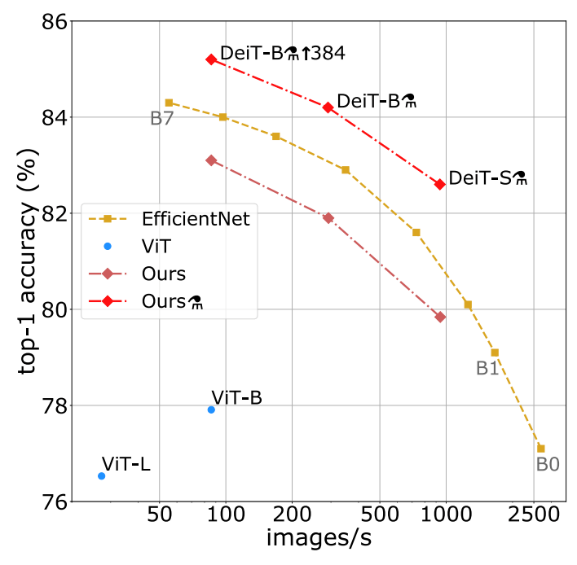
안녕하세요 segmentation의 마지막 주제인 Deit에 대한 발제를 맡게 된 16기 분석 이지혜입니다.

Introduction

이미지 분류와 같이 이미지를 이해하는 작업을 수행하는 데에 Vision Transformer들은 높은 성과를 보입니다. 그러나 이들은 큰 infrastructure, 즉 많은 양의 학습 데이터가 주어지는 환경에서 학습되어야 한다는 한계를 갖고 있다는 점을 시사하면서 논문을 시작합니다. 저자들은 Imagenet으로만 학습을 시킨, CNN을 이용하지 않은 기존의 VIT모델에 distillation을 도입하여 경쟁력 있는 transformer 기반의 모델인 DeiT를 소개합니다. 해당 모델은 SOTA CNN기반의 모델과 비슷한 성능을 보인다고 합니다. 이미지 영역에 있어서 transformer 기반의 모델은 이제 막 연구가 시작 되었는데, 비슷한 성능을 보인다는 점에서 해당 논문은 transformer의 가능성을 열어준다고 볼 수 있습니다.

왼쪽의 사진은 모든 모델들을 ImageNet으로만 학습시킨 이후에 EfficentNet, ViT, DeiT의 세가지 모델의 성능을 비교한 결과이고, DeiT의 성능이 제일 좋음을 확인할 수 있습니다.

