😀

Distilling the Knowledge in a Neural Network

Abstract

본 논문에서는 강력한 앙상블 모델 혹은 강력한 단일 모델의 예측 분포를 작은 단일 모델로 증류 시키는 새로운 앙상블 기법을 제안합니다. 추가로 종합적인 모델과 각 클래스에 세분화된 전문가 모델을 통한 앙상블 기법 또한 제안합니다.

Introduction

본 논문에서는 다양한 모델들이 서로 다른 Task를 수행하더라도 대부분 동일한 모델구조를 가지고 있다고 주장합니다. 그래서 본 논문은 특정 앙상블 모델 혹은 강력한 모델 처럼 특징 추출의 정확도를 높히는데 중점을 둔 모델을 기반으로 베포가 용이한 작은 모델로의 지식 증류를 통해서 베포의 용이성을 제안합니다.

하지만 일반적으로 지식을 증류한다는 개념은 파라미터 관점에서 학습된 가중치를 매핑 시킨다는 관점으로 이해하기에 모델의 구조가 바뀐다면 지식을 전달하는데 번거러움이 등장합니다. 하지만 본 논문에서는 모델의 구조나 파라미터가 바뀌어도 (입력 → 예측 출력) 의 함수적 관계 즉, 예측을 만들어내는 방식이 유지될 수 있다면 같은 지식을 보유하고 있다고 주장할 수 있다고 합니다.

또한 잘못된 오차확률 분포를 통해서도 모델이 어떻게 일반화를 하는지에 대한 정보를 얻을 수 있다고 주장합니다. 예를 들어 BMW 이미지를 쓰레기 트럭으로 예측할 확률이 0.003, 당근이라고 예측할 확률이 0.000001 이라면 쓰레기 트럭이 당근보다는 BMW와 유사한 특징을 보유하고 있다는 정보를 얻을 수 있습니다. 지식 증류는 이러한 오답 확률 분포가 작은 모델이 큰 모델의 미묘한 지식을 학습하는데 굉장히 중요하다고 주장합니다.

즉, 본 논문에서는 지식이란 파라미터의 가중치 그 자체가 아니라, 입력에 대한 출력 확률 분포가 유사하다면 유사한 지식을 가지고 있다고 판단하였습니다.

모델을 학습하는 경우 주로 일반화 성능을 향상 시키는 방향으로 학습시키게 됩니다. 하지만 처음부터 일반화 성능이 높은 학습 방법을 알지 못하기에 작은 모델은 단순히 거대 모델의 방식을 따로도록 합니다.

큰 모델로 작은 모델을 학습하는 가장 명백한 방법은 큰 모델이 만들어내는 클래스 확률을 작은 모델 학습의 'soft target'으로 사용하는 것입니다. 이에 거대 모델의 출력 값인 soft target은 entropy가 높을 수록 거대 모델의 디테일한 부분까지 학습할 수 있도록 유도합니다. ([0,1,0] 보다는 [0.2, 0.5, 0.3] 이 일반화에 도움이 된다)

본 논문에서는 MNIST의 거대 모델에 대해서는 정확도가 매우 높아 출력 확률 분포를 통해서 유의미한 결과를 얻기 어렵다고 하였습니다. 이에 Caruana 등의 접근법은 softmax를 넣기 전 logit 값을 사용하여 보다 일반화할 수 있다고 주장합니다. 그리고 본문에서는 Temperature를 통해서 softmax를 더욱 부드럽게 타켓을 설정하도록 하였다고 합니다.

그리고 본논문에서는 작은 모델을 비지도 학습과 지도 학습으로 진행했을때, 그래도 지도 학습을 통한 편향을 학습하는 것이 조금더 성능이 좋았다고 주장합니다.

Introudction 정리

거대 모델이 학습한 확률 분포를 모방함으로써, 작은 모델도 그와 유사한 지식을 갖추게 된다. 즉, 작은 모델이 거대 모델의 출력 확률 분포를 그대로 학습한다는 것은 곧 거대 모델의 지식을 증류받는 과정이라 할 수 있다.

Distillation

본 논문에서는 T라는 값을 사용하여 일반적은 Neural Network의 에측 확률 분포를 부드럽게 수정해줍니다.

