Ⅰ. Large-scale Video Classification with Convolutional Neural Networks[1]

❒ summary

❍ CNN 아키텍처 훈련을 위해 487개 스포츠 클래스 분류에 속하는 1백만 개의 YouTube 동영상 데이터 셋 구축

❍ CNN 구조에서 비디오의 로컬 모션 정보 활용을 위한 시간적 연결 패턴의 적합성, 추가 모션 정보가 CNN에 미치는 영향에 대해 초점을 둠.

❍ low-level feature를 용도 변경 하여 UCF-101 데이터 세트에서 성능 증가(41.3→65.4%)

❍ 과정

- local visual feature 추출
- 고정된 크기의 video-level description으로 결합(일반적으로 k-means dictionary 사용)
- Bag Of Words(BOW)를 통한 classifier(e.g., SVM) 훈련

❒ good points

❍ 입력 차원 감소로 중심 구조가 2~4배 더 빨라짐.

❍ 단일 프레임 모델에 대해 속도 향상이 있었음.

❒ disadvantages

❍ 계산상의 제약으로 인해 CNN은 최근까지 비교적 저해상도 이미지 인식 문제에 적용됨.

- e.g., MNIST, CIFAR-10, etc.

❍ 학습 중 validation error가 개선되지 않으면 직접 학습률을 변경해야 함.

❒ suggestions for the improvement

❍ 이미지의 회전이나 크기 조정으로 인한 변경이 생기면 식별이 어려워지는 것에 대한 해결책 필요

- 알려진 방법 중 하나는 4D 또는 6D 맵으로 AI를 학습시키는 것
- 그러나, 엄청난 비용을 감수해야 하는 방법임.

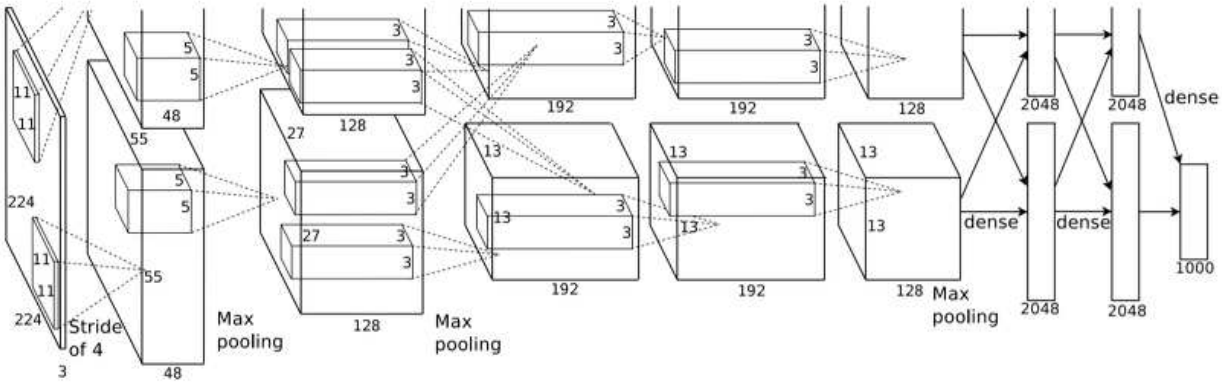
II. ImageNet classification with deep convolutional neural networks[2]

❒ summary

❍ ImageNet 120만 개의 데이터를 1,000개의 클래스로 분류하는 Deep CNN 학습

❍ 6,000만 개의 파라미터, 65만개의 뉴론, 5개의 Conv Layer, Max-pooling Layer 적용, 3개의 FC Layer를 통해 1,000개의 클래스 분류

❍ 학습 속도 향상을 위해 Convolution시 GPU 사용 및 드롭아웃을 통해 과적합 감소



<그림1> 논문에서 사용된 CNN 모델 구조

❒ good points

❍ ReLU 비선형 활성 함수를 사용하여 학습 속도 향상

❍ ReLu 적용 후 지역 반응 일반화 적용

- 1, 2번째 Conv Layer에 적용

❍ Overlapping Pooling 적용

- 1, 2, 5번째 Conv Layer에 적용

❒ disadvantages

❍ 누락되거나 명확하지 않은 이미지에서의 사용은 어려움.

- 고해상도의 고정된 크기의 이미지를 입력으로 사용한 결과물임.