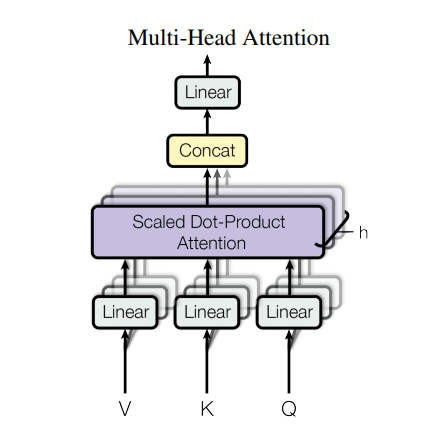
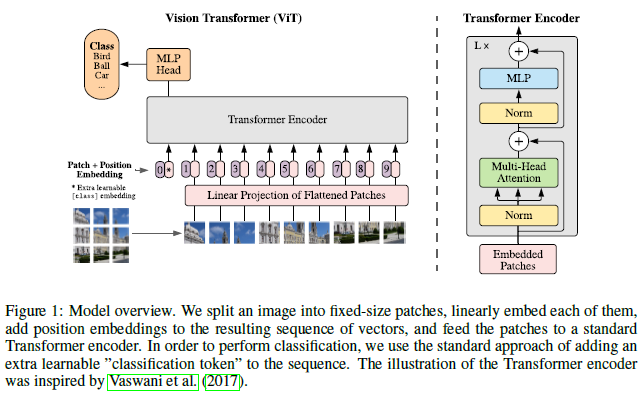
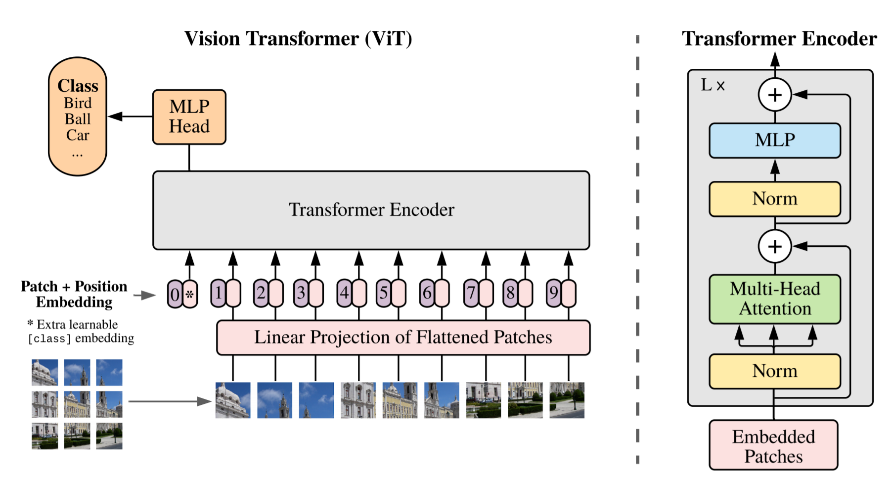
간단하게 transformer와 cnn의 차이점에 대해서, 그리고 자연어 처리에 사용이 되던 transformer이 이미지 철리에도 쓰일 수 있게 되는 이유에 대해 살펴보겠습니다. 위의 그림에서도 알 수 있듯이 CNN의 경우에는 이미지의 멀리 떨어진 픽셀 사이의 정보들을 하나로 통합하기 위해서 여러 개의 layer을 통과해야 하는 반면, transformer의 경우에는 하나의 layer만으로도 정보의 통합이 가능하게 됩니다. 즉, transformer은 멀리 떨어진 픽셀 사이의 상관관계의 분석을 위해서 1차원의 벡터로 바꾸어 주어서 value, query, key라는 값들을 이용해서 self-attention을 적용해서 이미지의 정보를 얻어낼 수 있는 것입니다.

이 시점에서 inductive-bias라는 개념에 대해서 짚고 넘어가고자 합니다. 이는 새로운 데이터에 대해서 좋은 성능을 내기 위해 모델에 사전적으로 주어지는 가정을 의미합니다. CNN의 경우 지역적인 정보를, RNN의 경우에는 순차적인 정보가 각각의 inductive bias가 될 것입니다. Transformer의 경우에는 1차원 벡터로 만든 후에 self-attention을 적용하기 때문에 2차원의 지역적인 정보를 유지해 주지는 못합니다. 때문에 CNN은 학습 후에 가중치가 고정되는 것과 달리 transformer은 가중치가 입력값에 따라 유동적으로 변한다는 특징을 갖습니다. 따라서 transformer의 경우에는 inductive bias가 낮지만 모델의 자유도가 높다고 할 수 있는 것입니다.

학습 데이터가 충분하지 않은 경우에는 CNN모델에 비해서 inductive bias가 낮아 성능이 감소하는 VIT의 단점을 극복한 모델이 오늘 소개할 DeiT인 것입니다.

Prerequisites

VISION TRANSFORMER (VIT)



논문을 이해하기 위한 첫번째 사전 지식인 vision transformer에 대해 간단하게 소개해 보도록 하겠습니다. VIT는 JFT-300M이라는 거대한 데이터를 이용해서 학습을 했습니다. 해당 모델은 SoTA 보다 뛰어난 성능을 약 1/15의 계산비용만으로 얻을 수 있다고 합니다.

