모델 선택 기법을 활용한 앙상블 기반 응용 트래픽 분류

백의준[1], 장윤성[2], 남승우[2], 김명섭[3]
[1]고려대학교 컴퓨터정보학과 박사과정
[2]고려대학교 컴퓨터정보학과 석사과정
[3]고려대학교 컴퓨터융합소프트웨어학과 교수
pb1069@korea.ac.kr, brave1094@korea.ac.kr, nam131119@korea.ac.kr, tmskim@korea.ac.kr

Ensemble-based Application Traffic Classification Using Model Selection

Ui-Jun Baek[1], Yoon-Seong Jang[1], Seung-Woo Nam[1], Myung-Sup Kim[2]
[1]Department of Computer and Information Science, Korea university
[2]Department of Computer Convergence Software, Korea university

요 약

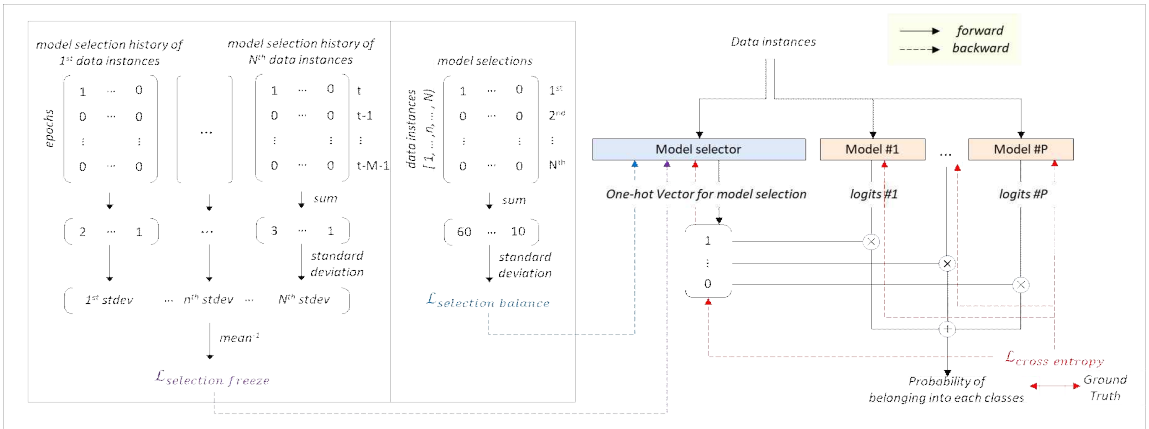
응용 트래픽이 점점 복잡해지고 방대해짐에 따라, 정확하고 효율적인 트래픽 분류에 대한 수요가 커지고 있습니다. 일반적으로 분류 모델의 크기와 분류 속도는 트레이드오프 관계이므로 두 개의 성능 목표를 달성하기는 어려우나 실제 환경에 분류 모델을 배포하기 위해선 더 빠르면서도 정확한 분류 기법이 필요합니다. 우리는 딥러닝 기반의 모델 선택기와 Gumbel-Softmax를 활용한 종단 간 앙상블 학습 방법을 제안합니다. 제안된 방법은 두 개의 공개 데이터셋을 사용하여 평가되었으며, 7개의 기준 모델들과 비교한 결과, 다른 방법에 비해 평균 4%의 정확도 향상을 보이면서도 합리적인 분류 속도를 유지했습니다.

1. 서론

트래픽 분류는 네트워크 트래픽을 다양한 응용 또는 서비스로 분류하는 작업으로 네트워크 관리자는 분류 결과를 통해 트래픽 패턴을 더 잘 이해하고, 네트워크 리소스를 효율적으로 할당하며, 악성 소프트웨어 또는 공격을 식별할 수 있습니다. 응용 트래픽 분류는 포트 기반 분류, 심층 패킷 검사, 행동 기반 분석과 같은 전통적인 방법에서 머신러닝, 딥러닝과 같은 보다 발전된 기술을 활용하고 있는 추세입니다. 최근에는 사전학습 기법을 활용하여 트래픽의 사전지식을 추출하고 학습하여 다양한 다운스트림 작업에 적용하고 있습니다. 이러한 발전에도 불구하고 실제 네트워크에서 트래픽 분류 모델을 배포하기 위해선 효율성, 배포가능성, 강건성, 신뢰성, 적응 가능성 등의 요구사항을 충족해야 합니다[1].

