동적 컨볼루션과 다중 스케일 시공간 주의 메커니즘을 사용한 비디오 분류 모델

김지하, 박현희*
명지대학교

{ yaki5896, hhpark*}@mju.ac.kr

Video Classification Model using Dynamic Convolution and Multi-Scale Spatio-Temporal Attention Mechanisms

Jiha Kim, Hyunhee Park*
Myongji Univ.

요 약

영상을 분류하기 위해서는 영상의 시공간 정보를 모두 학습해야 한다. 빠른 영상에서는 시퀀스가 조금만 변해도 프레임 내의 전역적인 변화를 보일 수 있으며, 정적인 영상에서는 프레임의 일부분만 변화하는 특징을 가지고 있기 때문이다. 따라서 본 논문에서는 영상의 시공간적 정보를 함께 처리하기 위한 dynamic multi-scale temporal spatio-temporal attention network (DyMSTA-Net)을 제안한다. DyMSTA-Net 은 동적 컨볼루션을 이용하여 비디오 시퀀스의 각 샘플에 맞춰 적응적인 필터를 사용한다.또한 multi-scale 을 통해 다양한 시간 프레임에서의 정보를 병렬로 학습하며, 시공간 어텐션 메커니즘을 이용하여 시간적 정보와 공간적 정보 중 어떤 부분을 집중할지 효율적으로 학습한다.

Ⅰ. 서 론

최근 영상정보의 활용은 다양한 분야에서 사용될 수 있다. 이에 따라 최근 영상 분류 모델에 대한 학습 성능이 크게 증가하고 있으며 다양한 학습 모델이 제안되고 있다. Dynamic convolution [1]에서는 기존의 단일 컨볼루션 필터 대신 여러 개의 병렬 필터를 사용하여 각 필터의 중요도를 동적으로 계산한다. TimeSformer[2]은 vision transformer (ViT)를 기반으로 한 모델이다. 이는 2D, 3D convolution neural network (CNN) 없이 순수하게 self-attention 만으로만 영상을 학습한다.

하지만 영상 내의 정보를 학습하기 위해서는 시간축의 정보를 이용해야 한다. 일반적으로 영상 데이터에서는 일정한 시간을 두고 연속된 프레임의 변화에 따라 영상의 속도감을 이해할 수 있다. 따라서 본 논문에서는 시공간의 정보를 적응적으로 학습할 수 있는 dynamic multi-scale temporal spatio-temporal attention network (DyMSTA-Net)을 제안한다. DyMAST-Net 은 동적 컨볼루션 필터와 multi-scale 을 통한 다양한 시간축에서의 특징, 시공간 어텐션 메커니즘을 통한 각 축에서의 집중도를 효율적으로 학습할 수 있다.

Ⅱ. 본론

DyMSTA-Net 은 제안하는 3 개의 모듈과 특징 추출을 위한 사전학습 모델, 최종 분류 모델로 이루어진다. 3 개의 모듈은 각각 동적 가중치 학습 (dynamic 3d convolution), 다중 시간 축 특성 학습 (multi-scale temporal block), 시공간 주의 메커니즘 (spatio-temporal attention)으로 구성한다.

Dynamic 3d convolution 은 입력 데이터 $\boldsymbol{X} \in \mathbb{R}^{B\times T\times H\times W\times C}$ 에 대하여 컨볼루션 필터 $\boldsymbol{K}_{\text{dynamic}}$ 를 생성한다. 여기서 B,T,H,W,C 는 각각 배치 크기, 입력 시간축, 높이, 너비를 의미한다. $\boldsymbol{K}_{\text{dynamic}}$ 는 base filter $\boldsymbol{K}$ 와 입력 데이터 $\boldsymbol{X}$ 에 대한 평균 값 $\overline{\boldsymbol{X}}$ 를 활성함수를 통해 비선형성을 더하여 동적 가중치 W_dynamic 을 생성한다. W_dynamic 을 base filter 에 곱하는 것으로 $\boldsymbol{K}_{\text{dynamic}}$ 을 생성한다.

