다중 스케일 특성 생성 네트워크*

이광한†, 신새별††, 우사이먼성일†††
성균관대학교 인공지능학과†
성균관대학교 소프트웨어융합학과††
성균관대학교 데이터사이언스융합학과†††
{ican0016, toquf930, swoo}@g.skku.edu

Efficient Multi-Scale Feature Generation Network

Gwanghan Lee†, Saebyeol Shin††, Simon S. Woo†††
Department of Artificial Intelligence, Sungkyunkwan University†
College of Computing and Informatics, Sungkyunkwan University††
Department of Applied Data Science, Sungkyunkwan University†††

요 약

조기 종료 네트워크(early-exit network)는 추론 시 동적으로 모델 복잡도를 낮춤으로써 신경망의 효율성을 높인다. 기존 연구들은 입력 샘플이나 모델 구조의 중복성(redundancy)을 줄이는 데 집중하였으나 고차원 특징 정보가 부족한 초기 분류기들이 전체 네트워크 성능에 치명적인 영향을 끼치는 문제를 해결하지 못했다. 본 연구는 중복성을 줄이는 것뿐만 아니라 합성곱 커널(convolution kernel) 중앙에서 가중치들을 공유하면서 효율적으로 다중 스케일(multi-scale) 특징을 생성하여 조기 종료 네트워크의 성능을 향상시킨다. 또한 이 논문의 게이팅 네트워크(gating network)는 네트워크의 서로 다른 위치에 있는 각 합성곱 레이어에 따라 최적의 다중 스케일 특징 비율을 결정하도록 학습된다.

1 서 론

심층 신경망(Deep Neural Network)이 이미지 인식, 객체 탐지와 같은 다양한 컴퓨터 비전 분야에서 사용되면서 성능 향상을 위해 더 많은 레이어들을 쌓아 모델 깊이를 증가시키는 방법들이 제안되었다. 그러나 모델 크기가 커져감에 따라 파라미터(parameter) 수와 연산량 또한 크게 증가하였고, 이로 인해 모델을 모바일 디바이스와 IoT 기기들에서 사용하기 어렵다는 문제가 대두되었다. 이러한 연산량을 줄이기 위해 가지치기(pruning), 양자화(quantization), 지식 증류(knowledge distillation) 등의 딥러닝 경량화 기법 [1, 2]이 제안되었다. 최근 위의 경량화 기법 외에 조기 종료 네트워크(early-exit network) [3, 4]가 주목을 받고 있는데, 조기 종료 네트워크는 레이어 중간에 보조 분류기(auxiliary classifier)가 구성되어 있어서 입력 샘플에 대한 예측 신뢰도(confidence)가 특정 임곗값보다 높으면 예측이 조기 종료되어 계산 비용을 절약할 수 있다. 하지만, 단순히 모델의 복잡도(블록, 레이어)를 줄이는 데에만 집중하고 고차원 특징 정보가 부족하여 성능이 낮은 얕은 보조 분류기(shallow auxiliary classifier)는 학습 시 네트워크 백본에 영향을 미쳐 전체적인 네트워크의 성능 저하를 유발한다는 문제가 있다.

본 논문에서는 위의 문제를 해결하기 위해 효율적인 다중 스케일 특성 생성 네트워크 (Efficient Multi-Scale Feature Generation Network)를 제안한다. 본 논문에서 제안하는 방법은 레이어마다 다중 스케일 특징들을 적절하게 생성하여 보조 분류기가 전체 네트워크 성능 저하에 미치는 영향을 줄였을 뿐만 아니라 보조 분류기의 성능을 향상시켰다. 또한, 효율적으로 다중 스케일 특징들을 만들어내는 EMG (Efficient Multi-Scale Feature Generation) Module을 제안한다. EMG Module은 5×5 컨볼루션 커널과 그 중앙의 파라미터를 공유하는 3×3 컨볼루션 커널을 사용하여 효율적으로 다중 스케일 특성을 생성할 뿐만 아니라 게이팅 네트워크 (gating network)를 통해 입력 샘플에 따라 그림 1과 같이 적절하게 다중 스케일 특징을 사용한다. 그 결과 기존의 조기 종료 네트워크에 비

