The Unreasonable Effectiveness of Deep Features as a Perceptual Metric

성균관대학교 인공지능융합전공 이나은

차례

Motivation

Difficulty in modeling human perceptual similarity judgment

컴퓨터 비전 분야에서 이미지 사이의 유사성을 판단하는 것은 매우 어렵다.

- high-dimensional
- subjective
- aiming to mimic human visual perception

Classic per-pixel measures

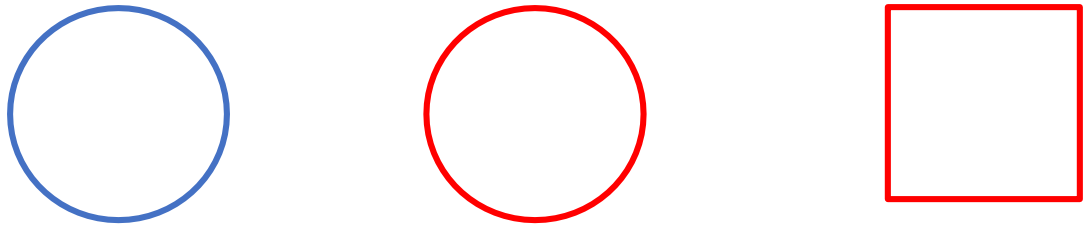
- l2 distance, Peak Signal-to-Noise Ratio (PSNR)
- Image compression
- Blurring

→ SSIM, MSSIM, FSIM, HDR-VDP

Difficulty in modeling human perceptual similarity judgment

인간의 인지적 판단과 부합하는 perceptual metric을 구축하는 것은 어렵다.

- depending on high-order image structure
- context-dependent
- may not actually constitute a distance metric



Context-dependent and pairwise nature of the judgments로 인해 human perceptual similarity를 모델링 하는 것은 어렵다.

There be a way to learn a notion of perceptual similarity without directly training for it?

Usefulness of deep convolutional networks

Deep convolutional network가 image classification 의 목적으로 훈련이 되었더라도 해당 네트워크의 internal activation이 다른 task에도 유용하게 쓰인다.

ex) Features from the VGG architecture are used to neural style transfer, image superresolution, and so on.

위와 같은 image regression problem에서는 VGG feature space에서 distance를 계산하여 “perceptual loss”로서 사용한다.

1. “Perceptual losses”가 실제로 human visual perception에 잘 대응될까?
2. Traditional perceptual image evaluation metrics과 어떻게 비교할까?
3. Network architecture가 중요할까?
4. 꼭 ImageNet classification task에서만 적용될까?
5. 네트워크가 훈련이 되어야 할까?

→ Berkeley-Adobe Perceptual Patch Similarity(BAPPS) dataset과
Learned Perceptual Image Patch Similarity (LPIPS) metric 제안

Similarity Judgments

2AFC & JND test

1. 2AFC (Two alternative forced choice)

한 개의 reference image와 두 개의 distorted images가 있을 때 distorted image 중 어떤 것이 reference와 유사한지 선택

2. JND (Just noticeable differences)

- 한 개의 reference image와 한 개의 distorted images가 있을 때 두 이미지가 서로 같은 지 혹은 다른 지 구분
- validation 용

