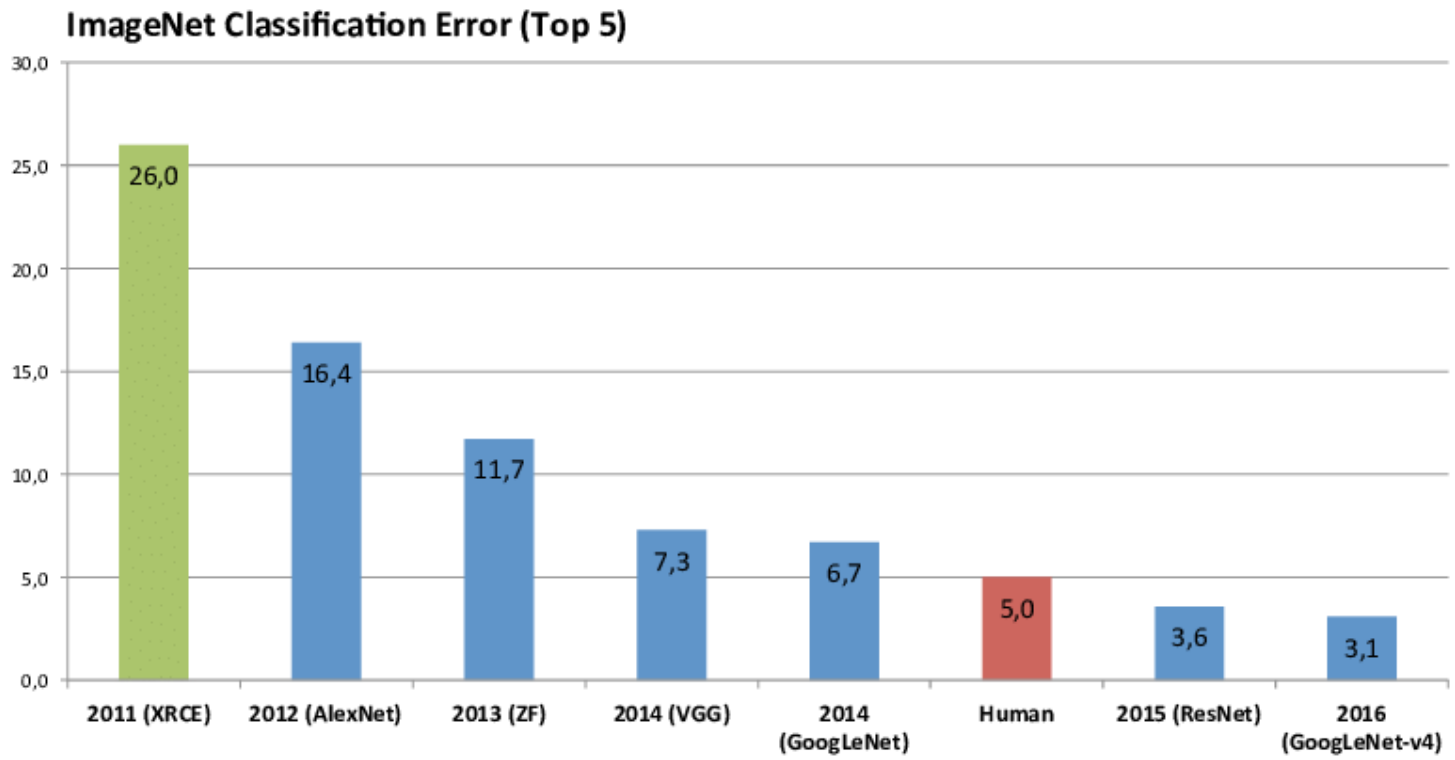
CNN 모델의 한계

Why do deep convolutional networks generalize so poorly to small image transformations?

CNN을 이용한 Classification

- 2012년 AlexNet을 시작으로 VGG, GoogleNet, ResNet 등 많은 발전이 있었음
- 특히 ResNet부터는 사람보다 이미지를 더 잘 분류한다



CNN의 특징

작은 필터 크기 + stride

- 크기가 작은 필터로 전체 이미지를 순회하면서 feature를 학습한다
- 여러 필터를 겹치면 수많은 feature 학습 가능

깊은 필터 크기

- 필터를 깊게 쌓으면 쌓을수록 더 개념적인 feature를 학습할 수 있다.