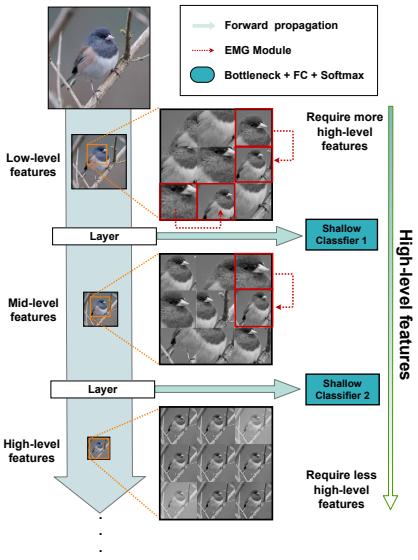


그림 1: 조기 종료 네트워크의 깊이에 따라 적절하게 고려되는 수용영역 (receptive field)

* 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이고(No. 2022-0-01199, 융합보안대학원(성균관대학교)), (No.2019-0-00421, 인공지능대학원지원(성균관대학교)), (No.2021-0-02309, 영상 품질 저하 조건 하에서의 저전력 · 고성능 객체 탐지 기술 연구) 과학기술정보통신부 · 광주광역시가 공동 지원한 '인공지능 중심 산업융합 집적단지 조성사업'으로 지원을 받아 수행된 연구 결과입니다.

해 CIFAR, ImageNet 데이터셋에서 연산량 대비 더 높은 정확도를 달성했다.

2 관련연구

최근 주목받는 조기 종료 네트워크는 보조 분류기들이 네트워크의 여러 위치에 연결되어 있는 형태를 가지며 각 입력 샘플들에 대해 예측 신뢰도에 따라 조기 종료를 수행한다. 보조 분류기를 효과적으로 학습하기 위해 SCANNet [4]은 어텐션 레이어를 추가하였으며 SD-ResNet [3]은 최종 분류기의 지식을 보조 분류기에 전달하기 위해 자가 증류(self distillation)를 도입했다. 하지만, 기존의 연구들은 네트워크에서 모델 복잡도를 줄이는 데에만 집중하였기 때문에 낮은 성능의 보조 분류기가 네트워크 전체의 성능 저하를 일으킨다는 문제점이 있었다.

3 방법론

3.1 조기 종료 네트워크

조기 종료 네트워크는 p 개의 보조 분류기가 네트워크의 깊이에 따라 구성되어 있으며 입력 샘플 x에 대해 p 개의 예측값을 생성하는데 아래와 같이 표현할 수 있다.

$$[y_1, ..., y_p] = f(x; w) = [f_1(x; w_1), ..., f_p(x; w_p)], \quad (1)$$

f_p 와 w_p는 각각 p번째 보조 분류기와 p번째 보조 분류기의 파라미터를 의미한다. 추론 시에는 특정 보조 분류기 예측값의 소프트맥스 값이 특정 임곗값보다 클 경우에 해당 위치에서 조기 예측을 하게 된다.

3.2 효율적인 다중 스케일 특성 생성 네트워크

기존 조기 종료 네트워크의 컨볼루션 레이어에서 이루어지는 연산은 $Y = X * f$와 같이 표현할 수 있고 $Y \in \mathbb{R}^{n\times w'\times h'}$, $X \in \mathbb{R}^{c\times w\times h}$, $f \in \mathbb{R}^{c\times k\times k\times n}$, 와 $*$는 각각 출력 특성, 입력 특성, 컨볼루션 필터 그리고 컨볼루션 연산을 의미한다. 또한, w, h, w', h', c 와 n은 각각 입력 너비, 입력 높이, 출력 너비, 출력 높이, 입력 채널수와 출력 채널수를 의미한다.

기존의 1×1 컨볼루션 레이어와 다르게, EMG module은 그림 2와 같이 1×1 컨볼루션 연산 후 다중 스케일 컨볼루션 연산을 하는 두 개의 과정으로 이루어진다. 본 논문의 1×1 컨볼루션 연산은 $Y' = X * f'$와 같이 표현할 수 있고 $Y' \in \mathbb{R}^{m\times w'\times h'}$, $f' \in \mathbb{R}^{c\times 1\times 1\times m}$는 각각 출력 피쳐와 컨볼루션 필터로 정의되며, m은 출력 채널 수로 본 논문의 실험에서는 $m = 0.5n$ 으로 설정했다. 1×1 컨볼루션 연산 후 다중 스케일 특성을 생성하기 위한 연산은

