[week12] 논문리뷰

Distilling the Knowledge in a Neural Network

1. Introduction

논문에서 다루고 있는 주제가 무엇인지와 해당 주제의 필요성이 무엇인가
논문에서 제안하는 방법이 기존 방법의 문제점에 대응되도록 제안 되었는가

논문이 다루는 분야

- model compression
- knowledge transfer
- Knowledge Distillation
- 주로 적용한 task = supervised 분류 문제 전반
  - MNIST 이미지 분류
  - 대규모 음성 인식용 음향 모델
  - 초대형 이미지 데이터셋(JFT) 분류

해당 task에서 기존 연구 한계점

1. 앙상블 / 대형 모델의 실용성 한계
    - 앙상블이나 매우 큰 네트워크는 학습 시에는 좋지만, 파라미터 수가 많고 메모리/연산량이 크고 응답 시간이 느려 실제 서비스(모바일, 온라인 서빙 등)에 쓰기에는 너무 비싸다.
2. 기존 model compression(=logit matching) 방식의 제한
    - 앙상블의 logit이나 예측값을 그대로 맞추는 식의 모델 압축은 존재했지만, “왜 잘 되는지”, “어떤 형태의 정보가 중요한지”를 체계적으로 해석할 수 있는 일반적인 프레임워크가 부족했다
    - 특히 softmax의 온도(temperature)를 조절하면서 어떤 온도에서 어떤 정보(클래스 간 유사도 구조)가 가장 잘 드러나는지 logit matching이 어떤 경우에 자연스럽게 나오는지를 이론적으로 연결해 설명해주지 못했다
3. 작은 모델을 직접 hard label로만 학습할 때의 한계
    - 작은 네트워크를 같은 데이터, 같은 hard label만으로 학습하면 대형 모델/앙상블이 학습한 “오답들 사이의 미묘한 확률 구조”를 사용하지 못한다.
4. Mixture of Experts / specialist 구조의 실용성 문제
    - Mixture of Experts처럼 여러 expert 모델을 두고 gating으로 라우팅하는 방식은 powerful 하지만 대규모 병렬화, 구현이 어렵고 매우 큰 데이터셋(JFT)을 다루기엔 **시스템적으로 복잡**하다.
5. 기존 speech distillation 연구의 한계
    - 온도 $T=1$T = 1T=1, unlabeled data 사용, teacher 분포를 단순 모사하는 형태로, 이 논문에서 제안한 방법에 비해 성능 격차를 충분히 줄이지 못한다.

논문의 contributions

1. 온도 기반 Softmax를 이용한 일반적인 지식 증류 프레임워크 제안
    - teacher와 student 모두 softmax with temperature T를 사용하고, 높은 T에서 teacher가 내는 soft한 확률 분포(soft targets)를 student가 그대로 모방하도록 학습하는 지식 증류(distillation) 방법을 제안
    - 실제 추론 시에는 student가 T=1을 사용하므로, 추론은 일반적인 softmax처럼 빠르면서 학습 과정에서는 teacher의 풍부한 구조 정보를 활용한다.
2. Soft target + Hard label을 함께 사용하는 학습 목적함수 설계
    - 라벨이 있는 경우, 손실 함수를

    - (1) 높은 T에서의 soft target에 대한 cross-entropy
    - (2) T=1에서의 hard label에 대한 cross-entropy