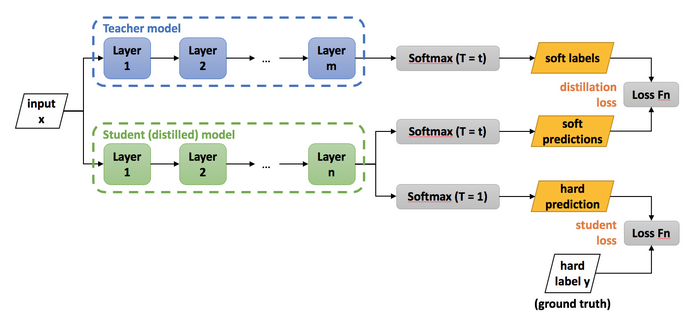
VIT는 결국 NLP에만 적용이 되던 transformer을 직접 이미지 처리에 적용하는 방법을 제안한 모델입니다. 기존의 이미지 self-attention을 사용하던 모델에 주로 도입되던 image-specific inductive bias를 사용하지 않고 각각의 이미지를 하나의 patch로 인식해서 transformer에서의 encoder층에 적용을 합니다. Transformer은 token embedding의 1차원의 시계열 데이터를 입력으로 받으며 이미지 patch로 처리된 데이터를 학습하기 위해 2차원의 이미지 patch 데이터로 바꾸어 줍니다.

왼쪽의 그림을 통해 구조를 파악하는 것이 가능한데, 여기서 원래 이미지의 해상도가 (H, W)였다면 이를 p의 제곱개의 조각, 즉 patch로 나누어서 일정한 길이의 시계열 형태의 데이터로 바꾸어 각각의 patch image의 해상도는 (p,p)가 될 것입니다. 모델은 transformer 모델의 encoder block과 두가지의 차이점을 제외하고 동일한 architecture을 사용합니다. 첫번째 차이점은 pre-norm, 즉 normalization을 하는 부분이 multi-head attention의 전에 존재한다는 것이고 두번째 차이점은 활성화 함수로 relu가 아닌 gelu를 사용한다는 점입니다. Transformer과 마찬가지로 multi head self attention과 mlp block을 통과한 image patch에 관한 정보와 문맥 정보를 다루는 positional encoding 정보를 합쳐서 가중치 행렬을 학습해 예측에 사용하게 되는 것입니다.

KNOWLEDGE DISTILLATION

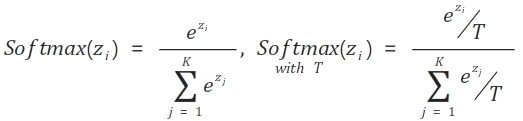
*Hinton et al., "Distilling the Knowledge in a Neural Network" (NIPS 2014)*

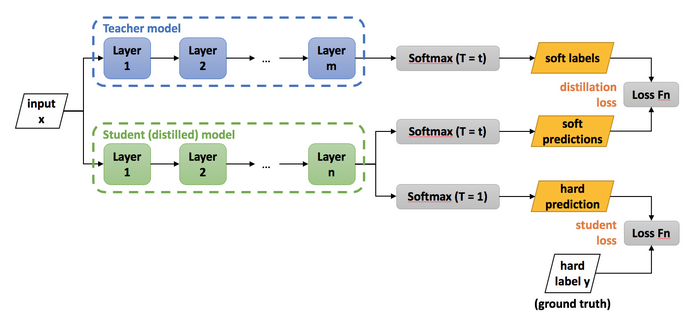
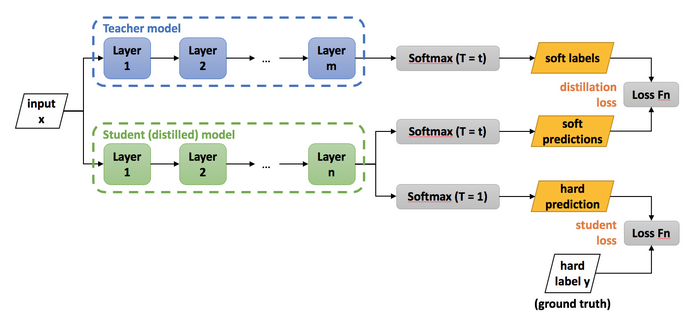
두번째 소개할 사전 지식은 knowledge distillation, 즉 지식 증류라고 불리는 것입니다. 이는 미리 잘 학습된 큰 네트워크, 즉 teacher network의 지식을 실제로 사용하고자 하는 작은 네트워크인 child network에게 전달하여 작은 네트워크에서도 큰 네트워크와 비슷한 성능을 내는 것이 목적인 학습 방법입니다. 일반적으로 teacher model은 특정 task와 데이터에 대해 학습이 잘 된, 파라미터의 개수가 많은 모델이며, student model은 파라미터의 개수가 적은 모델로 설정을 하게 됩니다.

