류 체계를 수립하였으며, ISCX VPN 2016을 대상으로 15개 정제 규칙을 제안하고 전체 세션의 98.5%가 노이즈에 해당함을 확인하였다. 4개의 실험군 설계 및 비교 분석을 통해, 노이즈가 포함된 환경에서는 노이즈 분포 특성에 따라 모델 성능이 과대 또는 과소 추정될 수 있음을 밝히고, 노이즈 제거가 2D-CNN과 ET-BERT 두 모델에서 일관되게 분류 성능 향상과 학습 수렴 속도 개선 효과를 가져옴을 검증하였다. 본 결과는 NTC 분야에서 데이터셋 노이즈 정제의 필요성을 실험적으로 뒷받침하며, 체계적인 노이즈 정제 기반의 NTC 연구를 위한 기초를 제공한다.

Ⅰ. 서 론

네트워크 트래픽의 암호화가 보편화됨에 따라 전통적인 NTC 방식의 한계가 명확해졌으며, 이를 극복하고자 패킷의 원시 바이트 수준 특징을 직접 학습하는 딥러닝 기반 NTC 연구가 활발히 진행되고 있다. 딥러닝 기반 분류 연구에서는 학습 데이터에 혼입된 노이즈가 모델의 일반화 성능을 저하시킨다는 사실이 이미 보고되어 왔으며 [1, 2, 3], 이러한 문제는 딥러닝 기반 NTC 연구에서 더욱 심각하게 나타난다. 실제 네트워크 환경에서 수집된 트래픽에는 DNS·ARP 등 네트워크 제어 트래픽, 백그라운드 서비스 트래픽, CDN·SSO 트래픽 등 응용 식별과 무관한 세션이 불가피하게 혼입될 수 있으며, 이미지나 자연어 도메인과 달리 네트워크 트래픽 노이즈는 연구자가 육안으로 식별하기 어렵다는 근본적인 차이가 존재한다. 이로 인해 기존 NTC 연구에서는 노이즈 혼입 여부에 대한 검토 없이 데이터셋이 활용되어 왔을 가능성이 있으며, 트래픽 도메인에서 노이즈를 체계적으로 정의·유형화하는 기준과 그 학습 효과에 대한 실증적 검증이 부재한 실정이다.

본 논문은 이에 대응하여 다음의 세 가지 기여를 제시한다.

- NTC 데이터셋 내 노이즈를 딥러닝 학습 관점에서 정의하고, 기원과 태스크 관련성에 따른 3수준의 분류 체계(Taxonomy)를 수립한다.
- 공개 데이터셋 ISCX VPN 2016을 대상으로 노이즈 유형별 실증 분석을 수행하고, 필수 및 선택 제거 기준을 제안한다.
- 노이즈 포함·제거 조합의 4개 실험군 설계 및 오답 인스턴스 분석을 통해, 노이즈 제거가 Target Signal 학습 효율화에 미치는 효과를 정량적으로 검증한다.

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(00235509, ICT융합 공공 서비스●인프라의 암호화 사이버위협에 대한 네트워크 행위기반 보안관제 기술 개발), 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0024177, 2023년 지역혁신클러스터육성)

Ⅱ. 관련 연구

딥러닝 기반 분류 연구에서 노이즈가 학습에 미치는 영향은 다양한 관점에서 연구되어 왔다. Label Noise 분야에서는 잘못 라벨링된 학습 데이터가 모델의 일반화 성능을 저하시키며 [1], 노이즈 비율이 증가할수록 결정 경계가 왜곡된다는 사실이 보고되어 있다 [2]. Feature Noise 관점에서도 입력 특징에 혼재하는 불필요한 패턴이 Target Signal과 무관한 상관관계를 학습하게 유도할 수 있음이 확인되었다 [3].