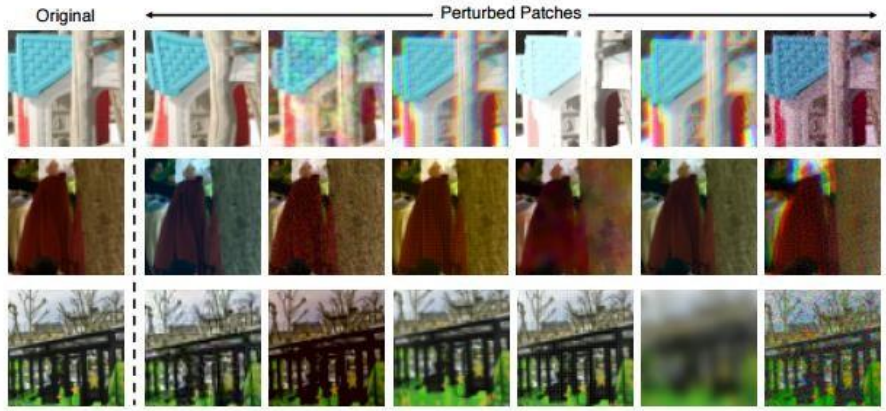
Distortions

Distortions

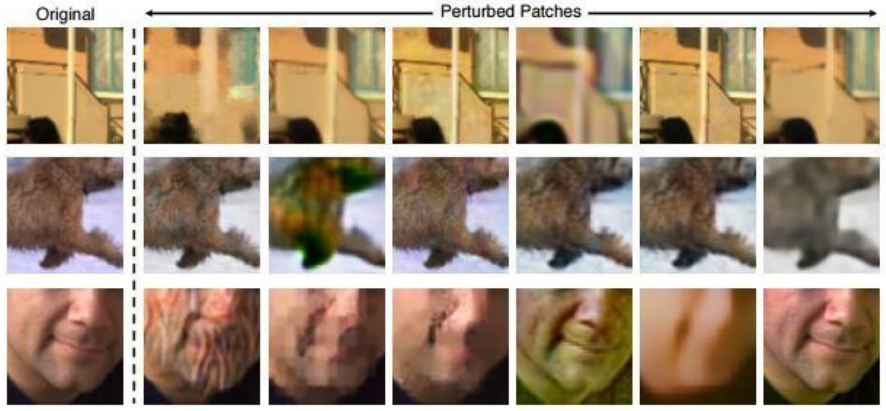
Sub-type	Distortion type
Photometric	lightness shift, color shift, contrast, saturation
Noise	uniform white noise, Gaussian white, pink, & blue noise, Gaussian colored (between violet and brown) noise, checkerboard artifact
Blur	Gaussian, bilateral filtering
Spatial	shifting, affine warp, homography, linear warping, cubic warping, ghosting, chromatic aberration,
Compression	jpeg

Parameter type	Parameters
Input corruption	null, pink noise, white noise, color removal, downsampling
Generator network architecture	# layers, # skip connections, # layers with dropout, force skip connection at highest layer, upsampling method, normalization method, first layer stride # channels in 1^{st} layer, max # channels
Discriminator	number of layers
Loss/Learning	weighting on oixel-wise (ℓ_1), VGG, discriminator losses, learning rate

다양한 distortion을 사용하는 이유? 현실 세계의 distortion (real algorithm output)을 구현



(a) **Traditional**



(b) **CNN-based**

BAPPS dataset

Dataset composition

1. 2AFC (Two alternative forced choice)

각 sub folder : ref, p0, p1, judge (0, 1)

train	traditional
	cnn
	mix
validation	traditional
	cnn
	superres
	deblur
	color
	frameinterp

* real algorithm output

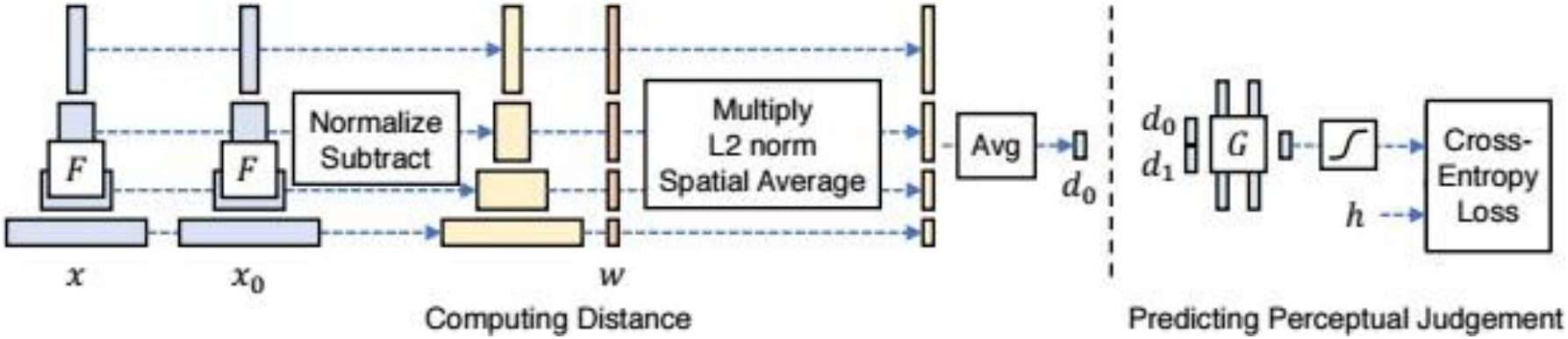
2. JND (Just noticeable differences)