❒ questions

❍ 결론 부분에 사전 훈련된 비지도 학습을 사용하지 않음을 명시

- 만약 사용했다면, 네트워크의 깊이가 더 증가했을 거라고 가정

- 비지도 학습을 사용했을 때 정확도가 더 증가했다면, 적용을 했을지 의문

❒ suggestions for the improvement

❍ 사전 훈련된 비지도 학습 적용과 그렇지 않았을 때의 비교 결과 첨부

Ⅲ. Convolutional neural networks for sentence classification[3]

❒ summary

❍ 기존의 CNN 구조를 모방하여 나온 모델

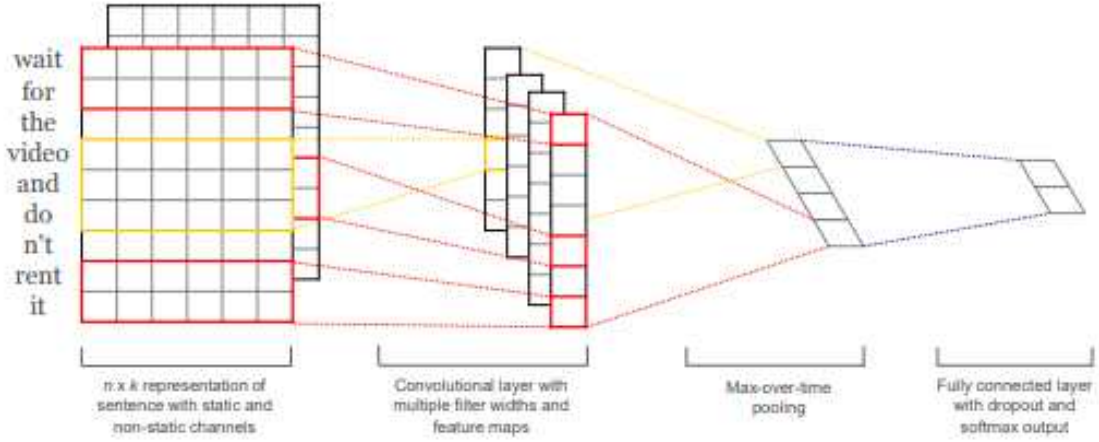
❍ 문장에 대한 단어의 개수 n, 차원 수 k일 때, n * k로 구성된 입력 채널

- 정적, 비정적 채널로 존재

❍ 2개의 채널에 대해 Convolution하여 하나의 Conv 층을 이룸.

❍ Conv 층으로부터 나온 feature map을 max pooling하여 penultimate Layer 생성

❍ FC Layer에 dropout, softmax를 적용하여 라벨에 대한 확률 분포를 만들어 최종 분류 수행



<그림2> 논문에서 사용한 CNN 구조

❒ good points

❍ CV 분야에서 집중적으로 사용하는 CNN을 NLP에 이용

❍ Max pooling을 통해 가장 큰 값(문장을 잘 표현하는 특징)만 사용

- 불필요 데이터를 제거하여 과적합 억제

- 전체 데이터 감소에 따른 계산 비용 저하, 속도 증가

❒ disadvantages

❍ early stopping이 추가된 것 외에 특별한 사항이 없음.

❍ 학습 결과, 데이터 셋에 따라 성능 상이

- 매 순간 multi-channel의 성능이 좋은 것은 아님.