NTC 분야에서도 데이터셋 품질 문제를 지적한 연구가 일부 존재한다. [4]는 기존 암호화 트래픽 분류 연구의 대다수가 레거시 데이터셋에 포함된 미암호화 트래픽을 활용해 왔음을 실증적으로 지적하였으며, [5]는 CIC-IDS2017에 대한 정밀 분석을 통해 피처 오계산, 라벨링 오류 등 다수의 품질 문제를 보고하였다. 그러나 이들 연구는 개별 데이터셋의 특정 문제를 사례로 보고하는 수준에 그치며, 딥러닝 학습 관점에서 노이즈를 체계적으로 정의·유형화하고 그 학습 효과를 실험 검증한 연구는 부재하다.

본 논문에서는 두 가지 딥러닝 기반 NTC 모델을 베이스라인으로 활용한다. 2D-CNN [6]은 원시 바이트를 2차원 이미지로 변환하여 분류하는 모델이며, ET-BERT [7]는 페이로드 바이트를 바이그램 토큰으로 변환하여 BERT 구조로 사전학습한 모델이다. 두 모델 모두 노이즈 혼입 여부 및 정제 과정에 대한 언급 없이 분류 성능을 보고하고 있으며, 본 논문에서는 두 모델을 활용하여 노이즈 제거 효과의 모델 독립성을 함께 검증한다.

Ⅲ. 본론

3.1 DL 기반 NTC에서의 노이즈 정의

딥러닝 기반 분류 모델의 학습 목표는 입력과 라벨 간의 관계를 가장

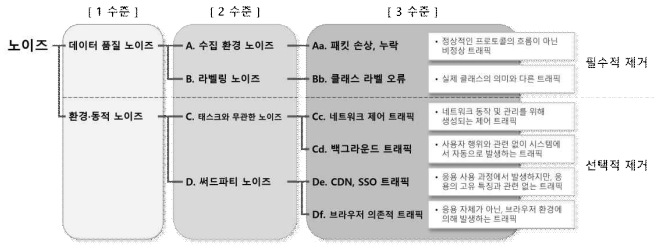


그림 1. NTC 분야 노이즈 분류 체계

잘 설명하는 함수, 즉 Target Signal을 학습하는 것이다. NTC 분야에서는 공격 트래픽 탐지, 공격 유형 분류, 응용 프로그램 분류, 서비스 분류, VPN·Tor 식별 등 다양한 태스크가 존재한다. 이러한 태스크별 클래스는 공격 기법부터 암호화 여부, 응용 프로그램에 이르기까지 다양하게 정의되며, 딥러닝 기반 NTC 모델의 학습 목표는 이러한 클래스를 구분하는 Target Signal을 학습하는 것이다. 본 논문에서 노이즈는 수집 및 라벨링 과정에서 혼입되었거나, 데이터셋에 정의된 클래스를 명확하게 구분하는 Target Signal의 학습을 저해하는 모든 세션으로 정의한다.

딥러닝 모델은 손실 최소화 방향으로 탐욕적으로 학습하는 특성상, 노이즈가 혼입된 데이터셋에서는 공격 유형이나 응용 프로그램 분류 자체가 아닌, 노이즈를 포함한 클래스 분류를 학습하게 될 수 있다. 본 문제는 4장의 실험을 통해 검증한다.

3.2 노이즈 분류 체계

데이터셋별로 노이즈의 종류와 존재 여부가 상이하고, 연구 목적 및 태스크에 따라 동일한 트래픽이 노이즈로 취급될 수도, 그렇지 않을 수도 있다. 본 논문에서는 그림 1과 같이 태스크와 데이터셋에 따라 노이즈 처리 여부를 유연하게 결정할 수 있도록 3수준의 분류 체계를 정의한다.