각 sub folder : p0, p1, same (0, 1)

validation	traditional
	cnn

Experiments & Results

Experiments



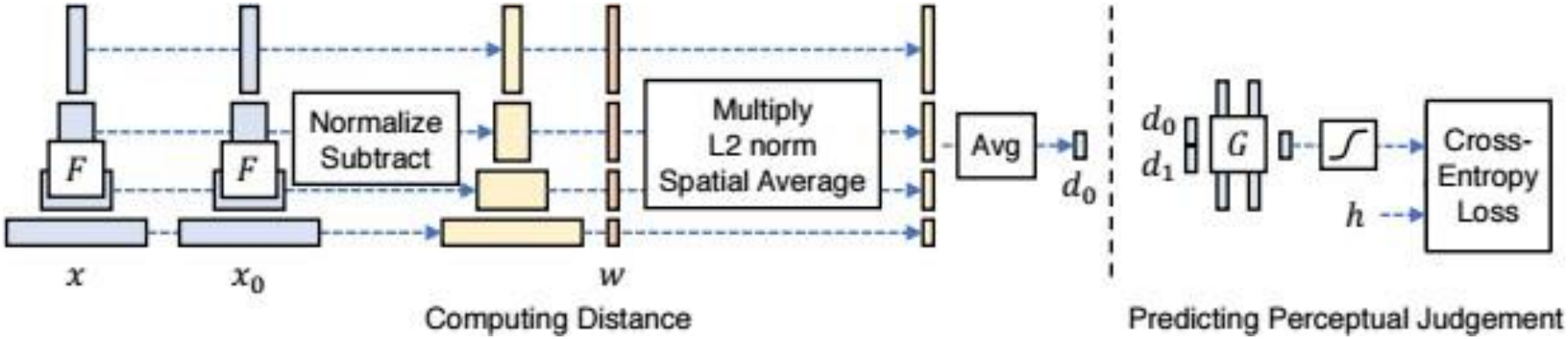
1. Network activations to distance

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2$$

x, x0 : reference, distorted image
F : network

1) y : feature들을 channel별로 unit normalization한 결과 -> ex) (R, G, B)
2) w_l : scaling vector
3) l2 distance 구하기 (spatially)
4) Spatially normalization
5) Sum channel-wise