딥러닝 기술의 발전은 전통적인 방법 및 머신러닝 기반 접근 방식과 관련된 많은 한계점을 분명히 해결했으며, 복잡한 문제를 처리하는 데 뛰어난 성능을 보여주고 있습니다. 그러나 딥러닝 기반의 단일 모델은 과적합의 위험을 안고 있으며, 최적의 성능을 달성하기 위해서는 광범위한 하이퍼파라미터 튜닝이 필요합니다. 이러한 제한 사항을 극복하기 위해, 머신러닝 분야에서 두드러진 성과를 보인 앙상블 기법과 딥러닝을 결합하려는 시도가 증가하고 있습니다. 앙상블 방법은 여러 기본 모델을 결합하여 단일 모델로 만드는 것으로, 데이터의 다양성과 기본 모델의 강점을 활용하여 개별 모델에 비해 우수한 성능을 낼 수 있으며 과적합을 효과적으로 완화할 수 있습니다. 기본 모델이 어떻게 훈련되고 결합되는지에 따라 앙상블 방법은 평균화, 배깅, 랜덤 포레스트, 스태킹 및 부스팅과 같은 기술을 포함하며, 대부분의 연구는 배깅 방법을 사용하고 있습니다[2]. 그러나 전통적인 앙상블 방법은 단일 최종 분류 결과를 위해 여러 분류기가 반드시 작동해야 하므로 앙상블 모델의 효율성이 크게 저하됩니다.

본 논문은 응용 트래픽 분류 작업을 위한 종단 간 앙상블 학습을 위한 미분가능한 모델 선택 방법을 제안합니다. 제안된 방법은 다양한 분류 모델의 고유한 능력을 극대화하는 손실 함수를 통합하고, 개별 기본 모델의 사전학습 없이 종단 간 학습을 지원합니다. 또한, 모델 선택 메커니즘은 추론 과정에서 단일 모델의 출력만을 사용하여 효율성 측면에서 명확한 이점을 제공합니다. 논문의 서론에 이어 관



(그림 1) 제안하는 학습 과정 개요

련 연구, 제안 방법, 실험 결과 및 결론으로 구성됩니다.

2. 관련 연구

본 장에서는 모델 선택 기반의 앙상블 기법과 이를 활용한 일부 연구를 소개합니다.

모델 선택 기법은 다양한 정확도 프로필을 가진 모델들의 출력을 집계하는 방법으로 기존의 앙상블과는 달리 여러 모델의 출력 중 하나의 모델 출력만을 최종출력으로 사용합니다. 딥러닝 기반의 모델 선택에서는 검벨-소프트맥스를 활용합니다. 검벨-소프트맥스는 연속 확률적 접근 방식을 사용하여 카테고리형 이산 변수의 샘플링을 근사하며, 이산 변수의 미분을 가능하게 합니다. 검벨-소프트맥스는 최초로 VAE(Variational Auto Encoder)에서 카테고리형 데이터의 샘플링을 미분 가능하게 만들기 위해 사용되었으며[3], 이후 다양한 분야에서 적용되었습니다[4]. [4]는 텍스트 생성을 위한 LSTM(Long Short Term Memory)기반 생성기와 CNN(Convolutional Neural Network) 판별기로 구성된 TextGAN을 제안하였으며 텍스트 생성 모델의 이산 변수 학습 과정에서 검벨-소프트맥스를 활용했습니다. 최근에는 모델 선택 과정을 학습 단계에 통합한 미분가능한 모델 선택 방법이 제안되었습니다[5]. [5]는 여러 사전 훈련된 모델의 분류 결과를 기반으로 최종 분류 모델을 선택하는 E2E CEL 프레임워크를 소개하여 네 개의 인기 있는 데이터셋에서 높은 정확도를 달성했습니다. 그러나, 사전 훈련된 기본 분류기는 서로 독립적으로 학습되며, 딥러닝 기반 학습 과정의 탐욕적 특성으로 인해 유사한 구조를 가진 기본 분류기는 유사한 분포를 생성하는 경향이 있습니다. 이상적으로 기본 모델은 학습 과정에서 보완적인 출력을 생성해야 하며 우리는 기본 모델의 학습과 모델 선택 과정을 통합하여 종단 간 앙상블 학습 방법을 제안합니다.