1수준은 태스크와의 관계성에 따라 데이터 품질 노이즈(필수 제거)와 환경·동적 노이즈(선택 제거)로 구분한다. 2수준은 노이즈 기원과 제거 방법론에 따라 A. 수집 환경 노이즈, B. 라벨링 노이즈, C. 태스크 무관 노이즈, D. 써드파티 노이즈의 4가지 범주로 세분화한다. 3수준은 트래픽의 구체적 특성에 따라 다음의 6가지 유형으로 세분화된다.

a. 패킷 손상·누락(Aa): TCP SYN 부재, Payload 없음, 패킷 손상·누락 경고 세션이 해당되며, 정상 트래픽 패턴을 반영하지 못하므로 필수 제거 대상이다.

b. 클래스 라벨 오류(Bb): 클래스 라벨과 실제 L7 프로토콜이 불일치하는 세션으로, 학습 오염을 유발하므로 필수 제거 대상이다.

c. 네트워크 제어 트래픽(Cc): ARP·ICMP 등 L3 제어 프로토콜, DNS·DHCP 등 주소 해석·할당 프로토콜, SMB·SSDP 등 Windows 인프라 프로토콜, SNMP·NTP 등 인프라 관리 프로토콜, 브로드캐스트 세션이 해당되며, 특정 응용에 귀속되지 않으므로 필수 제거 대상이다.

d. 백그라운드 트래픽(Cd): BT-DHT·uTorrent 등 P2P 백그라운드, WinHTTP·Watson 등 Windows 백그라운드, MS Delivery Optimization 세션이 해당되며, 태스크에 따라 선택적으로 제거한다.

e. CDN·SSO 트래픽(De): Google CDN 및 SSO 세션으로, 다양한 응용에 공통으로 발생하여 특정 클래스의 Target Signal로 귀속할 수 없으므로 선택적으로 제거한다.

f. 브라우저 의존적 트래픽(Df): 브라우저가 새 탭·홈페이지 로드 시 자동 발생시키는 MSN·Bing 요청 세션으로, 연구 목적에 따라 선택적으로 제거한다.

3.3 ISCX VPN 2016 데이터셋 적용 정제 규칙

본 논문은 ISCX VPN 2016을 대상으로 3.2절의 분류 체계에 근거한 15개 정제 규칙을 적용한다. 노이즈일 가능성이 있더라도 Target Signal의 일부일 가능성을 배제할 수 없는 경우에는 제거 대상에서 제외하는 보수적 기준을 채택하며, Aa → Bb → Cc → Cd → De → Df 순으로 순차 적용한다.

수집 환경 노이즈(Aa)는 TCP SYN 부재 세션(Aa_1), Payload 크기가 0인 세션(Aa_2), 패킷 손상·누락 경고가 포함된 세션(Aa_3)을 제거한다. 라벨링 노이즈(Bb)는 SCP 클래스 내 L7 프로토콜이 SSH가 아닌 세션(Bb_1)을 제거하며, L7 불일치가 명백히 확인되는 클래스에 한하여 적용한다. 태스크 무관 노이즈(Cc)는 ARP·ICMP 등 L3 제어 프로토콜(Cc_1), DNS·DHCP 등 주소 해석·할당 프로토콜(Cc_2), SMB·SSDP 등 Windows 인프라 프로토콜(Cc_3), SNMP·NTP 등 인프라 관리 프로토콜(Cc_4, VoIP 클래스의 RTCP 제외), 브로드캐스트·멀티캐스트 세션(Cc_5)을 제거한다. 백그라운드 트래픽(Cd)은 BT-DHT·uTorrent 백그라운드 세션(Cd_1), WinHTTP·Watson 등 OS 백그라운드 세션(Cd_2), MS Delivery Optimization 세션(Cd_3)을 제거한다. 써드파티 노이즈(De)는 Google CDN 세션(De_1, googlevideo 제외), SSO 관련 세션(De_2)을, 브라우저 의존적 트래픽(Df)은 MSN·Bing 요청 세션(Df_1)을 제거한다. 상기 15개 규칙을 전체 310,242개 세션에 순차 적용한 결과, 305,630개(98.5%)가 노이즈로 분류되었으며 4,612개(1.5%)만이 Clean 세션으로 확인되었다.