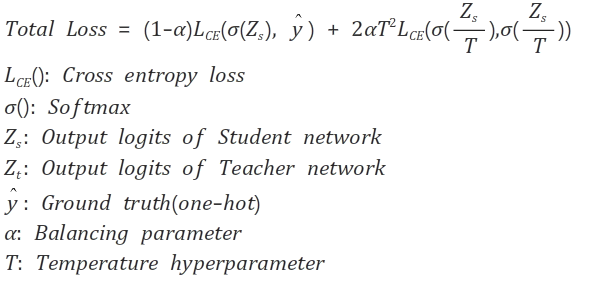


위의 네트워크를 보면 이해를 하는 것이 쉬울 것입니다. Child network의 분류 성능에 대한 loss는 실제 분류 결과와의 차이를 cross entropy loss를 이용해서 계산하고, teacher network와 child network의 분류 결과의 차이를 loss에 포함시키면서 계산하게 됩니다. 두 분류 결과가 같으면 둘 중에 더 작은 값을 선택하게 되는 것입니다. 그림에 표기된 ‘soft label’이란 각각의 class에 대한 예측도의 합이 1이 되도록 했다는 것인데, 이는 hard label처럼 one-hot encoding을 적용하는 것과 같은 방법보다 공존하는 특징을 잘 반영할 수 있습니다. 따라서 이러한 정보의 손실이 없이 두 network의 분류 결과를 비교해서 student network가 teacher network를 모방하도록 학습을 시킵니다.

------🡪





위에 주어진 전체 손실에 대한 식을 잠깐 살펴보겠습니다. 왼쪽항은 student loss를 계산하는 부분입니다. 알파값은 왼쪽항과 오른쪽 항에 대한 가중치이기 때문에 만약에 알파값이 크면 오른쪽 항의 loss를 더 중요하게 간주한다는 의미입니다. 그리고 오른쪽의 항은 distillation loss를 계산하는 부분입니다. 이 부분에서는 softmax 함수에 T, 즉 temperature을 적용하여 작은 값은 더 작게, 큰 값은 더 크게 함으로서 soft label을 사용하는 것에 대한 이점을 최대화합니다.

ARCHITECTURE

BAG OF TRICKS

VIT를 능가하는 성능을 보이게 한 요소는 다양한 차원에서 fine-tuning과정을 거쳤기 때문입니다. 사실 해당 논문에서 소개하는 DeiT라는 모델은 기존의 vision transformer의 구조 뿐만 아니라 학습 방법 또한 동일하게 적용하고 있다고 볼 수 있습니다. 그럼에도 이 모델이 ImageNet으로만 학습을 시켰음에도 불구하고 높은 성능을 보일 수 있었던 이유는 효과가 좋은 hyperparameter tuning의 과정을 거쳤기 때문입니다.

그중에서도 transformer은 많은 양의 데이터를 학습에 필요로 한다는 특징 때문에 data-augmentation과정에 힘을 썼다고 저자들은 밝힙니다. Auto-augment, Random-Augment, 그리고 random-erasing등의 dropout을 제외한 모든 augmentation 과정이 효과적이었다고 합니다.

----🡪

학습 과정에서의 hyperparameter 설정을 위해서는 cross-validation을 하면서 3개의 서로 다른 학습율과 가중치 감소치를 적용했으며, 그 결과 표에서 타원으로 표시한 해당 부분과 같은 수식을 적용해서 학습율을 설정하는 것이 제일 효과적이었다고 합니다.