3. 모델 선택 기법을 활용한 앙상블

제안하는 학습 방법의 개요는 그림 1과 같습니다. 그림 1의 오른쪽에는 모델의 간소화된 구조와 순방향 및 역방향 전파 과정이 설명되어 있습니다. 순방향 과정은 데이터 인스턴스를 각 기본 모델과 모델 선택기로 입력하는 것으로 시작됩니다. 데이터 인스턴스는 각 모델의 입력 형식에 맞게 전처리 됩니다. 각 기본 모델은 각 클래스에 속할 확률을 출력하며, 모델 선택기는 주어진 데이터 인스턴스에 대해 어떤 분류기의 출력을 사용할지를 결정합니다. 이를 위해 로짓을 출력하고, 이 로짓은 검벨-소프트맥스를 통과하여 최종적으로 원-핫 벡터를 생성합니다. 이후 각 모델의 로짓은 원-핫 벡터와 곱해지고, 결과는 각 클래스에 속할 최종 확률을 생성하기 위해 합산됩니다. 역방향 과정은 그림 2의 점선 화살표로 표시된 세 가지 손실 함수의 값을 기반으로 수행되며, 전체 손실 함수는 다음과 같습니다:

$$L_{total} = (L_{CE} \cdot W_{CE}) + (L_{SB} \cdot W_{SB}) + (L_{SF} \cdot W_{SF}) \quad (1)$$

첫 번째 손실 함수는 크로스 엔트로피 손실로, 실제 레이블의 확률 분포와 모델의 예측 확률 분포 간의 차이를 측정합니다. 수식은 다음과 같으며 p_i와 q_i는 각각 실제 레이블 확률 분포와 예측 확률 분포를 나타냅니다:

$$L_{CE}(p,q) = \sum_{i=1}^{k} p_i \log(q_i) \quad (2)$$

두 번째 손실 함수는 선택 균형(Selection Balance) 손실로, 각 기본 모델의 선택 빈도를 조절

하는 것을 목표로 합니다. 수식은 다음과 같으며 S_n은 n번째 데이터 인스턴스의 선택 결과의 원-핫 벡터이며 N은 데이터 인스턴스의 수를 나타냅니다:

$$L_{SB} = STDEV(\sum_{n=1}^{N} S_n) \quad (3)$$

세 번째 손실 함수는 선택 동결(Selection Freeze) 손실로, 각 데이터 인스턴스에 대한 선택 변경을 제한합니다. 각 학습 에포크에서 각 모델의 선택 빈도가 모든 데이터 인스턴스에 대해 집계되어 메모리 M에 저장되며 M은 5로 설정됩니다. 에포크 수가 M보다 커지면 메모리에 저장된 모델 선택 기록을 합산한 후 표준 편차를 계산하고 이를 손실로 사용합니다. 높은 표준 편차는 해당 인스턴스에 대한 모델 선택에 많은 변화가 있었음을 나타내며, 학습 과정에서 모델의 변동성을 증가시켜 학습을 방해할 수 있고 제안한 손실은 모델 선택 변동에 제약을 걸어 학습 과정의 안정성을 향상시킵니다. 수식은 다음과 같으며 S_{nm}은 인스턴스 n에 대해 m번째 메모리에 저장된 모델 선택 빈도를 나타내는 원-핫 벡터입니다:

$$L_{SF} = (\frac{1}{M}(\sum_{m=1}^{M} STDEV(\sum_{n=1}^{N} S_{nm})))^{-1} \quad (4)$$

제안하는 방법의 구현을 위해 공개된 딥러닝 기반 응용 분류 모델을 기본 모델로 사용하며 모델 선택 기는 트랜스포머 기반 경량 모델인 SAM을 사용하고 모든 입력 데이터는 각 기본 모델의 요구사항을 만족합니다.