Ⅳ. 실험 결과 및 분석

본 장에서는 노이즈 정제가 딥러닝 기반 NTC 모델의 학습에 미치는 영향을 실험적으로 검증한다. 4.1절에서는 실험군 설계 및 데이터셋 구성과 전체 실험 결과를 제시하고, 4.2절에서는 실험군 1과 실험군 3의 비교를 통해 노이즈 혼입 학습이 Target Signal 분류 성능 및 학습 수렴 속도에

표 1. ISCX VPN 2016 데이터셋 정제 규칙 별 제거된 데이터 개수

Type	Rule	aim	bit torrent	email	face book	ftp	gmail	hangout	icq	netflix	scp	sftp	skype	spotify	vimeo	voip buster	you tube	Total
origin		451	477	7,610	82,907	947	449	91,820	467	510	11,304	244	106,432	383	579	4,554	1,108	310,242
Aa	Aa_1	37	0	24	214	11	36	176	36	293	75	30	432	82	152	0	350	1,948
	Aa_2	5	2	14	36	8	2	55	4	23	502	1	113	47	25	31	35	903
	Aa_3	0	177	0	5	53	0	13	0	7	30	11	135	6	13	2	30	482
Bb	Bb_1	0	0	0	0	2	0	0	0	0	1,083	0	0	0	0	0	0	1,085
Cc	Cc_1	0	0	0	2	3	0	0	0	0	0	0	2	0	0	3	0	10
	Cc_2	367	112	5,728	76,323	90	341	85,840	384	98	9,528	112	94,516	120	172	3,329	436	277,496
	Cc_3	0	1	90	141	5	0	196	1	4	41	7	282	9	5	8	14	804
	Cc_4	6	1	12	6	6	3	6	7	0	2	2	24	1	0	33	0	109
	Cc_5	0	0	1,260	5,434	0	0	4,846	0	1	0	0	8,591	5	2	0	1	20,140
Cd	Cd_1	0	0	0	0	544	0	0	0	0	0	0	0	0	0	935	0	1,479
	Cd_2	3	0	0	0	0	0	2	0	0	0	0	3	0	0	9	0	17
	Cd_3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
De	De_1	0	0	294	107	5	26	489	0	13	0	5	66	40	23	14	45	1,127
	De_2	0	0	0	12	0	0	9	0	0	0	3	6	0	0	0	0	30
Df	Df_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
clean		33	184	188	627	220	41	188	35	71	43	73	2,262	73	187	190	197	4,612

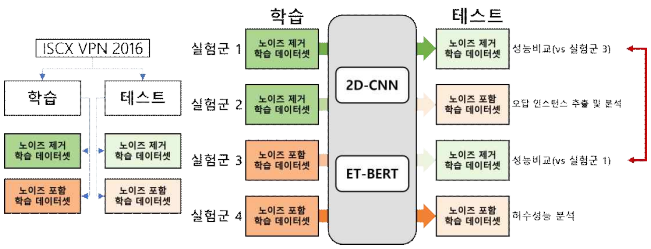


그림 2. 실험군 설계 및 데이터셋 구성

미치는 영향을 분석한다. 4.3절에서는 실험군 1과 실험군 4의 비교를 통해 노이즈 포함 평가 환경에서 발생하는 허수 성능 문제를 규명한다. 4.4절에서는 실험군 2의 오답 인스턴스를 분석하여 성능 저하의 원인이 데이터셋 자체의 분류 난이도가 아닌 노이즈에 있음을 직접적으로 입증한다.

4.1 실험 설계 및 데이터셋 구성

실험에는 ISCX VPN 2016 데이터셋을 사용한다. 해당 데이터셋은 VPN 및 비VPN 환경에서 수집된 암호화 트래픽으로 구성되며, 전체 310,242개 세션과 16개 응용 클래스로 구성된다. 3.3절의 15개 정제 규칙을 순차 적용한 결과를 표 1에 제시한다. 전체 세션의 98.5%에 해당하는 305,630개가 노이즈로 분류되었으며, 이 중 Cc_2(주소 해석·할당 프로토콜)가 277,496개(89.5%)로 가장 높은 비중을 차지하였다.