      의 가중합으로 두는 방식을 제안.
  - soft target의 gradient가 1/T^2에 비례하는 점을 분석하고, T^2를 곱해 두 손실 기여도를 안정적으로 조절하는 트릭까지 제안.
3. Logit matching이 "고온도 증류"의 특수한 경우임을 이론적으로 증명
  - teacher logit v_i, student logit z_i일 때, 높은 온도 T에서의 증류 loss의 gradient를 전개하면 $\propto (z_i - v_i)$꼴이 되어 제곱 오차로 logit matching하는 것과 동일해짐을 보여줌.
  - Caruana식 logit matching = distillation의 한계 케이스라는 해석을 제공
    → 기존 model compression 방법을 더 큰 이론 틀 안에 포함시킴.
4. MNIST, 음성 인식, JFT 등 다양한 규모의 실험으로 distillation 효과 입증
5. Specialist 네트워크 구성 및 결합 방법 제안
  - generalist의 예측값 공분산을 클러스터링하여 클래스 그룹을 만들고, 각 그룹에 대해 specialist를 학습하는 방법을 제안.
  - 추론 단계에서 generalist + 관련 specialist들의 출력 분포와의 KL divergence 합을 최소화하는 일관된 최종 분포를 gradient descent로 구하는 방법을 제안.
  - Mixture of Experts와 달리 specialist들을 완전히 독립적으로, 병렬로 학습할 수 있고, 이후 하나의 통합 분포로 결합하는 일반적 방법을 제공

2. 제안 방법론

Introduction에서 언급된 내용과 동일하게 작성되어 있는가
Introduction에서 언급한 제안 방법이 가지는 장점에 대한 근거가 있는가
제안 방법에 대한 설명이 구현 가능하도록 작성되어 있는가

3.1 **Core Distillation Framework**

- teacher와 student 모두 온도 T를 가진 softmax 출력을 사용한다

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

T가 커질수록 분포가 부드러워져(soft) 클래스들 사이의 상대적 구조 정보가 더 많이 드러난다.

- Soft target
  - transfer set의 각 샘플에 대해, teacher가 높은 온도 T에서 softmax 확률 분포를 출력하면, student는 같은 T로 이 soft 확률분포를 모방하도록 학습된다. 학습이 끝나면 student는 T=1로 추론한다.
- Soft + Hard label 혼합 학습:
  - 목적함수 =
    1. 높은 T에서 계산한 soft target에 대한 cross-entropy
    2. T=1에서 계산한 hard label에 대한 cross-entropy의 가중합
  - soft target으로부터 오는 gradient는 크기가 1/T^2에 비례하므로, 온도가 바뀌어도 두 손실의 비중이 크게 변하지 않도록 soft 손실 쪽 gradient에 T^2를 곱해준다.

3.2 **Logit Matching as a Special Case**

- teacher logit v_i, student logit z_i라 할 때, 온도 T에서의 증류 손실은 teacher soft target p_i와 student 분포 q_i 사이의 cross-entropy이고, 그 gradient는 다음과 같이 주어진다.

$$\frac{\partial C}{\partial z_i} = \frac{1}{T}(q_i - p_i)$$

- **고온도 한계**에서 (모든 logit이 T에 비해 작고, 각 샘플마다 logit의 평균을 0으로 맞춘 경우) 이 gradient는 다음과 같이 근사된다.

$$\frac{\partial C}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

- 이는 (zi−vi)^2를 최소화하는 것과 같다. 즉, logit matching은 고온도 증류의 특수한 경우임을 보인다.
- 중간 수준의 T에서는 매우 음수인 logit들의 영향이 줄어들어, 학습 시 거의 제약받지 않은 노이즈성 logit에 과도하게 맞추지 않으면서도 중요한 클래스들 간의 구조는 유지할 수 있다.

3.3 **Specialist Ensemble Construction**

- JFT처럼 (100M 이미지, 15,000 클래스) 매우 큰 데이터셋에서 많은 풀 모델을 앙상블로 학습하는 것은 계산량이 과도하다. 대신, 전체 클래스를 다루는 generalist 모델 1개와 혼동되는 일부 클래스들에만 집중하는 여러 개의 specialist 모델로 구성된 앙상블을 제안한다.
- 각 specialist는
  - generalist의 가중치를 초기값으로 사용해 하위 표현을 공유하고,
  - 자신의 특화 클래스들 + 나머지를 합친 dustbin 클래스로 구성된 작은 softmax를 가지며,
  - 학습 시 절반은 특화 클래스, 절반은 나머지 데이터로 학습한 뒤,
  - oversampling 효과를 보정하기 위해 dustbin logit에 상수 보정값을 더한다.
- specialist에 할당할 클래스 그룹은, generalist가 산출한 예측값들의 공분산 행렬의 열들을 클러스터링하여 얻는다. 이는 정답 라벨 없이도, 함께 자주 예측되는 클래스들을 모아서 specialist의 타깃 집합으로 사용할 수 있게 한다.
- 추론 시에는
  1. generalist가 top-n 클래스를 선택하고,
  2. 이들 중 하나 이상을 포함하는 specialist들을 active set으로 모은 뒤,
  3. generalist 분포 p_g와 active specialist 분포들 p_m과의 KL divergence 합을 최소화하는 전체 클래스 분포 q를, softmax logit에 대해 gradient descent로 찾는다.