제안하는 모델의 평가를 위해 ISCX-VPN 2016 [6]과 ISCX-Tor 2016 [7] 데이터셋을 사용하였습니다. 전처리 단계에서는 이더넷, IP, TCP, UDP 헤더를 모두 제거하고 페이로드만 남깁니다. 정제 과정에서는 응용의 동작과 관련 없는 DNS 프로토콜 세션을 제거하며 3-way 핸드셰이크 과정이 포함되지 않은 TCP 세션 또한 삭제됩니다. ISCX-VPN 2016 데이터셋은 약 만 개의 세션을 포함하며 ISCX-Tor 2016 데이터셋은 약 4만 개의 세션을 포함합니다.

4. 실험 및 평가

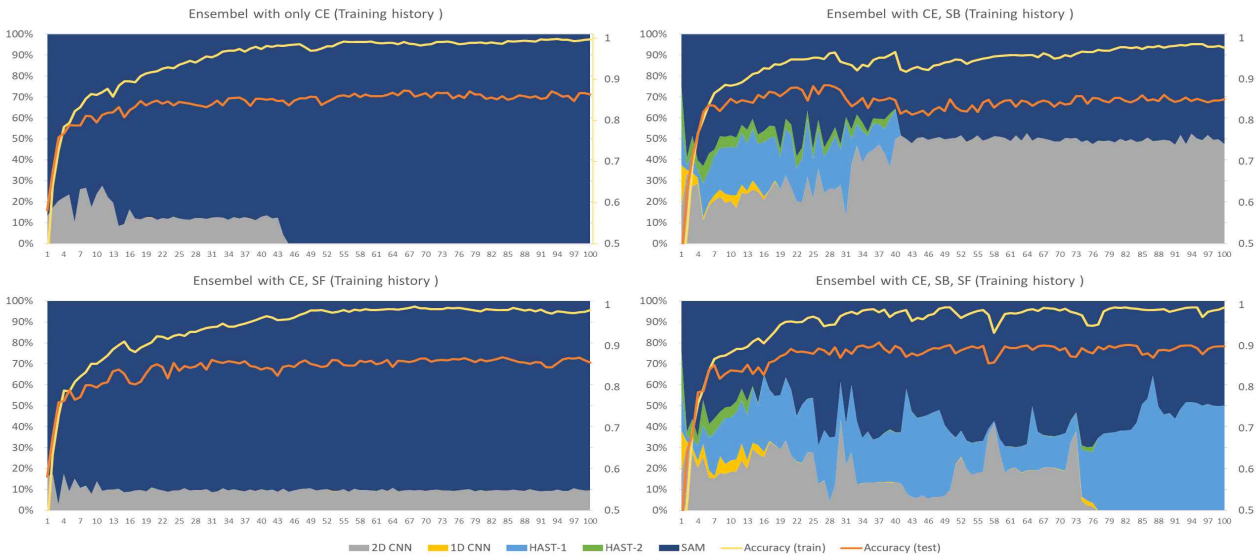
표 1은 제안된 방법과 기존 연구 간 정확도를 비교합니다. ET-BERT를 제외한 다섯 개 기본 모델의 결과를 결합한 Hard vote 방식이 앙상블 방법

(표1) 기존 연구와 제안하는 방법의 비교

기본 모델	ISCX VPN		ICSX Tor	
	응용 타입	응용	응용 타입	응용
2D CNN	78.6	71.3	96.0	94.9
1D CNN	79.4	72.9	91.3	89.7
HAST-1	86.8	78.2	97.2	96.1
HAST-2	83.9	79.9	96.6	96.4
SAM	84.6	81.0	95.0	94.8
ET-BERT	87.0	81.5	96.1	95.5
Hard vote	89.3	81.8	96.9	96.3
Soft vote	86.4	77.9	95.9	95.5
only CE	87.3	80.0	96.3	96.1
CE+SB	88.5	**82.6**	97.5	96.8
CE+SF	87.3	80.1	96.8	96.2
CE+SB+SF	**90.9**	81.9	**97.7**	**97.2**