❒ suggestions for the improvement

❍ 논문에서 Collobert(2011)의 word vector를 사용하면 성능이 낮다는 결과

- word vector 구축 문제인지, 1,000억 개의 구글 뉴스 데이터 문제인지 제시 필요

IV. 3D Convolutional neural networks for human action recognition

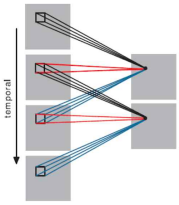
❒ summary

❍ 감시 카메라 영상에서의 사람 행동 인식 방법에 대한 연구

❍ 2차원 영상을 위한 CNN 모델을 시계열 데이터를 고려하여 3차원 CNN 모델로 제안

❍ 2D Convolution 개념과 인접한 프레임과의 Convolution 연산이 결합된 구조 제안

<그림3> 2D Convolution



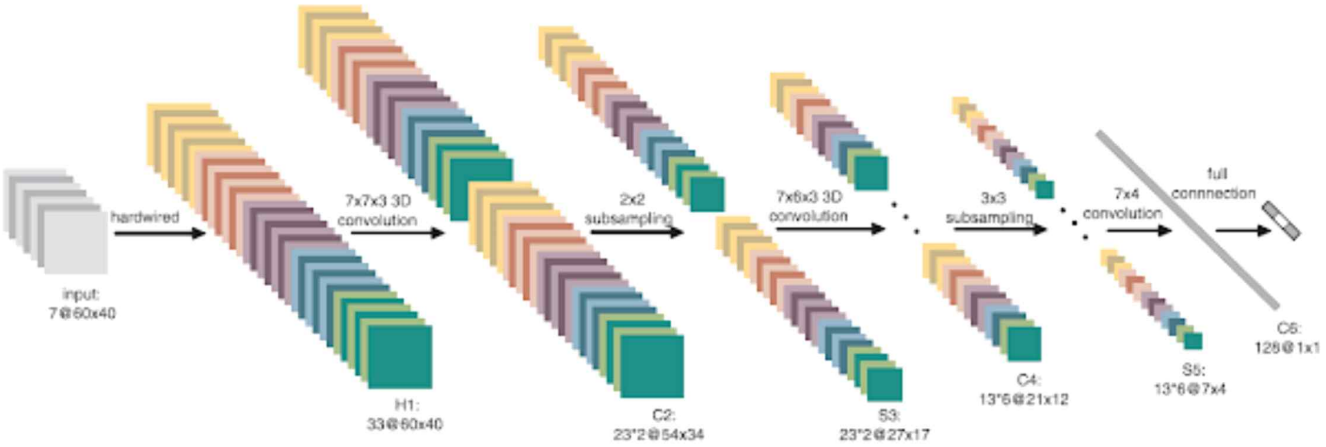
<그림4> 3D Convolution

❍ 제안된 3D CNN 모델은 weight가 공유되지 않는 형태로 Feature map 생성

❒ good points

❍ ‘hardwired’ 라는 커널을 사용하여 33개의 feature map 생성

- 입력 영상 전처리를 진행하는 커널
- 수평과 수직 방향 각각에 대해 회색조 영상의 기울기 값, optical flow 값을 구함.
- 논문 저자의 노하우가 반영된 커널로 생각됨.



<그림5> 3D CNN 구조

❒ disadvantages

❍ (상기 good points 참고) 33개의 feature map 중 28개의 feature map 설명되었지만 나머지 5장의 설명은 존재하지 않음.

V. Deep Residual Learning for Image Recognition(ResNet)[5]

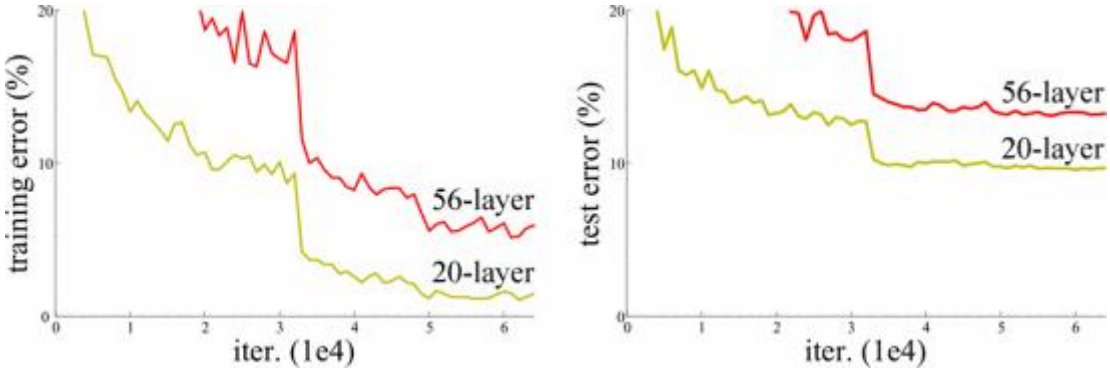
❒ summary

❍ 단순 다중 레이어는 네트워크의 성능을 좋게 하는 요인이 아님.

- 대표적으로 기울기 소실/ 폭발 문제 발생
- 네트워크 깊이가 깊어질수록 정확도가 떨어지는 현상 발생
- 이 문제는 과적합 문제가 아니기 때문임.