Pooling

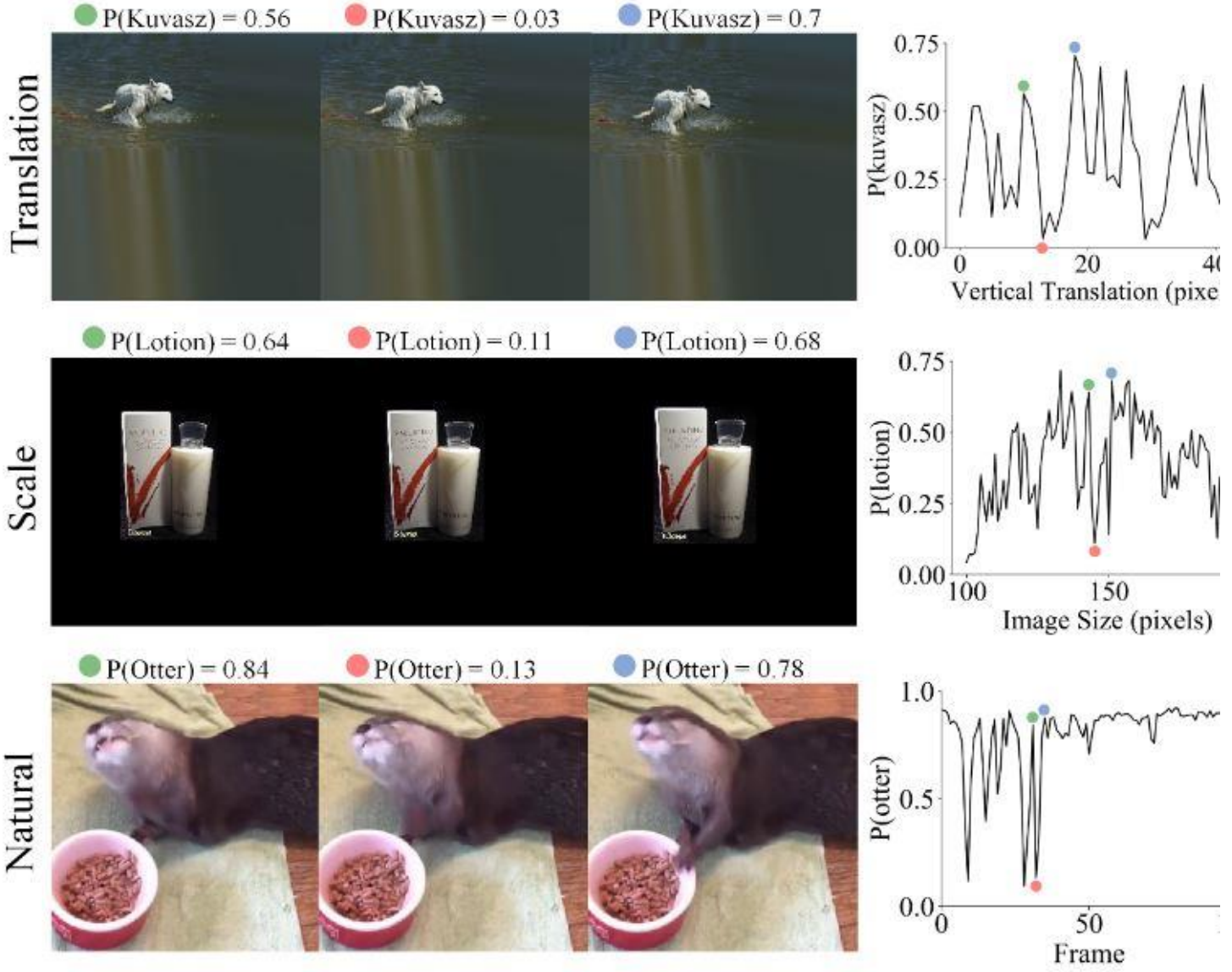
- 이미지에 변형이 있어도 강하다
- 위치가 조금 움직인 것은 Pooling에 의해 무시될 수 있다.

....?

- ??

CNN을 이용한 Classification의 문제

- 그런데 사람의 prediction에 영향을 한 미치는 작은 변형만 있어도 score가 크게 변한다
- 비디오의 연속된 프레임을 모델에 넣었을 뿐인데, score가 크게 변화하고 있다
- "위치가 조금 움직인 것은 Pooling에 의해 무시될 수 있다."....???