Experiments



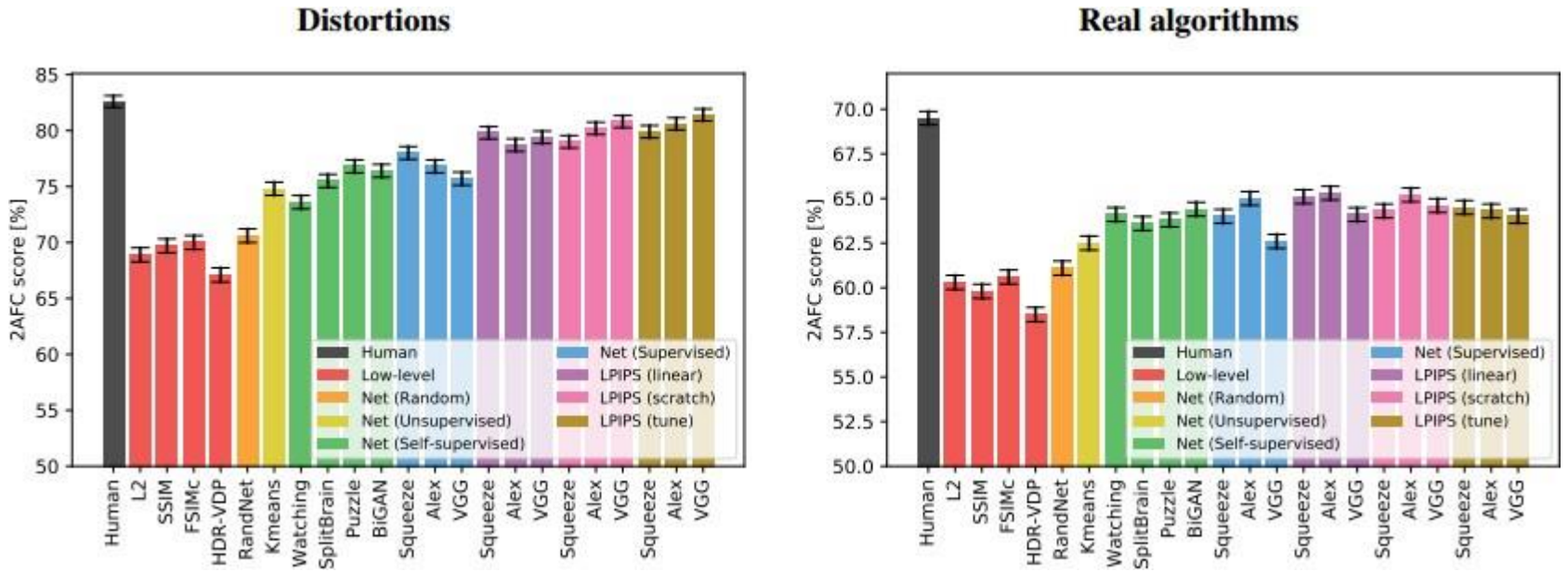
2. Predicting perceptual judgement

하나의 reference image와 두 개의 distorted image가 있을 때 두 번의 distance 거리를 측정 (이전 슬라이드)
-> G는 classification model로서 유사성을 판별하도록 함
-> h : perceptual judgment (0 or 1)
-> cross entropy loss : 0~1 사이의 범위를 가지며 분류 모델의 훈련 시 사용

Experiments

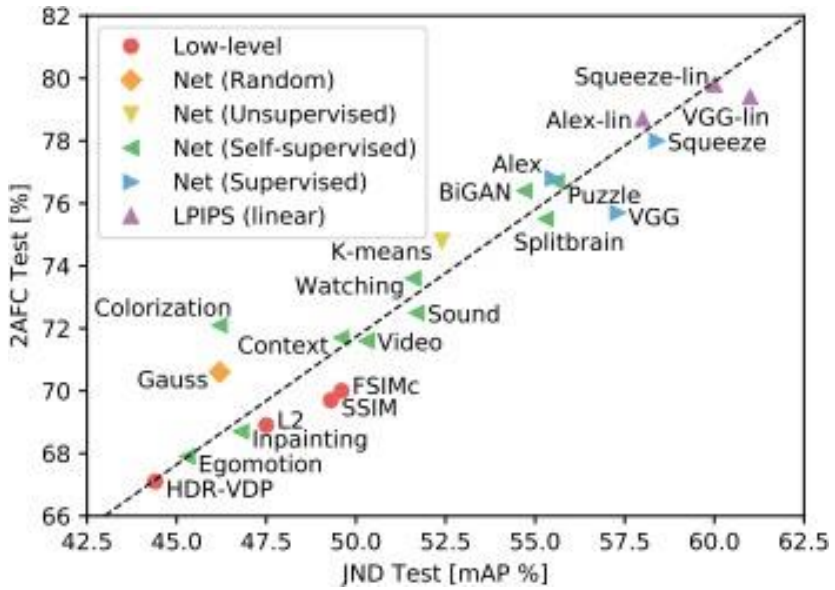
3. Learned Perceptual Image Patch Similarity (LPIPS)

1) lin (linear) : pre-train된 network의 가중치 F를 고정하고 linear한 가중치 w를 그 위에 학습
2) tune : pre-train된 classification model의 가중치로 초기화 후, network에 대한 모든 가중치를 fine-tuning
3) scratch : random한 Gaussian 가중치에서 초기화하고, judgment로 전체를 학습



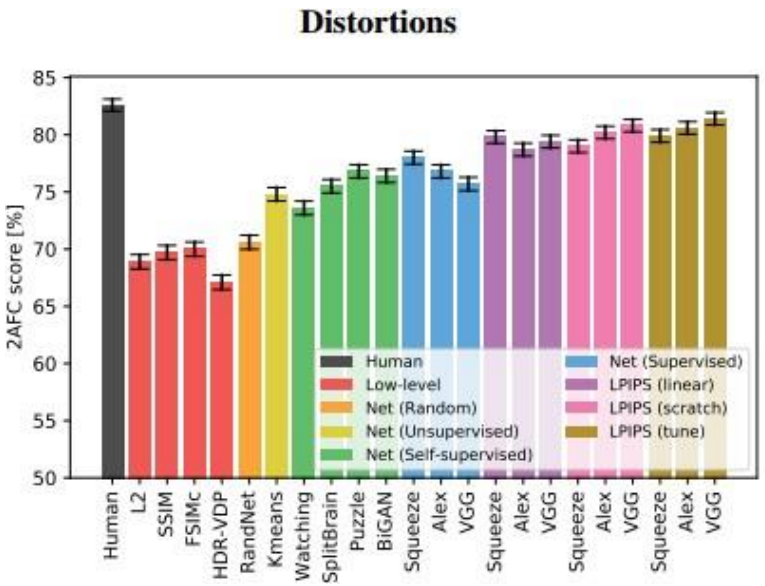
- Distortions에서 tune을 사용했을 때 가장 좋은 성능
- Real algorithm (실제 왜곡된 데이터) : 일반화한 결과로 볼 수 있으며 lin과 scratch가 좋은 성능을 보임. 반면 tune은 다소 낮은 성능을 보임
-> 논문의 한계

