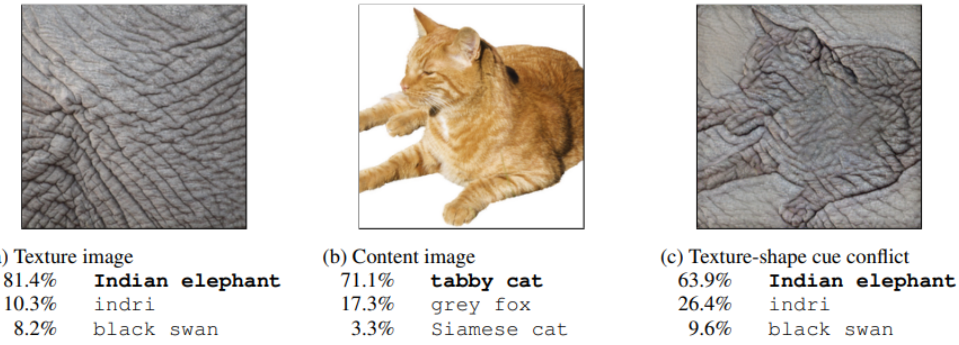
AI502 Final Project Proposal

20233189 김준영, 20233339 박현서

Ⅰ. Problem and Goal

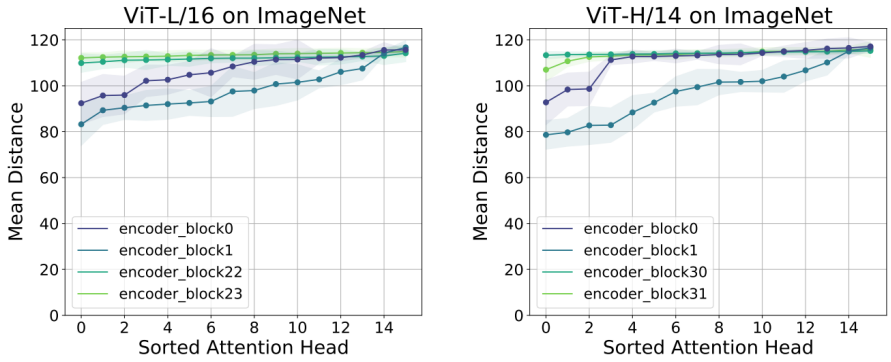
1. Problem

"*ImageNet-Trained CNNs are Biased Towards Texture; Increasing Shape Bias Improves Accuracy and Robustness*"라는 논문은 ImageNet으로 학습한 CNN 모델이 object의 shape보다 texture에 집중해서 classify한다는 결과를 제시했다. CNN의 구조는 얕은



layer에서는 image의 local 정보를 학습하고, 깊은 layer에서는 image의 global 정보를 학습하도록 설계되었다. 만약 이 논문에서 제시한 결과가 CNN의 구조와 관련이 있다면 CNN이 얕은 layer에서 검출한 texture, color와 같은 local 정보가 classification에 더 dominant하게 영향을 줬다고 추측해볼 수 있다. 이 가설을 확인하기 위해서 CNN과 다른 구조를 가지고 image classification task를 수행하는 Vision Transformer에서는 어떤 결과가 나오는지 알아보고자 한다.

Vision Transformer는 self-attention을 통해 모든 image patch들 사이의 연관도를 배울 수 있다. 즉, Vision Transformer는 image의 global context를 이해할 수 있다. "*Do Vision Transformers See Like Convolutional Neural Networks?*"라는 논문이 제시한 결과 중 하나에 의하면, ImageNet으로 학습한 Vision Transformer는 얕은 layer에서도 global 정보를



학습한다. 위 그림에서 Mean Distance가 높을수록 멀리 있는 image patch와의 attention 값이 높다는 뜻이다. 이 결과로부터 Vision Transformer는 CNN과 달리 object의 texture보다 global 정보에 해당하는 shape에 집중해서 classify할 것이라고 기대해볼 수 있다.

2. Goal

ImageNet 21K로 pretrained된 ResNet50과 ViT(Vision Transformer)를 비교했을 때, ResNet50은 texture-biased decision을 내릴 것이고 ViT는 그렇지 않을 것이다. 이 가설을 확인함으로써 CNN과 ViT의 학습 방법의 차이에 대해 이해할 수 있을 것이다.

Ⅱ. Dataset

- **Cue conflict using Stylized-ImageNet (for test)**

Stylized-ImageNet generation framework을 사용하여 shape와 texture를 합성한 데이터인 'Cue conflict'를 생성한다.

- Dataset for shape: ImageNet 1k
- Dataset for texture: Paint by numbers

Ⅲ. Evaluation metric

1. Test accuracy for Imagenet 1k

먼저 ResNet50과 ViT의 test accuracy를 통해 두 모델이 얼마나 image classification을 잘 수행하는지 확인한다.

2. Test accuracy for Cue conflict

ResNet50과 ViT에 cue conflict를 입력으로 넣었을 때 test accuracy 측정한다. 이때 모델이 shape dataset에서 잘 예측한 image에 대해서만 cue conflict를 생성하고, ground truth label은 shape dataset의 label을 사용한다.