Self-Distillation 을 활용한 Few-Shot 학습 개선

김태훈, 주재걸
고려대학교 컴퓨터학과
e-mail : sunshine1kim@korea.ac.kr, jchoo@korea.ac.kr

Improving Few-Shot Learning through Self-Distillation

Tae-Hun Kim, Jae-Gul Choo
Dept. of Computer Engineering, Korea University

요 약

딥러닝 기술에 있어서 대량의 학습 데이터가 필요하다는 한계점을 극복하기 위한 시도로서, 적은 데이터 만으로도 좋은 성능을 낼 수 있는 few-shot 학습 모델이 꾸준히 발전하고 있다. 하지만 few-shot 학습 모델의 가장 큰 단점인 적은 데이터로 인한 과적합 문제는 여전히 어려운 숙제로 남아있다. 본 논문에서는 모델 압축에 사용되는 distillation 기법을 활용하여 few-shot 학습 모델의 학습 문제를 개선하고자 한다. 이를 위해 대표적인 few-shot 모델인 Siamese Networks, Prototypical Networks, Matching Networks 에 각각 distillation 을 적용하였다. 본 논문의 실험결과로써 단순히 결과값에 대한 참/거짓 뿐만 아니라, 참/거짓에 대한 신뢰도까지 같이 학습함으로써 few-shot 모델의 학습 문제 개선에 도움이 된다는 것을 실험적으로 증명하였다.

1. 서론

급격한 발전을 거듭하고 있는 기계학습은 알파고의 핵심 모델이었던 강화 학습을 시작으로 이미지 인식과 자연어 처리, 자율 주행 등 다양한 분야에서 두각을 나타내고 있다. 특히 충분한 데이터를 확보한 환경에서의 지도 학습은 인간의 인지 능력을 넘어선 성능까지도 보여주고 있다. 하지만 실제 산업 현장에서는 여러가지 제약들로 인해 충분한 데이터를 확보하지 못한 경우가 대부분이라 기존 모델들이 좋은 성능을 보장하기 어렵다. 이러한 제한된 환경에서의 데이터 부족을 문제를 해결하기 위해 제안된 모델이 few-shot 학습 모델이다.

한정된 데이터만 존재하는 환경에서 좋은 성능을 낼 수 있는 few-shot 학습 모델은 다양한 발전을 거듭하면서 메모리 저장소를 활용 [7]하거나 메타 러닝을 활용 [6]하는 등 발전을 거듭하고 있지만 적은 데이터로 인해 학습 시 발생하는 과적합 등의 문제는 여전히 중요한 숙제로 남아있다. 이를 해결하기 위해 본 논문에서는 정규화 기능을 포함하는 것으로 알려진 self-distillation 기법을 도입하여 이 문제를 개선해 보고자 한다.

Distillation 은 혼합 용액에서 중요한 성분을 추출하는 증류 과정과 같이 복잡한 모델이 모바일과 같은 한정된 자원의 장치에서 구동될 수 있도록 중요한 부분만을 추출하여 최적화된 모델로 만들어주는 기술이다. 이 기술은 교사가 학생을 가르치는 형태와 같이 기존의 모델에서 새로운 모델로 핵심이 되는 내용들을 전달하기 때문에 교사-학생 네트워크 형태를 띄고 있다. 특히나 self-distillation 의 경우 이러한 내용 전달 과정에서 기존 모델의 필수적인 기능은 전달하되 불필요한 편향성은 전달하지 않으려고 함으로써 국소 최저점을 회피할 수 있다고 알려져 있다. 이러한 정규화 기능을 가지고 있는 self-distillation 이 few-shot 학습 모델에도 일반적으로 적용될 수 있는지 알아보기 위해 본 논문에서는 대표적인 few-shot 모델인 Siamese Networks[3], Prototypical Networks[2], Matching Networks[5]에 self-distillation 기법을 적용해 보고자 한다. 이를 통해 self-distillation 기법이 few-shot 모델에서의 과적합 문제를 완화하여 학습 성능을 개선할 수 있는 일반적인 해결책이 될 수 있는지 알아보고자 한다.