Multi-scale temporal block 은 입력 데이터 X 에 대하여 3D 컨볼루션 연산을 수행한다. 3D 컨볼루션 연산에 사용되는 필터는 $T_{short}, T_{medium}, T_{long}; (T_{short} < T_{medium} < T_{long})$ 로 각각 시간축에서 단기, 중기, 장기 정보를 추출하는 필터로 사용한다. 추가적으로 max pooling 을 통해

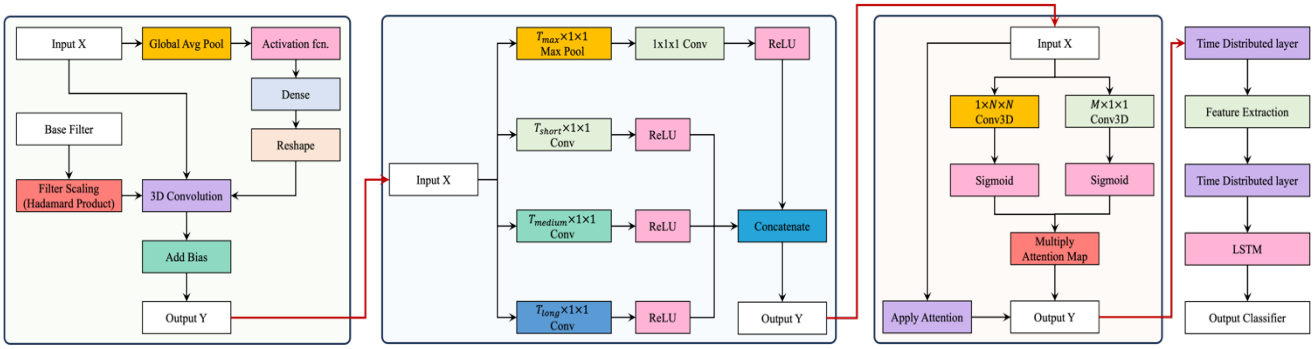

그림 1. DyMSTA-Net architecture

모든 시간축에서의 정보를 전체 범위에서 압축하여 하나의 시간적 특징만 남긴다. 이를 통해 multi-scale temporal block 은 여러 시간 범위에서 발생하는 패턴을 종합하여 영상 데이터의 다양한 시간적 특성을 학습할 수 있다.

Spatio-temporal attention 메커니즘은 비디오 내에서 중요한 시간적, 공간적 위치를 모델이 병렬적으로 학습할 수 있다. 이는 특정 프레임 내에서의 중요한 공간적 위치 (예: 움직임이 일어나는 부분)와 전체 시퀀스에서 중요한 시간적 정보 (예: 중요한 사건이 일어나는 프레임)을 강조할 수 있다. 이를 통해 영상 데이터에 대한 속도감을 모델이 학습하여 특징으로 사용할 수 있다.

Ⅲ. 시뮬레이션 결과

학습에 사용된 데이터는 UCF 101 데이터셋[3]으로, 101 가지 행동으로 분류되는 영상 데이터이다. 학습 과정은 모델의 성능이 saturation 되는 시점인 100 epoch 까지 진행한다.