중 가장 우수한 성능을 보입니다. 제안하는 방법은 두 가지 새롭게 제안된 손실 함수의 적용에 따라 4가지 경우로 나뉘며, 모든 손실 함수를 사용하는 것이 단 한 가지 경우를 제외하고 가장 높은 정확도를 보입니다. SB 손실은 CE 손실만 사용할 때에 비해 정확도를 크게 향상시키며 SF 손실은 CE 손실과 사용할 때에 비해 정확도를 향상시키지 않습니다. 다만, SB 손실과 함께 사용했을 때 약간의 정확도 향상을 보입니다. 이는 SF 손실이 학습 안정성을 강화하도록 설계된 반면, SB 손실은 학습의 다양성을 증가시키는 것을 목표로 하기 때문입니다. 이러한 결과는 그림 2의 학습 이력 비교에서 더 자세히 논의됩니다.

그림 2는 다양한 손실 함수 적용에 따른 학습 이력을 비교합니다. CE만 사용한 경우에는 처음엔 두 개의 모델이 사용되지만 훈련 중반부터는 하나의 모델만 사용합니다. CE와 SB를 사용한 경우는 CE만 사용한 경우 대비 학습 과정에서 다양한 모델의 출력을 사용하는 것을 관찰할 수 있습니다. CE와 SF를 사용한 경우는 다양한 모델의 출력을 사용하진 않지만 모델 선택 변동이 적습니다. 모든 손실 함수를 사용한 경우 학습 초기 모든 모델이 사용되지만, 중간 단계에서는 세 개의 모델만 활용됩니다. 이후 중후반에는 다시 모든 모델을 사용하려는 시도가 있으며, 결국 두 개의 모델만을 사용합니다. 이는 모델이 다른 경우들처럼 지역 최적값에 머물지 않고 지속적으로 전역 최적값을 추구한다는 것을 시사하며, 우리가 설계한 손실 함수의 적절성을 뒷받침합니다.



(그림 2) 손실 함수 적용에 따른 학습 과정 비교

5. 결론

본 논문은 모델 선택 기법을 활용한 종단 간 앙상블 학습 방법을 제안합니다. 제안된 방법은 두 개의 데이터셋에서 평가되었으며 설계된 두 가지 손실 함수가 의도한 대로 동작함을 실험적으로 확인했습니다. 제안한 방법은 기존 연구 대비 높은 정확도를 달성했을 뿐만 아니라 기존의 앙상블 방법과 달리 수평 모델 확장 과정에서 약간의 오버헤드만 요구하여 모델의 효율성을 크게 향상합니다. 오픈소스 분류 모델뿐만 아니라 최신 고성능의 응용 트래픽 분류 모델을 활용한다면 더욱 높은 정확도를 달성할 수 있을 것이라 기대되며 2가지 향후 연구가 요구됩니다. 첫 번째는 기본 모델 및 모델 선택기의 동작 과정을 설명할 수 있는 설명가능한 인공지능에 대한 연구이며 두 번째는 모델 훈련의 안정성과 다양성에 대한 추가적인 연구입니다.

Acknowledgements

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) 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과이고 (2021RIS-004) 본 논문은 2024년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력 기반 지역혁신 사업 (2021RIS-004)과 2023년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원(P0024177, 2023년 지역혁신클러스터육성)을 받아 수행된 연구입니다.

참고문헌

[1] ACETO, Giuseppe, et al. AI-powered Internet Traffic Classification: Past, Present, and Future. IEEE Communications Magazine, 2023.

[2] MIENYE, Ibomoiye Domor, et al. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 2022, 10: 99129-99149.

[3] JANG, Eric, et al. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:1611.01144, 2016.

[4] ZHANG, Yizhe, et al. Adversarial feature matching for text generation. In: International conference on machine learning. PMLR, 2017. p. 4006-4015.

[5] KOTARY, James, et al. Differentiable model selection for ensemble learning. arXiv preprint arXiv:2211.00251, 2022.

[6] GIL, Gerard Drapper, et al. Characterization of encrypted and VPN traffic using time-related features. In: Proceedings of the 2nd international conference on information systems security and privacy (ICISSP 2016). SciTePress, 2016. p. 407-414.

[7] LASHKARI, Arash Habibi, et al. Characterization of tor traffic using time based features. In: International Conference on Information Systems Security and Privacy. SciTePress, 2017. p. 253-262.