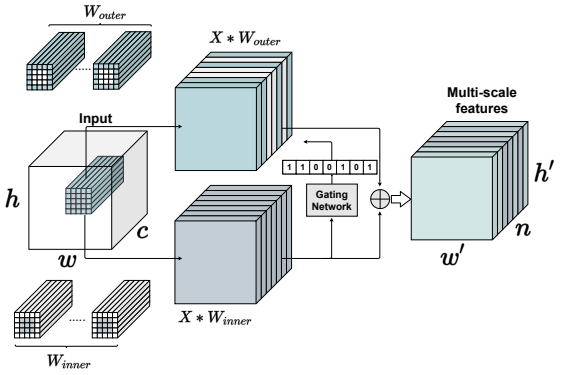


그림 2: EMG module의 전체 구조

아래와 같이 정의한다.

$$Y_i'' = \begin{cases} (Y' * W_{inner})_i, & \text{if } D_i = 0 \\ (Y' * W_{inner} + Y' * W_{outer})_i, & \text{otherwise} \end{cases}, \quad (2)$$

Y'', W_{outer}, W_{inner} 와 i는 각각 다중 스케일 출력 특성, 5×5 컨볼루션 커널, 3×3 컨볼루션 커널과 출력 채널 인덱스를 의미한다. 본 논문의 방법은 3×3 컨볼루션 커널로 연산 후 게이팅 네트워크를 통해 나온 D 텐서가 더 큰 스케일의 특성이 필요한지 판단하고 필요하다고 판단될 경우 5×5 컨볼루션 커널로 추가 연산을 하여 최종적으로는 1×1, 3×3 그리고 5×5 컨볼루션 연산을 통해 생성된 다중 스케일 특성을 출력한다. 이때, 5×5 컨볼루션 연산은 3×3 컨볼루션과 가중치를 공유하기 때문에 출력 특성을 효율적으로 재사용하여 더 큰 스케일의 특성을 생성할 수 있다.

3.3 네트워크 학습

본 논문에서 제안하는 EMG module은 입력 샘플에 따라 다중 스케일 특성을 적절하게 생성하기 위해 게이팅 네트워크를 사용하고, 아래와 같이 정의한다.

$$H(x)_i = \begin{cases} 1, & \text{if } x_i \geq 0 \\ 0, & \text{otherwise} \end{cases}, \quad (3)$$

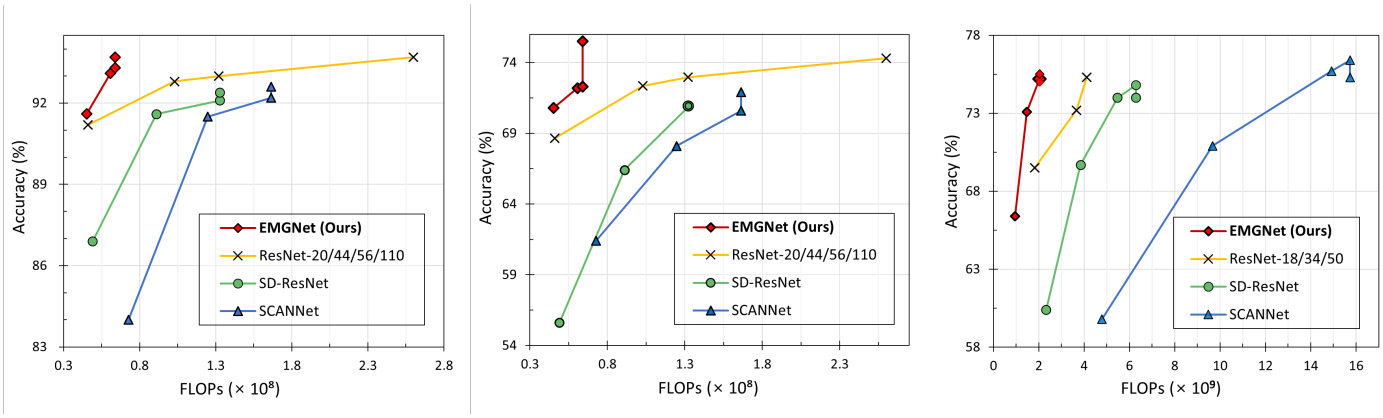
$$S(I, t)_i = H(\sigma(GlobalAvgPool(|I| - t) - 0.5)_i), \quad (4)$$

$$D_i = S(I, t)_i, \quad (5)$$

H, S, σ, $t \in \mathbb{R}^m$ 와 I는 각각 이진 단계 함수, 게이트 함수, 시그모이드 함수, 임곗값 그리고 $Y' * W_{inner}$을 의미한다.