2. 관련 연구

2.1 Self-Distillation

Distillation 기법은 여러 가지 형태로 제안되어 왔는데 널리 알려진 distillation 관련 논문은 Hinton et al.의 연구[1]로, 이 논문에서는 dark knowledge 개념을 사용하여 좀 더 효율적으로 distillation 을 할 수 있는 방법을 제안한다. 즉, 일반적으로 클래스 분류 학습에서 사용하는 목표 값인 1(참)과 0(거짓)의 정보 이외에 각각의 클래스들이 얼마나 참과 멀리 떨어져 있는지에 대한 정보를 학습하는 것이 또한 중요하다고 주장하는데, 여기서 후자에 해당하는 정보가 dark knowledge 이다. 교사 모델과 학생 모델 간의 학습에 단순한 1, 0 보다 좀 더 풍부한 정보를 학습하게 함으로써 학습의 정확도를 좀 더 높일 수 있다는 것이다. 이를 학습하기 위해 각 클래스별로 참과 거짓에 대한 신뢰도가 포함되어 있는 확률 분포를 학습하게 하면

서 distillation 학습을 진행하게 된다. 이 논문 이후에 단순히 softmax 의 결과 뿐만 아니라 다양한 레이어의 확률 분포값을 학습 대상으로 선정해서 distillation 을 진행하기도 하고[9] 해당 분포값에 noise 를 추가해서 학습 성능을 향상시키기도 하였다 [10]. 이러한 distillation 을 복잡한 모델을 최적화 시키는 기능에만 사용하는 데서 나아가 모델 자체의 성능 개선에도 사용을 해보고자 하는 Born Again Neural Network [4]의 개념이 제안되었고 여기서 중점적으로 사용된 아이디어가 self-distillation 이다. 자신의 모델을 그대로 본 따 새로운 동일 모델을 만들면서 이전에 학습했던 내용 중에 편향이 심하거나 국소 최저점에 이르는 내용은 회피하고 실제 모델의 역할에 맞는 핵심적인 부분만 전달하자는 것이었다. 이를 통해 self-distillation 이 정규화처럼 동작하게 되고 모델의 성능도 향상시킬 수 있다는 것이다.

2.2 Few-Shot Learning

이미지나 텍스트와 같은 분야에서 기계 학습의 경우 많은 노력으로 양질의 데이터셋이 구축되어 ImageNet, MNIST 등과 같은 데이터를 쉽게 접근할 수 있고 그에 따라 지도 학습의 경우 높은 성능을 내는 모델들이 다수 제안되었다. 하지만 실제 산업 현장에서는 물적, 인적 자원의 한계와 각 산업 현장의 특수성 때문에 이러한 양질의 데이터셋을 확보하기가 현실적으로 힘들다. 이렇게 적은 데이터를 확보한 상황에서도 좋은 성능을 낼 수 있도록 제안된 모델이 few-shot 학습 모델이다. few-shot 모델은 일반적으로 클래스별로 하나의 샘플만을 보여주는 1-shot 학습이나 5 번 보여주는 5-shot 이 일반적[3,6,7]이지만 극단적으로 zero-shot 학습 모델[8]도 제안되고 있다. few-shot 모델은 1 차적으로 학습 데이터들을 통해 해당 도메인의 특징들을 추출하고 각 클래스별로 해당 특징들이 어떤 분포를 가지고 있는지 정의한다. 해당 학습이 끝나면 입력으로 들어오는 테스트 샘플에 대해 가장 유사한 특징 분포를 가지는 클래스로 판단을 하게 된다. 유사한 알고리즘으로 쉬운 예가 최근접 이웃 알고리즘이다 [11]. 이러한 학습의 특성상 특징 분포를 잘 학습할 수 있는 학습 데이터가 많을수록 정확도는 더 향상될 수 있지만 적은 데이터로도 특징을 쉽게 추출할 수 있는 모델을 설계하게 된다. 하지만 적은 데이터로 인한 과적합 문제를 가지고 있어 이 논문에서는 해당 문제의 개선을 위해 self-distillation 을 제안한다.