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

T = 1 이 기본 softmax이며, T 가 1보다 크면 클수록 더욱 부드러운 예측 확률 분포를 얻을 수 있습니다. 그래서 거대 모델의 에측 확률 분포를 얻는 경우 높은 T를 사용하고, 작은 모델 학습 시에도 높은 T를 사용하여 학습합니다. 그리고 Inference시 T = 1 로 설정하여 예측한다고 주장합니다.

그리고 transfer set이 라벨이 없어도 되지만 라벨이 있는 경우 높은 성능을 발휘한다고 합니다. 만일 라벨이 존재하는 경우 2가지 목적함수를 가중합 하여 사용하게 되는데,

(1) 높은 T를 사용하여 거대 모델의 soft target 간의 cross entropy 계산

(2) 라벨과의 cross entropy 계산

$$\alpha(1) + (1-\alpha)(2)$$

다음과 같은 가중합을 통해서 손실을 계산하게 됩니다. 그리고 주로 α값이 작은 경우 높은 성능을 달성한다고 주장합니다.

그리고 Backpropagation시 (1) 방법을 사용하게 되면 gradient가 $1/T^2$ 만큼 커지기에 gradient 계산시 (2)번 과의 일관성을 맞춰주기 위해서 T^2 을 곱해줘야 한다고 합니다.

Matching logit is a special case of distillation

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i) = \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{T}\left(\frac{1+\frac{z_i}{T}}{N+\sum_j \frac{z_j}{T}} - \frac{1+\frac{v_i}{T}}{N+\sum_j \frac{v_j}{T}}\right)$$

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{N\,T^2}\left(z_i - v_i\right)$$

< 위의 수식 풀이 >

$$\frac{\partial C}{\partial z_i} = \frac{\partial\left(-\sum_k p_k \log q_k\right)}{\partial z_i}$$ // p_k는 z_i와 무관.

$$= -\sum_{k=1}^{n} p_k \cdot \frac{1}{q_k} \cdot \frac{\partial q_k}{\partial z_i}$$ // $q_k = \frac{e^{z_j/T}}{\sum_j e^{z_j/T}}$

↳ $\begin{cases} q_k(1-q_k)\frac{1}{T} \\ -q_k q_i \cdot \frac{1}{T} \end{cases}$

$$= \frac{1}{T}(q_i - p_i)$$ // $q_i = \frac{e^{z_j/T}}{\sum_j e^{z_j/T}}$, $p_i = \frac{e^{v_j/T}}{\sum_j e^{v_j/T}}$

S.T "테일러 급수 적용 (T↑ 인 경우) 1차 근사 가능 ⇒ $e^x = 1+x$

$$= \frac{1}{T}\left(\frac{1+z_j/T}{N+\sum_j z_j/T} - \frac{1+v_j/T}{N+\sum_j v_j/T}\right)$$

S.T $\sum_j z_j = \sum_j v_j = 0$으로 가정한다.

$$= \frac{1}{T}\left(\frac{T+z_i}{N} - \frac{T+v_i}{N}\right) = \frac{1}{T}\left(\frac{(z_i - v_i)}{NT}\right) = \frac{1}{NT^2}(z_i - v_i)$$

T를 높게 설정할 경우

1차 테일러 근사 적용를 적용하여 T가 충분히 크면, $\exp\left(\frac{z_i}{T}\right)$ 는 1차 테일러 근사 $1 + \frac{z_i}{T}$ 로 근사할 수 있습니다. (테일러 근사는 $e^x = 1 + x + x^2/2! + \ldots + x^p/p!$ 로 정의되

고, 이때 T가 충분히 크게되면 z_i/T 값이 작아져 2차항 부터는 0에 가까운 수가 되기에 단순히 근사가 가능합니다.)

각 전이 케이스마다 로그릿 벡터 $\{z_i\}$ 가 제로 평균 ($\sum_j z_j = 0$)으로 정규화되면, 최종적으로 크로스 엔트로피 기울기가 $\frac{1}{T}(q_i - p_i) \approx \frac{1}{NT^2}(z_i - v_i)$ 로 근사되어, 결과적으로 Student 모델이 Teacher 모델의 로짓과의 차이 $\frac{1}{2}(z_i - v_i)^2$ 를 최소화하는 문제로 전환됩니다.

장점:

모든 클래스에 대한 확률이 평평하게 나와, 각 클래스 간의 미세한 관계(유사성)를 Student 모델이 학습하는 데 유리합니다.