❍ 이 문제는 과적합 문제가 아님

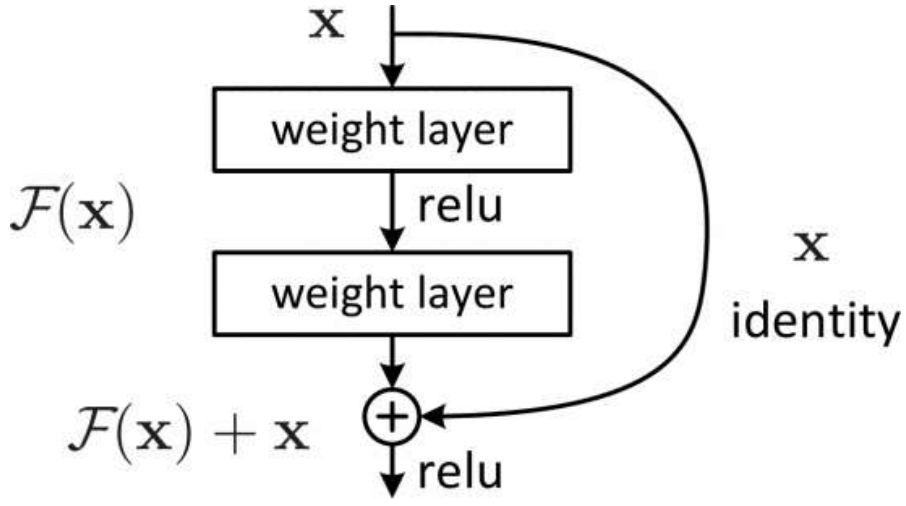
- 일반적으로 깊은 층의 학습 정확도는 높고 실험 정확도는 낮아야 함.
- 상기 문제의 경우 학습, 실험 정확도 모두 낮음.



<그림6> 20, 56개 네트워크를 사용하는 CIFAR-10의 훈련 오류(좌) 및 테스트 오류(우)

❍ ‘Deep Residual Learning framework’ 개념 도입

- 기본 매핑 $H_{(x)}$일 때, 누적된 비선형 레이어가 $F_{(x)} := H_{(x)} - x$의 다른 매핑에 맞도록 함.
- 참조되지 않은 원래 매핑을 최적화하는 것보다 잔여 매핑을 최적화하는 것이 더 쉽다는 가정
- 원래 매핑은 $F_{(x)} + x$로 캐스팅 되어 Shortcut Connection과 동일하고, 하나 이상의 레이어를 스킵
- Shortcut Connection은 추가적인 파라미터, 복잡한 곱셈 연산 불필요



<그림7> Residual learning: a building block.

❒ good points

❍ 심층 네트워크에 최적화가 쉽게 잘 됨.

- 1~2자리 수 단위 층에서 100단 단위 층까지 생성 가능

❍ inception, bottle-neck 구조를 통해 파라미터를 획기적으로 감소

❍ short-cut(출력+=입력) 도입

- 최종 출력에서 얻어내는 receptive field의 다양화

❒ disadvantages

❍ 입력, 출력의 차원을 통일시켜야 함.

❍ 층수가 1000개를 넘어서면 다시 정확도가 떨어지는 현상 발생

❒ suggestions for the improvement

❍ 기존의 개선된 형태와 부족했던 수식적 설명에 대한 보충이 다음 논문에서 진행[6]

Reference paper list

[1] Karpathy, A., Toderici, G., Shetty, S., Sukthankar, R., Li, F.-F.(2014). Large-scale video classification with convolutional neural networks. In *CVPR*, 6909619, pp. 1725-1732

[2] Krizhevsky, A., Sutskever, I., Hinton, G.E.(2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), pp. 84-90

[3] Kim, Y.(2014). Convolutional neural networks for sentence classification. In *EMNLP* 2014, Proceedings of the Conference, pp. 1746-1751

[4] Ji, S., Xu, W., Yang, M., Yu, K.(2013). 3D Convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1),6165309, pp. 221-231

[5] He, K., Zhang, X., Ren, S., Sun, J.(2016). Deep residual learning for image recognition. In *CVPR*, 7780459, pp. 770-778

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun.(2016). Identity mappings in deep residual networks. *Lecture Notes in Computer Science*, 9908 LNCS, pp. 630-645