본 논문의 네트워크를 학습시키기 위한 최종 로스 함수는 아래와 같이 표현할 수 있다.

$$\mathcal{L}_r = \sum_{j=1}^{E} \left(\sum_{i=1}^{m} exp(-t_{ij}) \right), \quad (6)$$



(a) CIFAR-10 (b) CIFAR-100 (c) ImageNet

그림 3: CIFAR-10, CIFAR-100 그리고 ImageNet 데이터셋에서 각 네트워크의 보조 분류기별 정확도(%)와 부동소수점 연산(FLOPs)을 통한 비교.

$$\mathcal{L} = \sum_{i=1}^{p} (\mathcal{L}_{ce})_i + \alpha \mathcal{L}_r, \tag{7}$$

$\mathcal{L}_{ce}$, $\mathcal{L}_r$, α 와 E는 각각 분류에 쓰이는 크로스엔트로피 로스, 다중 스케일 특성의 비율을 조절하는 정규화 로스, 스케일 팩터 그리고 컨볼루션 레이어의 수를 의미한다. 본 논문에서는 α에 따라 다중 스케일 특성의 비율을 조절했으며, 가장 효과적인 α를 찾기 위해 그리드 서치를 수행하였다.

4 실험 결과

베이스라인 모델은 SD-ResNet [3], SCANNet [4] 그리고 ResNet [5] 총 3가지 모델로 설정하며 EMGNet에 있는 각 분류기의 결과를 그림 3에 베이스라인 모델들의 조기 종료 값들과 비교하고 정확도와 FLOPs로 나타낸다. 그림 3에서 실선에 나타난 각각의 점은 네트워크 중간에 붙은 각 보조 분류기 성능을 나타내고 마지막 점은 보조 분류기의 예측값을 앙상블한 성능을 나타낸다.

우선 CIFAR 데이터셋의 경우 EMGNet은 동일한 보조 분류기 위치에서 SD-ResNet에 비해 1.2%에서 15.2%, SCANNet에 비해서는 1.1% 에서 9.4% 만큼의 정확도 향상을 보였다. 또한, ImageNet 실험에서 EMGNet은 동일한 각 보조 분류기에서 SD-ResNet에 비해 0.2% 에서 6%, SCANNet에 비해 0.2% 6.6% 만큼 정확도가 향상되었다. 또한 FLOPs를 비교해보면 ImagNet에서 EMGNet은 SD-ResNet과 SCANNet에 비해 각각 44.91%와 73.22% 만큼 감소되었다.

5 결론

기존의 조기 종료 네트워크는 모델의 복잡도를 낮추는 방법에만 집중했고 이에 따라, 보조 분류기의 낮은 성능이 네트워크 전체의 성능 저하를 일으키는 문제를 해결하지 못했다. 본 논문에서는 새로운 EMGNet을 제안하여 얕은 보조 분류기의 성능을 향상시키기 위해 EMG module을 통해 부족한 다중 스케일의 특성을 생성하였고, 보조 분류기의 위치에 따라 특성을 적절하게 사용할 수 있도록 게이팅 네트워크를 제안하였다. 제안된 방법은 CIFAR, ImageNet 데이터셋에서 기존 연구에 비해 보조 분류기들의 성능이 향상되었고, 향후 본 논문의 방법론은 모바일 디바이스와 IoT 기기와 같은 모바일 환경에서 효율적으로 사용될 수 있을 것으로 기대된다.

참고 문헌

[1] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[3] L. Zhang, J. Song, A. Gao, J. Chen, C. Bao, and K. Ma, "Be your own teacher: Improve the performance of convolutional neural networks via self distillation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3713–3722, 2019.

[4] L. Zhang, Z. Tan, J. Song, J. Chen, C. Bao, and K. Ma, "Scan: A scalable neural networks framework towards compact and efficient models," *arXiv preprint arXiv:1906.03951*, 2019.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.