단점:

실제로 중요한 큰 차이를 갖는 로짓 값 간의 차이가 감소되어, Teacher 모델이 갖고 있는 강한 확신(confidence) 또는 분류 경계 등 중요한 정보가 약화될 수 있습니다.

T를 낮게 설정할 경우

뾰족한 확률 분포를 갖습니다. 낮은 T에서는 소프트맥스 출력이 거의 원-핫 형태에 가깝게 나오므로, 정답 클래스의 확률이 매우 높고 나머지는 거의 0에 가까워집니다.

장점:

정답 분포(하드 타깃)에 집중하게 되어 노이즈를 제거하는 효과를 얻을 수 있습니다.

단점:

Teacher 모델의 아주 작은 값의 로짓이라도 중요한 클래스 간 관계 정보가 포함되어 있을 수 있는데, 낮은 T에서는 이런 미세한 관계가 반영되지 않으므로, Student 모델이 Teacher 모델의 일반화 패턴을 온전히 학습하지 못할 위험이 있습니다.

중간 단계의 T 설정 권장.

본 논문에서는 만약 Student 모델이 Teacher 모델의 모든 복잡한 패턴을 학습하기엔 용량이 부족하다면, 너무 높지도 낮지도 않은 중간 온도가 최적의 성능을 보입니다. 이는 중간 온도가 Teacher 모델의 미세한 유사성 정보를 일정 부분 전달하면서도, 너무 평탄해지는 문제를 피할 수 있기 때문입니다.

Preliminary experiments on MNIST

본 논문에서는 MNIST에 대해서 실험을 진행하였고, 1200개의 레이어를 갖는 강력한 모델 학습시 테스트시 67개의 오류를 기록했다고 합니다. 그리고 800개의 레이어를 갖는 모델의 경우 146개의 오류를 보였고, 이를 거대 모델의 soft target을 사용하여T = 20으로 학습한 경우 틀린 갯수가 74개로 줄었다고 주장합니다. 그리고 극단적으로 층이 30개 정도인 모델은 T를 2.5 ~ 4 범위를 사용하는게 효과가 더욱 좋았다고 합니다.

이후 본논문에서는 transfer set에서 3에 대한 데이터를 지우고 실험 한 결과 거대 모델에서 받은 soft target으로 충분히 3에 대한 지식을 학습할 수 있었음을 시사합니다. 단, 오류율이 높았지만 bias를 적당히 조정하게 되면 높은 성능을 달성한다고 주장합니다. 이에 지식 증류에 있어, 작은 모델의 용량, T의 크기 , bias 조정이 중요한 역할을 한다는 것을 시사합니다.

Result

실험 결과 단일 Hard taget이 제공할 수 없는 풍부한 정보가 soft target에 담겨 있어 보다 일반화된 패턴을 학습할 수 있다고 주장합니다. 특히 매우 큰 네트워크의 경우 전체 앙상블 학습이 어려울때, 다수의 전문가 모델을 독립적으로 학습 시켜 지식을 전달하는 전략이 효과적이라고 주장하며, 최종적으로 이러한 전문가모델의 지식을 다시 단일 모델로 증류 시키는 연구가 진행중에 있다고 합니다.

한줄 정리

본 논문에서는 지식을 파라미터의 가중치가 아닌 입력에 따른 출력의 확률 분포로 정의함. 이에 높은 연산량으로 학습된 모델의 출력 확률 분포를 모방하는 작은 모델을 학습 시키면 지식이 증류된다고 주장함. 이에 실험을 통해서 Hard label 학습 + soft lable target 으로 학습하는 경우 더 높은 성능을 발휘함을 입증함.

이에 강력한 모델로 부터 베포에 유리하도록 작은 모델로 지식을 증류 할 수 있음을 주장하였습니다.

나의 생각

지식을 전달하는 개념을 출력의 확률 분포로 정의한것은 매우 좋은 접근 인 것 같다. 하지만 뭔가 모델의 내부적인 구조나 진행하는 Task에 따라서 적합한 Teacher - student 모델의 쌍이 존재할것 같다고 생각이 된다.

추가로 최종 확률 분포 이외에도 중간 확률 분포 또한 적용하면 더 높은 성능을 발휘 할수 있지 않을까? 라는 생각이 든다.