기초 개념

Sampling Theorem

- 샘플링 주파수가 신호의 최대 주파수 두 배 이상이라면 표본으로부터 원래 신호를 완전하게 구성할 수 있다.
- 예: 사람이 들을 수 있는 소리의 최대 주파수는 20kHz이다. 그래서 녹음할 때 샘플링 주파수는 2배 이상인 44.1kHz이다.

Stride

- 큰 이미지에 대해 작은 필터를 적용할 때, stride만큼 밀면서 이미지 전체를 순회한다.

Invariant(불변량, 불변성)

- 어떤 유형의 변형이 객체에 적용될 때 변경되지 않고 보존되는 속성

Why?

- CNN의 striding이 subsampling이다
  - 필터가 한 번에 이미지의 일부만 적용되므로
- Convolution과 subsampling에 기반한 시스템에선 translation invariant를 보장하지 못한다.
  - 논문: Simoncelli et al.
  - Convolution과 subsampling에 기반한 시스템 = CNN 모델
  - translation invariant를 보장하지 못한다 = 1픽셀만 shift해도 결과 보장을 못 한다
- 증명은 생략
  - 통계학, 신호처리이론 등을 완비하면 이해 가능.

결론

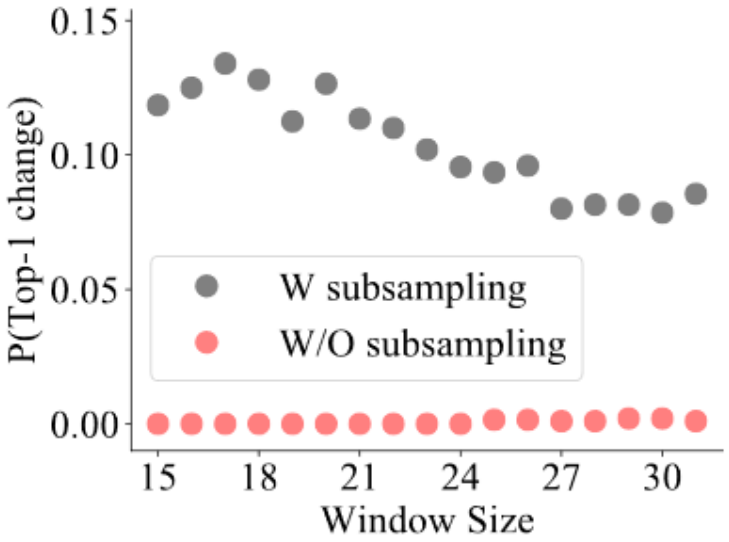
striding을 사용하는 대부분의 CNN 모델은, 작은 translation만 변형해도 이미지 고유의 특성이 변할 수 있다.

이미지를 1픽셀만큼만 shift해도 모델 결과가 막 변할 수 있다.

Pooling 덕분에 CNN 모델은 이미지 변형에 강하다는 것은 거짓말.