만약에 전통적인 image scaling technique인 bilinear interpolation같은 기법을 사용하게 되면 l2 정규화가 주변의 영역에 비해 감소하게 된다는 단점이 발생하게 됩니다. 따라서 이를 방지하기 위해 논문의 저자들은 벡터들의 norm을 최대한 유지할 수 있는 방향을 제공해 주는 bicubic interpolation을 사용하였습니다. Bicubic interpolation과 bilinear interpolation은 둘 다 2차원 interpolation과정은 맞지만, bilinear은 오직 4개의 픽셀을 이용하는 것과 달리 16개의 픽셀을 사용함으로서 이미지를 더 정교하게 재구성할 수 있게 되는 것입니다.

KNOWLEDGE DISTILLATION

기존에 사용하던 distillation은 위와 소개한 것과 같은 방법을 사용하지만, 논문의 저자들이 사용한 distillation 방법론은 조금 변화를 주었다고 합니다.

---------🡪

해당 그림은 deit의 구조인데, 왼쪽 부분은 vit의 구조와 완전히 동일하고 오른쪽 부분은 추가한 distillation token입니다. Class token과 비슷하게 self-attention을 통해서 다른 embedding과 상호작용합니다. 그러나 class token과의 차이로는 이 distillation token은 마지막 층의 출력값이며, 추가함으로서 기대하는 효과는 실제 예측해야 하는, 학습과정에서 주어지는 입력 target class가 아닌 teacher network가 distillation과정에서 예측하는 hard label을 학생 네트워크가 쉽게 모방할 수 있도록 하는 것입니다.

-------🡪

이렇게 hard label에 대한 정보를 추가적으로 제공해 주는 방법을 택함으로서 보이는 고양이 그림과 같은 문제를 해결 할 수 있었다고 논문의 저자들은 언급합니다. 고양이 문제란 데이터를 cropping하는 등의 전처리 과정에서 예측하게 된 이미지 데이터의 영역이 실제 값과 벗어나는 경우를 teacher network가 예측한 정보를 통해서 학생 네트워크가 오류를 범하는 경우에 일종의 규제를 가해 줄 수 있는 것입니다.

EXPERIMENTS

마지막으로 논문에서 언급한 실험들과 그 결과의 해석에 대해서 설명해 보도록 하겠습니다.

DISTILLATION

Distillation method를 이용하는 방법에 대해서는 크게 세가지의 결론을 도출 할 수 있었습니다. 우선 어떤 모델을 teacher model로 사용하는지에 대해서 학습을 한 결과, 같은 종류의 transformer 모델인 deit-B를 사용하였을 때에 비해서 합성곱 신경망믈 사용하는 전통적인 이미지 분류 모델인 RegNet을 사용했을 때 오히려 예측의 정확도가 높았음이 확인 가능합니다.

그리고 distillation에 hard distillation을 추가하였을 때 더 높은 정확도를 보임으로서 Deit 모델의 성능을 다시 한번 입증할 수 있었습니다.

EFFICIENCY vs ACCURACY

딥러닝 학습에 있어서 흔히 발생하는 문제 중 하나는 효율성과 정확성에 있습니다. 이 두 부분이 모두 높은 값을 보이도록 하는 것이 모델을 개선하는데 제일 중요한데, Deit-B의 경우에는 이를 측정하기 위해 이미지 처리 속도와 정확도의 두 수치를 이용하였습니다. 논문에서 제공한 표를 바탕으로 주요한 내용을 표로 정리한 결과를 보시면 data augmentation과 knowledge distillation을 모두 적용한 deit 모델이 imagenet를 이용해서 실험한 efficient net의 정확도와 이미지 처리 속도에 있어서 모두 개선된 모습을 보였음이 확인 가능했다고 합니다.

이상으로 DeiT모델에 대한 발표를 마치도록 하겠습니다.

감사합니다.