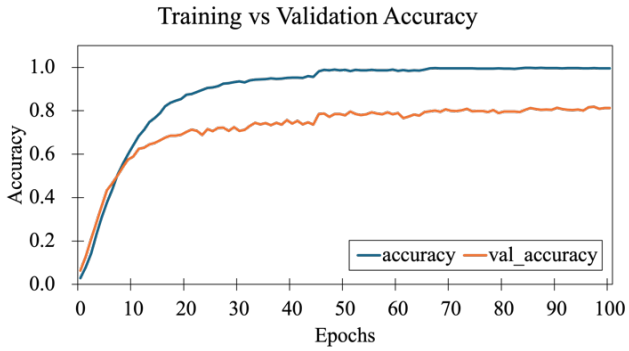

그림 2. 학습 과정 중 정확도 변화

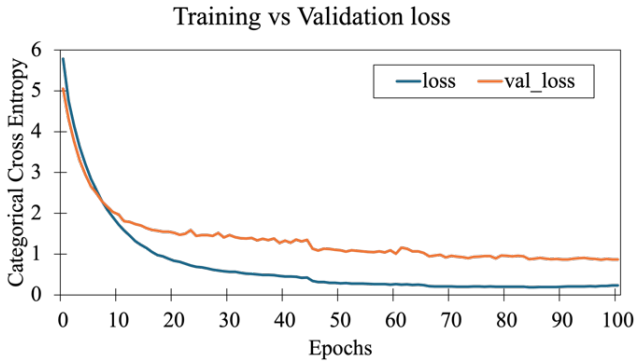

그림 3. 학습 과정 중 손실 변화

그림 2 와 그림 3 은 학습 과정에서의 성능 변화를 보여준다. 학습 결과 모델은 초기 단계에서 매우 빠르게 학습하여 saturation 될 수 있는것을 알 수 있다. 훈련 데이터에서는 90%가 넘는 학습 성능을 보여주고 있다. 하지만 10 epoch 부터는 과적적합이 발생하는 것으로 보인다. 이를 위해 learning rate 를 동적으로 조율하여 과적합이 발생함에도 불구하고 검증 데이터를 통한 일반화 성능이 향상되는 것을 알 수 있다. 또한 학습 과정에서 손실 감소 패턴으로 보아, learning rate 가 적절히 조정되었음을 알 수 있다. 특히, 학습 후반에 안정적인 성능을 보이는 것은 learning rate 를 조절하는 것이 효과가 있음을 알수 있다.

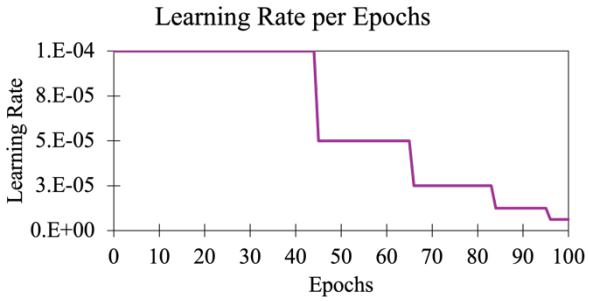

그림 4. 학습 진행에 따른 learning rate 변화

표 1 은 precision, recall, f1-score 성능지표를 이용한 테스트 결과를 보여준다. 3 가지 성능지표 모두 비슷한 수준 (0.79~0.80)으로 균등한 성능을 보여준다. 이는 클래스 불균형 문제가 어느정도 해소되었음을 알 수 있다. 특히 f1-score 가 0.79 인 것으로 보아 모델은 어느정도 실용적인 성능을 보인다는 것을 알 수 있다. 최대 상위 k 개의 정확도를 측정하는 top-k accuracy 에서도 k 가 5 일 때 정확도 0.9294 로 측정된다.

표 1. Performance metrics

	Precision	Recall	F1-score
Accuracy			0.80
Macro avg	0.79	0.79	0.78
Weighted avg	0.80	0.80	0.79

Ⅳ. 결론

본논문에서는 영상 분류를 위한 DyMSTA-Net 모델을 제안한다. DyMSTA-Net 은 dynamic 3d convolution, multi-scale temporal block, spatio-temporal attention 의 3 개의 모듈로 시공간적 데이터 학습을 효과적으로 할 수 있다. 학습 결과 검증 데이터에서의 정확도는 최대 약 0.8128, 손실은 최소 약 0.8685 로 일반화 된 성능을 보여준다. 하지만 UCF 101 데이터셋은 단조로운 동작의 영상이 많기 때문에 동작 변화에 민감도에 따른 성능은 아직 부족한 것으로보인다. 향후 HMDB 51, Kinetics 과 같은 역동적인 데이터셋을 활용할 연구를 진행할 계획이다.

ACKNOWLEDGMENT

이 논문은 정부(교육부)의 재원으로 한국연구재단과 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2022R1A2C2005705, 분산 머신 러닝 기반 지능형 플라잉 기지국을 위한 AI-MAC 프로토콜, No. 2021-0-00990, 설명가능한 인공지능 기반 무선랜 네트워크 시스템 고도화 핵심 기술 연구)

참 고 문 헌

[1] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., and Liu, Z. "Dynamic convolution: Attention over convolution kernels," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11030-11039, 2020.

[2] Bertasius, G., Wang, H., and Torresani, L. "Is space-time attention all you need for video understanding?," Proc. ICML, Vol. 2, No. 3, p. 4, July 2021.

[3] Soomro, K., Zamir, A. R., and Shah, M. "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," CRCV-TR-12-01, Nov. 2012.