노이즈 정제 효과를 다각도로 검증하기 위해 학습·테스트 데이터의 노이즈 포함 여부를 조합한 4개의 실험군을 설계한다. 데이터셋은 층화 추출(80:20) 방식으로 분할하며, 분할 이후 각각에 동일한 정제 규칙을 독립적으로 적용하여 노이즈 포함·제거 버전을 생성한다. 평가 모델로는 2D-CNN [6]과 ET-BERT [7]를 사용하며, 분류 태스크는 서비스 분류(Service, 7클래스)와 응용 분류(Application, 16클래스)의 두 가지를 대상으로 한다. 실험군 설계는 그림 2와 같다.

각 실험군의 테스트 정확도 및 실험군 1 대비 성능 변화를 표 2에 제시한다. 세부 분석은 4.2절부터 4.4절에서 수행한다.

4.2 실험군 1 vs 실험군 3 :

노이즈 혼입 학습이 Target Signal 분류 성능에 미치는 영향

실험군 1(노이즈 제거 학습·노이즈 제거 테스트)과 실험군 3(노이즈 포함 학습·노이즈 제거 테스트)은 동일한 테스트셋에서 학습 데이터의 노이즈 포함 여부만 달리한 구성이다. 실험군 3의 학습 데이터 양이 더 많음에도 표 2와 같이 2D-CNN 기준 Service -5.51%p, Application -6.67%p의

표 2. 노이즈 포함 여부에 따른 성능 차이 비교

데이터셋 (태스크)	실험군	학습	테스트	CNN	Δ CNN	ET-BERT	Δ ET-BERT
ISCX VPN 2016 (Service)	1	제거	제거	79.35%	-	92.97%	-
	2	제거	포함	41.65%	-37.70%p	23.10%	-69.87%p
	3	포함	제거	73.84%	-5.51%p	92.65%	-0.32%p
	4	포함	포함	78.12%	-1.23%p	79.54%	-13.43%p
ISCX VPN 2016 (Application)	1	제거	제거	87.62%	-	93.22%	-
	2	제거	포함	26.18%	-61.44%p	11.74%	-81.48%p
	3	포함	제거	80.95%	-6.67%p	92.36%	-0.86%p
	4	포함	포함	60.30%	-27.32%p	64.88%	-28.34%p

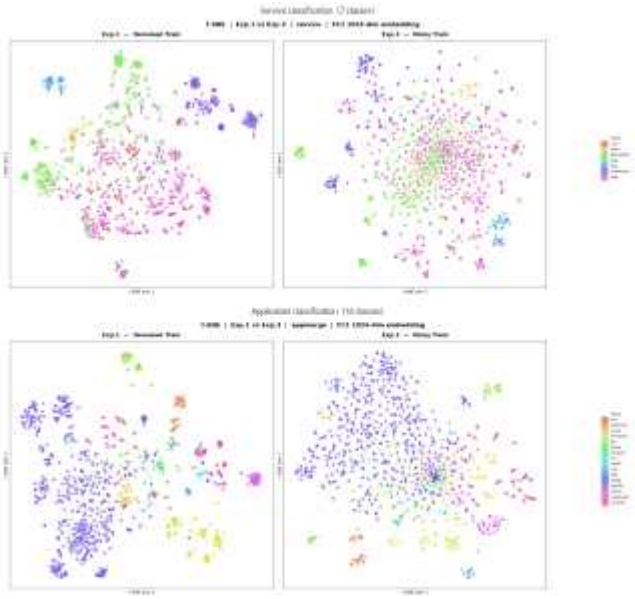


그림 4. 시각화 결과 – 실험군 1 vs. 실험군 3
(위 Service 분류/ 아래 Application 분류)

성능 저하가 나타났으며, 이는 데이터 양이 아닌 노이즈 혼입 자체가 Target Signal 학습을 저해함을 보여준다. ET-BERT는 성능 차이가 Service -0.32%p, Application -0.86%p로 비교적 영향이 덜 하지만, 그럼에도 노이즈 혼입에 따른 성능 저하가 확인된다.