3. Related Work

💡 Introduction에서 언급한 기존 연구들에 대해 어떻게 서술하는가
제안 방법의 차별성을 어떻게 표현하고 있는가

- **앙상블 기반 모델 압축**: Buciluǎ–Caruana–Niculescu-Mizil은 **앙상블의 예측값**을 정답으로 삼아 작은 모델을 학습함으로써 앙상블의 지식을 단일 모델로 압축할 수 있음을 보였다. 이 논문은 여기에 **softmax 온도 조절**을 도입해 보다 일반적인 "증류" 방법을 제시하고, logit 매칭이 그 특수한 경우임을 수식으로 보여준다.
- **드롭아웃 = 암묵적 앙상블**: 드롭아웃은 파라미터를 공유하는 매우 많은 얇은 네트워크들의 앙상블로 해석될 수 있고, 이들 전체를 대표하는 하나의 큰 네트워크를 teacher로 사용할 수 있다.

- **대형 음성 모델 → 소형 모델**: Li 등(2014)은 큰 음성 인식 모델의 출력 분포를 작은 모델이 따라가도록 학습하는 유사한 아이디어를 사용하지만, **온도 1**, **대규모 비라벨 데이터**에 의존하며, 성능 격차를 줄이는 정도가 이 논문의 방법보다 작다고 보고된다.
- **Mixtures of Experts와의 관계**: 전통적인 Mixture of Experts는 gating 네트워크가 각 예제를 어떤 expert에게 보낼지 학습하는 구조이지만, 이 방식은 대규모 병렬화가 어렵다. 반면 이 논문의 "specialist" 네트워크들은 **generalist 모델을 먼저 학습한 뒤**, 그 혼동 구조를 기반으로 **완전히 독립적으로 병렬 학습**될 수 있다.

4. 실험 및 결과

💡 Introduction에서 언급한 제안 방법의 장점을 검증하기 위한 실험이 있는가

4.1 **MNIST**

- teacher: 은닉층 1200 ReLU 두 층, 드롭아웃과 가중치 제약, 입력 이미지의 ±2픽셀 jittering을 사용한 큰 네트워크로, MNIST에서 **테스트 오류 67개**를 달성한다.
- 작은 baseline: 은닉층 800 ReLU 두 층, 정규화 없이 hard label만으로 학습하면 **테스트 오류 146개**.
- 같은 구조의 student를 teacher의 **온도 T=20**soft target(및 일부 hard label)으로 증류 학습하면 **테스트 오류 74개**로, 작은 baseline보다 훨씬 teacher에 가까운 성능을 얻는다.
- 온도 영향: 층당 300개 이상 유닛이면 T≥8이면 비슷한 결과지만, 층당 30개처럼 매우 작은 네트워크에서는 T≈2.5–4의 **중간 온도**가 가장 좋았다. 이는 너무 큰 T에서 매우 음수인 logit 정보를 거의 버리는 것이, 용량이 작은 모델에는 오히려 손해일 수 있음을 시사한다.
- **클래스를 아예 안 보여준 실험(3 제외)**:
  - transfer set에서 3을 완전히 제거해도 증류된 모델은 처음에 **206개** 오류(그 중 133개가 3),
  - 3 클래스의 bias를 +3.5만큼 올리면 전체 오류 **109개**, 이 중 3에 대한 오류는 14개 → **학습 중 3을 본 적이 없음에도 3에 대해 98.6% 정확도**를 보인다.
  - transfer set이 7과 8만 포함해도 bias 조정만으로 오류율을 크게 낮출 수 있음을 보인다.