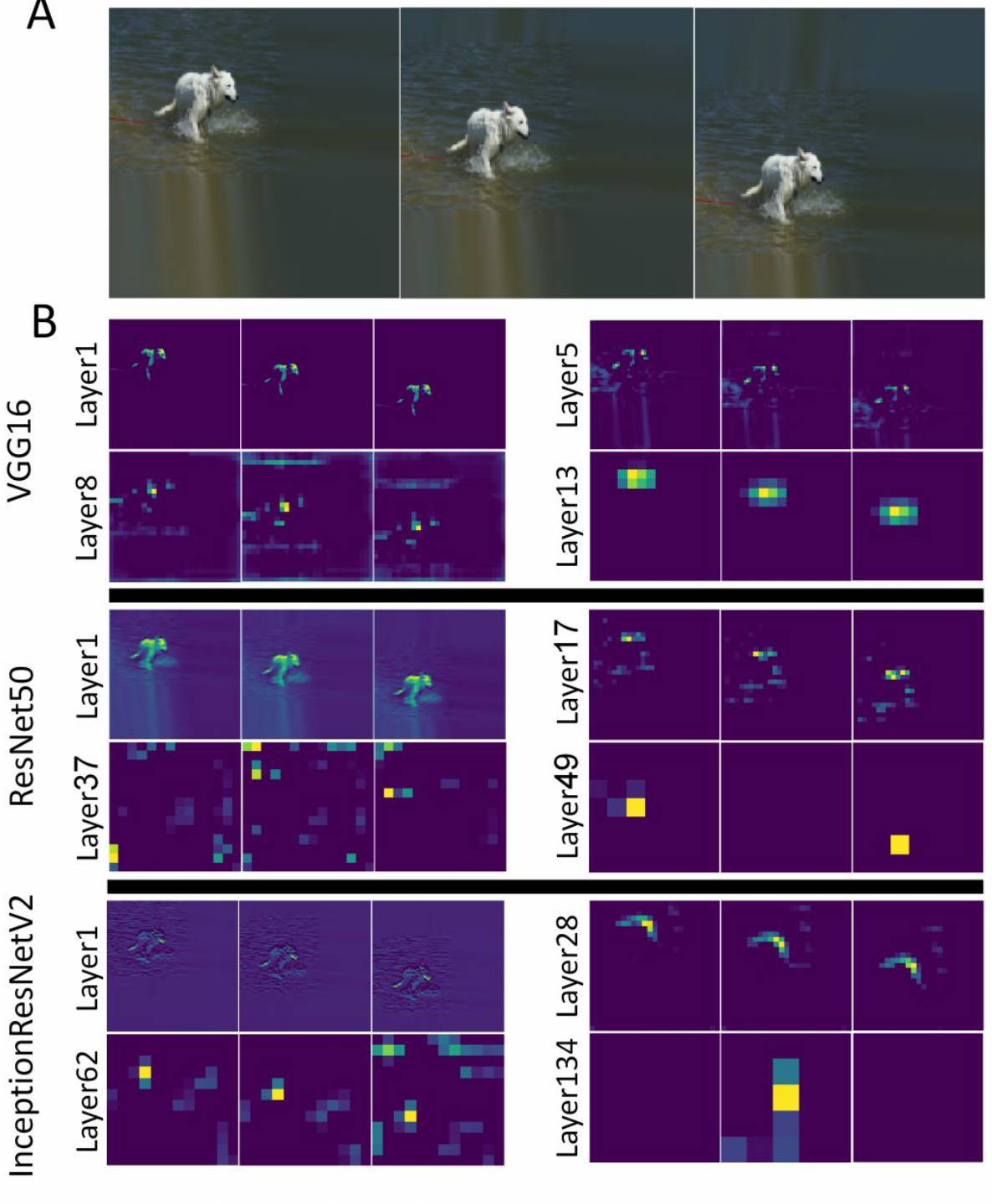
Subsampling 여부에 따른 실험

- 실험
  - Alex 비슷한 모델 사용
  - CIFAR-10 dataset classification
- 결과
  - Accuracy: 두 모델 모두 비슷
  - Subsampling을 사용하면 translation 발생시 top-1 class가 변할 확률이 0.1 이상