그림 3의 학습 정확도 곡선에서 2D-CNN 실험군 1은 초기 수십 에폭 내에 100%에 근접하여 수렴하는 반면, 실험군 3은 60% 내외의 상한에 갇혀 1,000 에폭 내내 유의미한 향상을 보이지 못한다. ET-BERT는 두 실험군 모두 최종 수렴하나 실험군 1이 일관되게 더 빠른 수렴 속도를 보여, 사전학습 모델에서도 학습 효율성 차이는 존재함을 확인할 수 있다.

그림 4의 t-SNE 시각화에서 실험군 1은 클래스별 군집이 명확하게 형성되는 반면, 실험군 3은 데이터 포인트가 임베딩 공간 중앙으로 밀집되어 클래스 간 경계가 붕괴되는 양상을 보인다. 이는 노이즈 혼입 학습이 분류 성능뿐 아니라 모델의 내부 표현 공간 자체를 왜곡하는 것을 의미하며, 해당 패턴은 Service와 Application 태스크 모두에서 일관되게 나타난다.

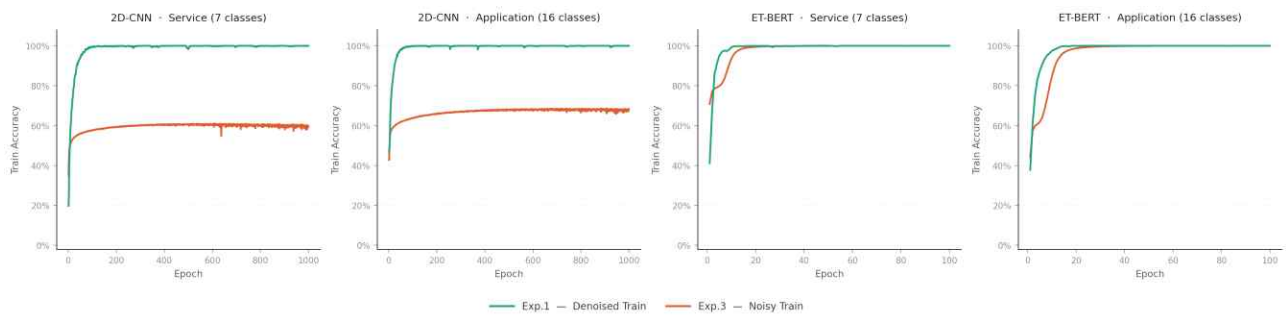
4.3 실험군 1 vs 실험군 4 :

노이즈 포함 평가 환경에서의 허수 성능

실험군 4(노이즈 포함 학습·노이즈 포함 테스트)는 노이즈 정제 없이 수행된 기존 연구 방식과 동일한 구성이다. 두 실험군의 테스트셋이 서로 다르므로, 단순 수치 비교가 아닌 각 실험군의 성능이 실제 태스크 수행 능력을 얼마나 정확하게 반영하는지를 검토한다.