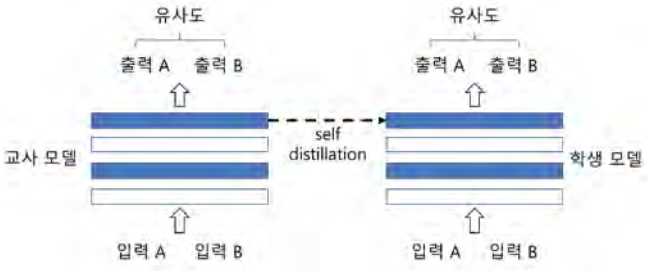
3. Self-Distillation in Few-Shot Learning

Self-Distillation 의 학습은 2 가지로 이루어진다. 첫 번째는 일반적인 사물 분류 모델처럼 라벨을 통해 출력 값의 차이(L_{label})를 판단하고 그 차이값을 역전파를 통해 네트워크에 흘려 줌으로써 모델을 학습시키는 부분이고 두번째는 앞서 이야기한 dark-knowledge 를 활용하기 위해 교사 모델에서 출력되는 클래스별 확률값과 학생 모델에서 출력되는 클래스별 확률값의 차이(L_{dist})를 계산하여 학생 모델의 확률 분포를 교사 모델의 분포에 맞추는 방향으로 학습이 진행되는 부분이다. self-distillation 모델은 위 두 가지 손실을 최소화하는 방향으로 진행하기 위해 두 손실을 더한 값으로 최종 손실값(1)을 가지게 된다.

$$L_{total} = L_{label} + \alpha L_{dist} \quad (1)$$

(1)의 우측의 첫번째 항은 일반적인 분류 모델의 학습과 같이 cross entropy loss 를 사용해서 학습을 진행하고 두번째 항은 두 모델의 분포차를 학습하기 위해 Kullback-Leibler divergence 를 사용한다. 이 중 두번째 항은 [4]에서 언급된 바와 같이 L2 와 같은 정규화 항처럼 작용하면서 편향성을 억제하고 국소 최저점을 회피할 수 있는 역할을 하게 된다. 다만 적용할 모델별로 해당 정규화 작용의 가중치가 조절되어야 할 필요가 있기 때문에 손실 가중치로 하이퍼 파라미터 α 를 사용하였다. (1)을 바탕으로 각 few-shot 학습 모델에서 어떻게 동작하는지 알아보기 위해 대표적인 모델인 Siamese Networks[3], Prototypical Networks[2], Matching Networks[5] 에 적용해 보았다.

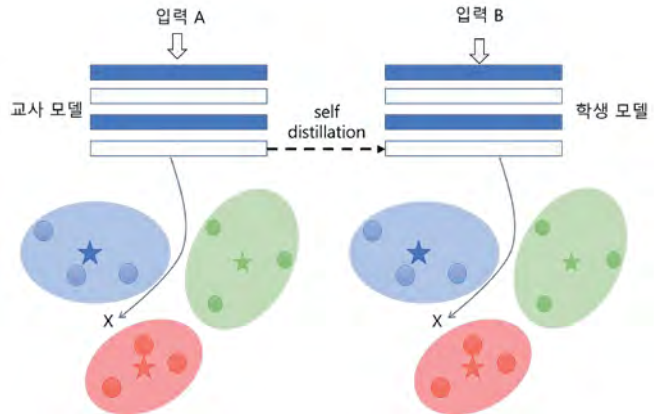
Siamese Networks 분류 대상이 되는 사물을 여러 개의 후보군과 각각 1:1 로 비교 하면서 유사도가 높은 후보군의 클래스로 분류하는 모델이다. 비교 대상이 되는 두 사물의 특징을 동일한 구조의 Convolutional Neural Networks 를 통해 각각 추출하고 해당 특징을 바탕으로 fully connected layer 를 통해 유사도를 측정한다. 해당 모델에 Self-Distillation 을 적용하기 위해 우선 교사 모델을 기존의 손실 함수(L_{label})로 학습시켰고 이후 동일한 구조를 가지는 학생 모델을 새로이 생성하여 기존의 손실 함수(L_{label})와 교사 모델과 학생 모델의 출력값 간의 분포 차이를 나타내는 손실 함수(L_{dist})를 더한 값으로 학습을 진행하였다. L_{dist}의 기여도에 대한 값인 손실 가중치 α값은 실험을 통해 0.1 로 설정하였다.