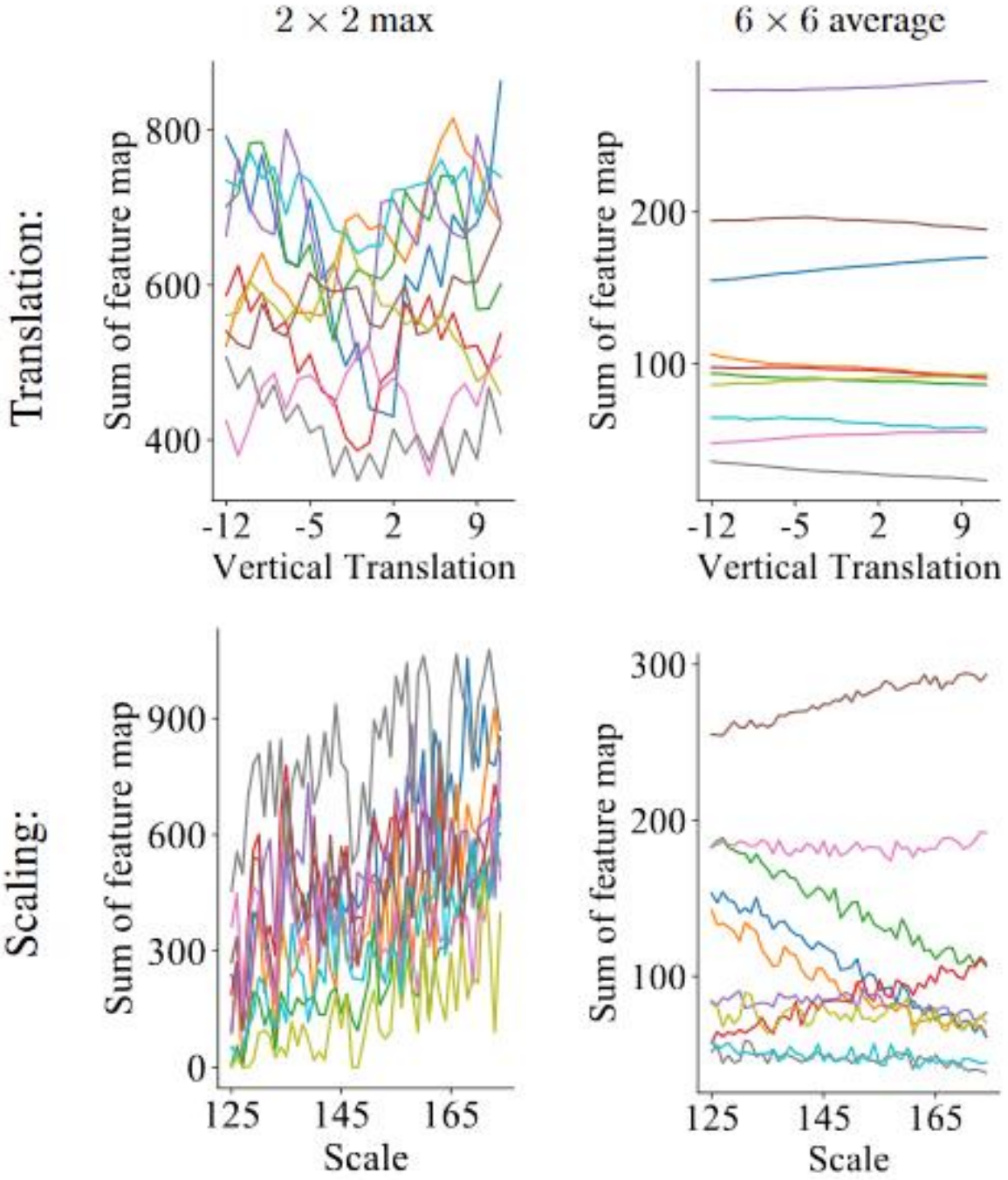
Translation에 따른 layer activation

- 연속된 비디오의 세 프레임을 모델에 넣어보자
- 세 모델 모두 마지막 layer는 포함하였다.
- VGG16는 마지막 layer activation이 물체 이동에 따라 어느 정도 유지되고 있다.
- 더 깊은 모델(ResNet, InceptionResNet)은 물체가 옆으로 이동하면 마지막 layer activation이 들쭉날쭉



VGG를 고쳐 보자

- Pre-trained된 VGG16 모델을 가져온다.
- 2X2 Max Pooling을 6X6 Average Pooling으로 바꾼 뒤, 마지막 layer만 재학습 후 테스트
- Vertical translation을 해도 feature의 activation은 경향이 비슷
- Sacle시에는 feature activation이 확연하게 변함
- Top-1 Accuracy가 0.8에서 0.3으로 감수



Modern CNN model이 invariant한 이유

Dataset Bias

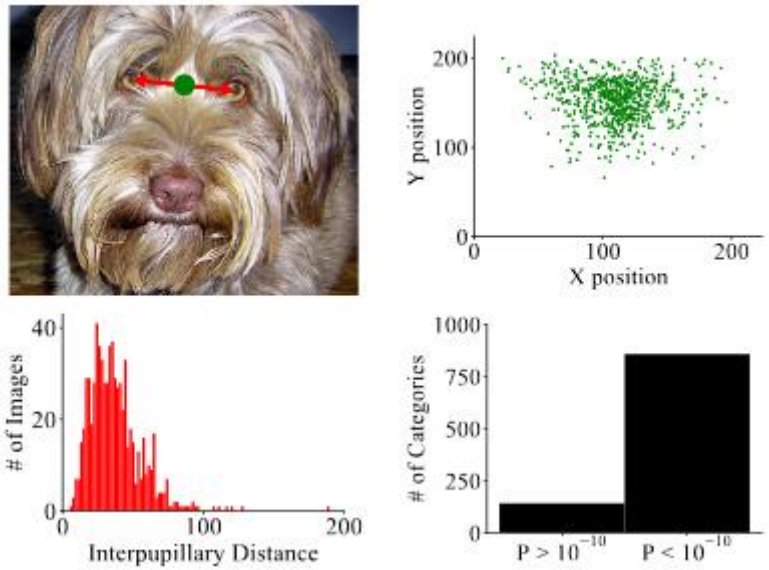
- ImageNet 데이터는 object 위치와 크기가 극히 편향되어 있다
- 찍힌 물체가 사진 중앙에 있다거나, 물체 크기가 비슷비슷하다거나 등등

