**SMS: Stochastic Mask Selection for MaskedKD**

**1. 서론**

Knowledge distillation(KD)은 모델의 경량화에 유용한 방법론이나, 학습과정에서 teacher의 훈련 데이터셋에 대한 추론을 위해 추가적인 비용을 필요로 합니다. MaskedKD[1]는 transformer 기반 모델 간 student의 attension score를 활용해서 비용과 성능 측면에서 보다 효율적인 KD를 수행할 수 있는 mask selection mechanism을 제안했습니다. 이번 연구계획서에서는 stochastic mask selection(SMS)을 제안하고, 그에 대한 가정과 필요성, 효과에 대한 간단한 실험과 결과를 담았습니다. SMS는 deterministic하지 않은 soft-selection으로 student 모델이 특정 patch에 대한 학습에만 집중하여 발생할 수 있는 overfitting을 줄이고, 특히 attention score 간 편차가 크지 않은 초기 학습단계나 특정 데이터셋의 학습환경에서 local minima 문제를 벗어날 수 있을 것이라고 기대됩니다.

**2. 관련연구**

**Distillting ViTs.** DeiT[2]는 distillation token을 teacher 모델으로 학습시키는 것으로 teacher의 inductive bias를 전달하는 방법을 제안했습니다. 이는 기존에 ViT에서 inductive bias에 취약한 점을 보완하기 위해 대규모 데이터셋을 활용한 것과 다른 방식이며, KD가 단순히 모델의 경량화 측면 뿐만 아니라 성능 향상에도 도움을 준다는 것을 보여주었습니다.

**Mask Selection.** MaskedKD[1]는 transformer 기반 모델 간 distillting 학습에 필요한 teacher 모델 추론에서 비용은 절감하고 학습 성능은 올릴 수 있는 mask selection을 제안했습니다. Student 모델이 추론에 사용한 attention score를 mask selection에서 그대로 사용하기 때문에 추가적인 비용 없이, teacher는 주요하지 않은 부분이 mask된 patches로 추론을 하기 때문에, 예측에 소요되는 비용과 보다 정확한 예측 결과를 얻을 수 있었습니다.

**3. 제안방법**

**3.1. Review of MaskedKD**

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Distillting 학습은 student가 true label을 학습할 수 있도록 하는 크로스 엔트로피 손실 와 student의 distillation token과 teacher의 classification token의 출력분포가 닮을 수 있도록 학습하는 KLD 손실 을 더한 최종 손실(1)을 최소화하는 것을 목표로 합니다. 이때 student의 추론으로 마지막 multi head attention

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

레이어의 attention score(2)를 더한 으로 top-k selection에 따라 가장 낮은 순서로 n-patches를 masking 합니다. Attention score를 기반으로 하기 선택하기 때문에 추론에 꼭 필요한 patch가 mask되지 않고 teacher는 더 적지만 핵심 patch들로 추론을 하기 때문에 추론 속도와 학습 성능에서 우위를 보여줍니다.

3.2. Stochastic Mask Selection

Attention score를 top-k 알고리즘으로 선택하게 되면 학습이 진행될수록 student가 학습한 편향에 따라 동일한 patch들만 mask되게 되어서 overfitting 될 수 있다고 생각했습니다. 또한, head의 attention의 총합을 사용하여 hard-selection할 경우 일반적으로 일부 patch가 대부분의 정보를 담고 있는 long-tail 한 상황이 아닌, multi-object detection과 같은 head 간 분포 차이가 클 것으로 예상되는 학습환경이라면 손실될 수 있는 정보가 더 많아질 수 있다고 가정했습니다.

그래서 soft-selection으로

**4. 결론**

이번 연구계획서에서는 data shortage와 마찬가지로 현실에서 발생할 수 있을 imbalanced data issue 상황에서 regression을 부분적으로 classification으로 변형해서 문제를 해결해갈 수 있을지에 대한 가정을 담았습니다. 기존에 연구하신 TSM methods와 충돌하는 부분이 없다면 접목해서 shortage & imbalanced 상황에서 어떤 결과가 나올지도 연구해보고 싶습니다. “Cost-Effective Learning and Inference”이나 다른 data driven learning 등 methodological research에 대한 관심과 이를 응용하는 분야에 관심이 있습니다.

**References**

[1]Son, S., Lee, N., and Lee, J., “MaskedKD: Efficient Distillation of Vision Transformers with Masked Images”, 2023.

[2]Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H., “Training data-efficient image transformers & distillation through attention”, 2020.