표 2와 같이 2D-CNN 기준 Service 분류에서 실험군 4(78.12%)는 실험군 1(79.35%)과 유사한 수치를 보이나, 이는 실제 분류 능력이 아닌 허수 성능에 해당한다. 표 1에 따르면, voip 클래스가 전체 세션의 67.9%를 차지

그림 3. 2D-CNN, ET-BERT 모델 에폭별 학습 정확도 – 실험군 1 vs. 실험군 3



하고 Cc_2 노이즈의 69.8%가 voip에 집중 분포하는 구조에서, 모델이 DNS·DHCP 등 노이즈 패턴을 voip의 핵심 특징으로 잘못 학습하여 높은 정확도를 유지하는 것이다. 반면 Application 분류(-27.32%p)와 ET-BERT의 Service(-13.43%p), Application(-28.34%p)에서는 클래스별 노이즈 분포가 편중되지 않아 노이즈가 분류 단서로 활용되지 못하고 성능이 과소 추정된다. 이처럼 노이즈 분포 특성에 따라 평가 수치의 왜곡 방향조차 예측하기 어려우며, 이는 기존 NTC 연구 평가 방식이 내포하는 근본적인 한계이다.

4.4 실험군 2 오답 인스턴스 분석: 성능 저하 원인 규명

실험군 2(노이즈 제거 학습·노이즈 포함 테스트)는 오답 인스턴스 분석을 통해 성능 저하의 원인이 데이터셋의 분류 난이도가 아닌 노이즈에 있음을 직접 규명한다. 표 2와 같이 실험군 2의 정확도는 실험군 1 대비 2D-CNN 기준 Service -37.70%p, Application -61.44%p, ET-BERT 기준 Service -69.87%p, Application -81.48%p의 급격한 하락을 보이며, ET-BERT의 하락폭이 더 큰 것은 Target Signal에 특화된 모델일수록 노이즈 세션에 대한 예측 실패가 두드러지기 때문이다.

2D-CNN 기준 오답 인스턴스 중 노이즈 세션 비율은 Service 99.5%, Application 99.7%에 달하는 반면, Clean 세션의 오답률은 각각 20.8%, 13.6%에 불과하다. 이는 성능 하락의 원인이 데이터셋 고유의 분류 난이도가 아닌 노이즈 세션의 대량 오분류에 있음을 직접 입증한다. 그림 5의 혼동 행렬에서 2D-CNN과 ET-BERT 모두 voip 클래스의 노이즈 세션이 streaming, filetransfer 등으로 대규모 오분류되는 동일한 패턴이 관찰되며, 이는 해당 오분류가 모델의 구조적 한계가 아닌 노이즈 세션 자체의 속성에 기인함을 보여주는 모델 독립적인 근거이다.

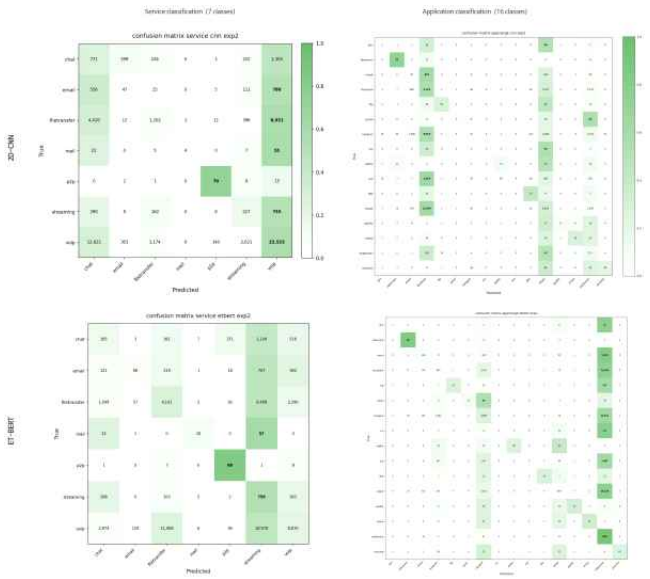


그림 5. 실험군 2 혼동 행렬 – Service 분류 (2D-CNN / ET-BERT)