4.2 **Speech Recognition (Acoustic Models)**

- 과제: 대규모 음성 인식을 위한 DNN 기반 **음향 모델** (Android voice search와 유사).
  - 구조: 은닉층 8개, 층당 2560 ReLU, 출력층 softmax 14,000개 HMM 상태.
  - 입력: 40차 Mel filterbank 특징 26프레임(10ms 간격)을 넣고, 21번째 프레임의 상태를 예측.
  - 파라미터 약 8,500만 개, 2000시간(7억 프레임) 음성 데이터 → baseline frame accuracy **58.9%**, WER **10.9%**.
- 동일 구조의 모델을 초기값만 다르게 해서 **10개 학습**한 앙상블은 frame accuracy **61.1%**, WER 10.7%로 성능이 향상된다.
- 이 10개 모델 앙상블의 soft target으로 동일 구조의 single 모델을 증류하면, frame accuracy **60.8%**, WER 10.7%를 얻어, **앙상블이 baseline 대비 얻은 이득의 80% 이상을 단일 모델로 회수**한다

4.3 **Soft Targets as Regularizers (Data Efficiency)**

- 전체 데이터의 **3%만 사용**해 acoustic 모델을 학습하는 실험:
  - hard target만 사용: train 67.3%, test 44.5%로 심각한 overfitting(early stopping 필요).
  - soft target 사용(teacher는 전체 데이터로 학습): train 65.4%, test 57.0%로, 전체 데이터(58.9%)와 거의 비슷한 성능을 달성한다.
- 이는 soft target이 전체 데이터에서 학습된 **규칙성과 구조를 강하게 내포**하고 있어, 일부 데이터만 가지고도 일반화를 잘 하게 만드는 **강력한 정규화 효과**를 보여준다.

4.4 **JFT Specialists**

- 데이터셋: Google 내부 JFT (이미지 1억 장, 라벨 1.5만 개). baseline은 수개월 동안 대규모 분산 학습을 거친 깊은 CNN이다.
- baseline generalist를 초기값으로, 약 300개 클래스 + dustbin 클래스를 갖는 **specialist 61개**를 독립적으로 병렬 학습한다.
- 결과(Top-1 정확도):
  - baseline: 조건부 정확도 43.1%, 전체 정확도 25.0%.
  - specialist 61개 추가 시: 조건부 45.9%, 전체 26.1%로, **전체 정확도 기준 약 4.4% 상대 향상**.
  - 하나의 클래스에 대해 specialist가 많이 겹칠수록 상대적인 향상이 커지며, 여러 specialist의 중첩이 유리함을 보여준다.

5. 결론 (배운점)

💡 연구의 의의 및 한계점, 본인이 생각하는 좋았던/아쉬웠던 점 (배운점)

- 지식 증류(distillation)는 큰 앙상블이나 강하게 정규화된 teacher 모델의 지식을 작은 student 모델로 옮겨, 단일 모델의 효율성을 유지하면서도 앙상블의 성능 이득 대부분을 회수할 수 있는 방법임을 보인다.
- MNIST에서는 transfer set에 특정 숫자가 아예 없어도, soft target으로 학습한 뒤 bias만 조정하면 그 숫자에 대해 매우 높은 정확도를 얻을 수 있다. 이는 soft target이 클래스 간 유사도 구조를 풍부하게 담고 있음을 보여준다.
- 음성 인식에서는 baseline과 같은 크기의 단일 모델이 10개 앙상블이 제공하는 정확도 향상의 대부분을 증류를 통해 가져올 수 있고, soft target은 3% 데이터만 사용해도 좋은 일반화를 만드는 강력한 정규화 수단이 된다.
- 초대형 비전 과제(JFT)에서는 enormous generalist 네트워크 위에 specialist ensemble을 추가해 성능을 끌어올릴 수 있으며, 각 specialist는 빠르고 독립적으로 병렬 학습할 수 있다. 향후 과제로, 이 specialist들의 지식을 다시 단일 모델로 역증류(distill back)하는 방향이 제안된다.