(그림 1) Siamese Networks + Self Distillation

Matching Networks 분류 후보가 되는 클래스들의 대표 지점을 샘플 데이터를 통해 구성한 뒤 입력되는 사물이 어느 클래스의 대표 지점과 가까운지를 판별하여 분류하는 모델이다. 이를 위해 사전 학습을 통해 각 사물을 가상의 클래스 공간으로 투영시킬 수 있는 투영 신경망을 학습시킨다. 학습된 신경망을 통해 샘플 데이터들을 각각 가상 공간으로 투영시킨 뒤 샘플들의 중심 지점을 해당 클래스를 대표하는 지점으로 간주한다. 이후 클래스 판단이 필요한 사물을

가상 공간으로 투영시킨 위치가 어느 클래스의 대표 지점과 가까운지를 비교하여 가장 가까운 클래스로 분류하게 된다. 이 모델도 이전 실험과 마찬가지로 L_{label}와 L_{dist}를 더한 손실 값으로 진행하였으며 손실 가중치 α값은 0.05 로 설정하였다.



(그림 2) Matching Networks + Self Distillation

Prototypical Networks 기본적인 개념은 Matching Netowkrs 와 유사하지만 Matching Networks 에서 입력 값들의 특징을 추출할 때 bidirectional long short-term memory(Bi-LSTM)을 사용하고 가상 공간에서의 거리를 계산할 때 Consine Distance 를 사용한 반면 Prototypical Networks 에서는 Convolutional Neural Networks 를 통해 특징을 추출하고 Euclidean distance 를 사용하였다. 손실 가중치 α값은 0.05 로 설정하였다.

4. 실험

모든 실험은 일반적인 few-shot 학습 모델의 실험에서 가장 널리 사용되는 Omniglet [12] 데이터셋을 활용하였다. 해당 데이터셋은 50 가지 언어, 1600 여개 문자로 이루어져 있고 각 문자를 20 명이 수기로 적어 데이터를 구성하였다. 이 중 30 개의 언어를 통해 학습을 진행하고 20 개의 언어로 테스트를 진행하였다. 정답 클래스에 대해 한 개의 샘플만 제공하고 후보 클래스 5 종류 중 하나를 선택하는 1-shot 5-way 실험을 진행하였다.

목표 손실값(L_{label})과 분산 손실값(L_{dist})의 학습 반영 비율을 결정하는 손실 가중치 α값은 각 실험마다 모델에 맞게 최적화시켜 진행하였다. 실험의 baseline 들은 재구현 된 결과이고 실험의 공정성을 위해 baseline 과 self-distillation 을 추가한 모든 모델은 α값 이외에 모든 하이퍼 파라메터들(batch size, learning rate, number of epochs)을 같게 설정하여 실험을 진행하였다.

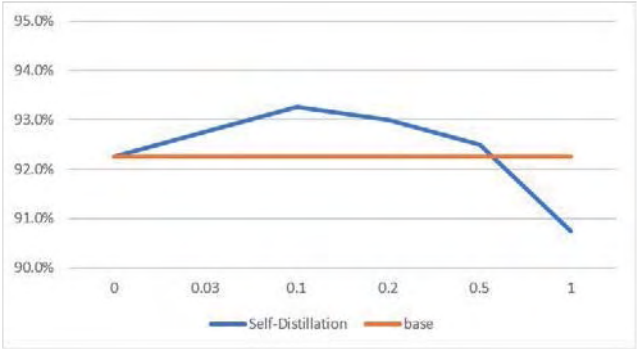
5. 결과 및 분석

Self-distillation 이 적용된 3 개의 모델 모두 <표 1>에서와 같이 각각 1.0%, 0.6%, 0.1%의 분류 정확도가 향상된 것을 볼 수 있었고 모델 변경이 적다는 점을 감안한다면 높은 정확도 향상이라고 볼 수 있다. 그 중 Siamese Networks 의 성능 향상폭이 가장 컸는데 비교 대상인 few-shot 학습모델 중 가장 초기 모델인 만큼 과적합에 대한 모델의 취약성이 가장 컸기 때문에 self-distillation 에 따른 정확도 향상이 가장 큰 것으로 판단된다. Prototypical Networks 의 경우 이미 높은 수준의 분류 정확도를 가지고 있어 그 영향이 적은 것을 볼 수 있었다. 이처럼 대표적인 few-shot 이라고 할 수 있는 3 개 모델에 모두 긍정적인 실험 결과를 보임으로써 self-distillation 이 일반적으로 few-shot 학습 모델에 적용될 수 있는 기술임을 확인하였다.