V. 결론

본 논문은 딥러닝 기반 NTC 연구에서 충분한 검토 없이 이루어져 온 성능 검증 실험의 한계를 지적하고, 데이터셋 내 노이즈 제거의 중요성을 실험적으로 규명하였다. 본 논문의 세 가지 주요 기여는 다음과 같다. 첫째, NTC 분야에서 노이즈를 딥러닝 학습 관점에서 정의하고, 기원과 태스크 관련성에 따른 3수준의 분류 체계를 수립하였다. 둘째, ISCX VPN 2016을 대상으로 15개 정제 규칙을 적용하여 전체 세션의 98.5%가 노이즈에 해당함을 확인하고, 노이즈 제거가 모델 학습에 미치는 영향을 다각도로 검증하였다. 학습 수렴 속도 측면에서 노이즈 혼입 학습 환경은 최적화 방향을 왜곡하여 2D-CNN의 경우 1,000 에폭이 경과하도록 60% 수준의 상한에 수렴하지 못하는 반면, 노이즈 제거 시 초기 수십 에폭 내에 안정적으로 수렴하였다. 테스트 정확도 측면에서는 2D-CNN 기준 Service 5.51%p, Application 6.67%p, ET-BERT 기준 Service 0.32%p, Application 0.86%p의 성능 향상이 확인되었다. 클래스 간 특징 경계 학습 측면에서는 노이즈 혼입 학습 시 임베딩 공간 내 클래스 군집이 붕괴되어 클래스 간 경계가 불분명해지는 반면, 노이즈 제거 학습 시 클래스별 군집이 명확하게 형성됨을 t-SNE 시각화를 통해 확인하였다. 셋째, 노이즈가 포함된 평가 환경에서는 노이즈 분포 특성에 따라 성능이 과대 또는 과소 추정될 수 있으며, 오답의 99% 이상이 노이즈 세션에 기인함을 두 모델에서 일관되게 확인하였다.

이러한 결과를 바탕으로 데이터셋 수집자는 수집 환경과 라벨링 방식을 충분히 공개하여 사용자가 태스크에 따른 노이즈를 식별하고 제거할 수 있도록 해야 하며, 데이터셋 사용자는 태스크별 노이즈 유형을 파악하고 정제한 뒤 학습 및 평가를 수행해야 한다. 본 논문의 실험은 단일 데이터셋과 두 가지 모델을 대상으로 수행되었다는 한계가 있으며, 향후 더 다양한 데이터셋과 모델 아키텍처로 검증 범위를 확장함으로써 본 논문에서 제안한 노이즈 분류 체계의 일반화 가능성을 검토하는 연구가 이어지기를 기대한다. 이를 통해 NTC 연구 분야에서 보다 정확하고 신뢰할 수 있는 성능 검증이 이루어지기를 바란다.

참 고 문 헌

[1] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8135–8153, Nov. 2023.

[2] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 25, no. 5, pp. 845–869, May 2014.

[3] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zeiler, F. A. Wichmann, and W. Brendel, "Shortcut learning in deep neural networks," Nature Machine Intelligence, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[4] N. Wickramasinghe, A. Shaghaghi, G. Tsudik, and S. Jha, "SoK: Decoding the enigma of encrypted network traffic classifiers," arXiv preprint arXiv:2503.20093, May 2025.

[5] A. Rosay, E. Cheval, F. Carlier, and P. Leroux, "Network intrusion detection: A comprehensive analysis of CIC-IDS2017," in Proceedings of the 8th International Conference on Information Systems Security and Privacy (ICISSP), Feb. 2022, pp. 25–36.

[6] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in Proceedings of the 2017 International Conference on Information Networking (ICOIN), Da Nang, Vietnam, Jan. 2017, pp. 712–717.

[7] X. Lin, G. Xiong, G. Gou, Z. Li, J. Shi, and J. Yu, "ET-BERT: A contextualized datagram representation with pre-training transformers for encrypted traffic classification," in Proceedings of the ACM Web Conference (WWW), Lyon, France, Apr. 2022, pp. 633–642.

[8] Microsoft, "Delivery Optimization workflow, privacy, security, and endpoints," [Online]. Available: https://learn.microsoft.com/en-us/windows/deployment/do/delivery-optimization-workflow. [Accessed: Apr. 2026].