Data Argumentation을 사용해도 불충분

- Data argumentation: rescaling, translation, rotation 등등 이미지 변형
- 최신 모델의 subsampling facto가 60
- translation에 대한 invariant를 학습하는 데만도 60*60=3600개의 data argumentation된 이미지 필요
- Rotation, rescaling까지 하려면…? 3600을 더 곱해야 함

ImageNet 데이터의 편향성

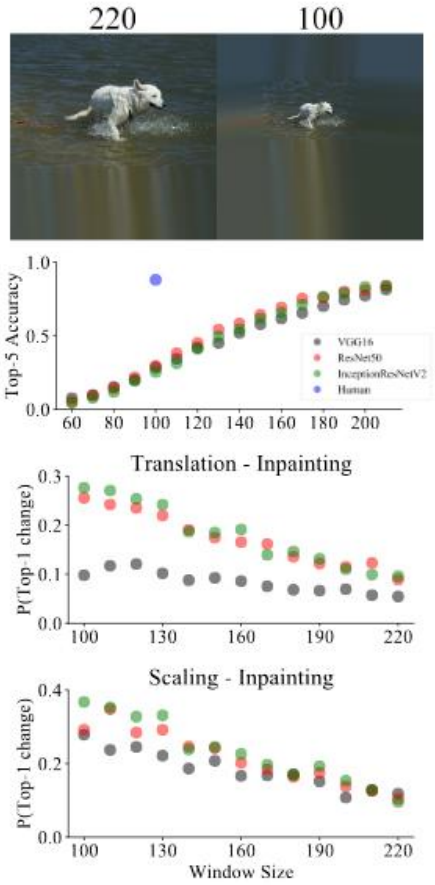
- ImageNet 데이터는 물체 위치나 물체 크기가 편향되어 있다는 연구 결과가 있다.
- 그러나 데이터 편향성을 해결한다고 해서 CNN 모델이 translation에 강해지진 않음.
- 편향된 데이터로 학습 후, 편향된 데이터로 테스트해서 classification Accuracy가 높을 수도 있다.



제한된 transform invariant라도 챙기자

- 이미지 embedding size를 모델이 예측한 사이즈, 즉 모델 입력 크기로 조정한다.
- 이미지 크기가 220에 가까워질수록 accuracy가 커진다
- translation, Scaling: 이미지/window size가 220에 가까울수록 top-1이 바뀔 확률이 줄어든다

.



결론

CNN 모델로 학습을 시키면 편향성이 생긴다

- 데이터 자체에 편향성이 있으면 모델이 편향성을 학습한다.
- Data argumentation을 아무리 많이 해도 invariant를 학습할 만큼의 데이터는 모을 수 없다.

그러나 학습 데이터에 최대한 편향성을 없애자.

- 하지만 사실상 불가능.

모델의 Input 데이터는 최대한 학습 데이터와 유사하게 만들자.

- 이미지 크기를 모델이 가정한 크기로 한다.
- 이미지의 경우, 물체 위치나 물체 크기를 비슷하게 맞춰보자.

논의

CNN 모델의 invariant는 거저 얻어지지 않는다

ImageNet 데이터셋은 사진가의 bias에 의해 편향되어 있어서, ImageNet으로 학습된 모델은 invariant할 확률이 크다

게다가 Sampling Theorem에 의해 CNN 모델은 Transform에 대해 편향적일 수 밖에 없다.

train data와 닮지 않을수록(이미지 크기 차이 등) CNN 모델의 성능은 떨어질 것이다

끝