<표 1> self-distillation 적용에 따른 분류 정확도

	base	base + self distillation	차이
Siamese Networks	92.3%*	93.3%	1.0%
Matching Networks	94.2%	94.8%	0.6%
Prototypical Networks	98.3%*	98.4%	0.1%

추가적으로 실험에서 가장 분류 정확도 향상 폭이 높았던 Siamese Networks 에서 self-distillation 의 역할을 판단하기 위해 손실 가중치 α값을 여러 가지로 변화시켜 최적값을 찾는 실험을 수행하였다.



(그림 3) Hyperparameter 별 판단 정확도

(그림 3)의 주황색 실선이 Siamese Networks 원모델의 분류 정확도 이고 파란색 실선이 해당 모델에 self-distillation 을 적용한 결과이다. 이 그래프에서 볼 수 있는 것처럼 손실 가중치 α값이 0~0.5 사이에 있을 때 분산 손실값은 모두 긍정적으로 Siamese 모델의 정확도에 기여하였다. 하지만 0.5 를 넘어가는 시점에서부터 부정적인 영향을 미쳤다. 또한 기존 목표 손실값(L_{label}) 없이 분산 손실값(L_{dist})만으로 학습을 진행한 별도의 실험에서도 원 모델의 정확도보다 확연히 낮은 결과를 보여주었다. 이런 점들로 미루어 볼 때 분산 손실값이 모델을 학습 시키는 주요 손실값으로는 부족하다고 판단된다. 하지만 우리가 의도했던 부분이 주요 손실값으로써의 기능이 아닌 정규화 항으로써 few-shot 학습의 과적합을 방지하는 기능이라는 점을 고려한다면 의도에 부합하는 역할

을 적절히 수행하고 있음을 보여 주었다.

6. 결론 및 향후 방향

기계 학습이 점점 더 정확도를 높여가고 다방면에 적용됨에 따라 다양한 상황을 접하게 되고 그 중 충분한 데이터를 확보하지 못한 상황에서의 학습이 중요한 문제로 대두되고 있다. 이와 관련하여 few-shot learning 이 좋은 해결책이 될 것으로 기대되고 있지만 이를 위해서는 적은 데이터로 인해 발생하는 과적합과 같은 여러가지 문제들이 해결되어야 한다. 본 연구에서 확인한 것처럼 self-distillation 이 대표적인 few-shot 학습 모델들에 긍정적인 효과를 보여 줌으로써 이러한 문제를 위한 일반적 해결책이 될 수 있음을 확인하였다. 더불어 큰 변경점 없이 적용이 가능하고 다양한 모델에 손쉽게 적용될 수 있다는 장점도 가지고 있다.

정규화 기능으로써의 self-distillation 은 신규 생성 모델이 원래 모델과 구조적으로 일치하다는 특성 덕분에 쉽게 다양한 모델에 적용될 수 있다. 때문에 추후 해당 기능을 메타 학습이나 강화 학습에도 확장해서 적용하고자 한다.

참고문헌

[1] G. Hinton, O. Vinyals, and J. Dean. “Distilling the knowledge in a neural network.”, 2015.

[2] J. Snell, K. Swersky, and R. S. Zemel. “Prototypical networks for few-shot learning.” In Advances in Neural Information Processing Systems, 2017.

[3] G. Koch, R. Zemel, and R. Salakhutdinov. “Siamese neural networks for one-shot image recognition.” In ICML Deep Learning Workshop, 2015.

[4] T. Furlanello, Z. C. Lipton, L. Itti, and A. Anandkumar. “Born again neural networks.” In NIPS Workshop on Meta Learning, 2017.

[5] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, et al. “Matching networks for one shot learning.” In NIPS, 2016.

[6] Ravi, S. and Larochelle, H. “Optimization as a model for few-shot learning.” In the International Conference on Learning Representations (ICLR). 2017.

[7] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap. “One-shot learning with memory-augmented neural networks.” In 33rd International Conference on Machine Learning, 2016.

[8] H. Larochelle, D. Erhan, and Y. Bengio, “Zero-data learning of new tasks,” in AAAI, 2008.

[9] Ba, J. and Caruana, R. “Do deep nets really need to be deep?” In NIPS, pp. 2654–2662. 2014.

[10] Sau, B.B., Balasubramanian, V.N., “Deep model compression: Distilling knowledge from noisy teachers.” 2016.

[11] D. Hand, H. Mannila, P. Smyth., “Principles of Data Mining.” The MIT Press. 2001.

[12] Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. “Human-level concept learning through probabilistic program induction.” Science, 350(6266):1332–1338, 2015.