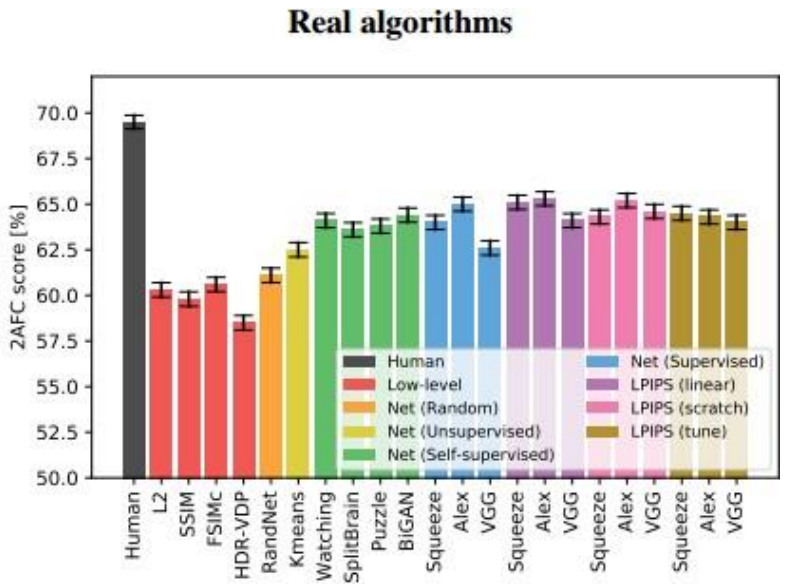
Results

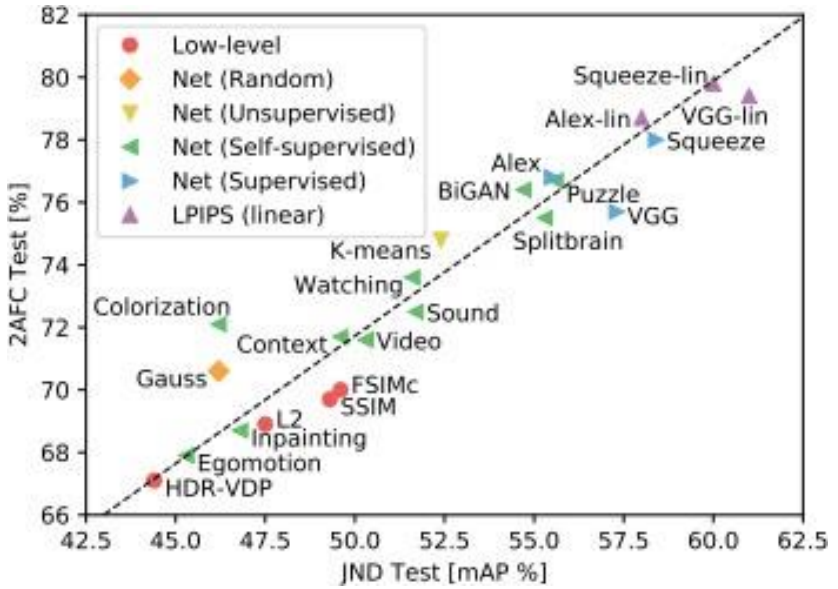


1. Low-level metric보다 LPIPS가 더 성능이 좋다.
2. Supervised, self-supervised, unsupervised 목표로 학습된 deep feature들이 perceptual similarity를 잘 모델링한다. 이는 이전의 널리 사용된 metric들을 뛰어넘는 결과를 보였다.
3. Network architecture만으로는 성능을 설명하지 못한다.

Results







1. “Perceptual losses”가 실제로 human visual perception에 잘 대응될까? (figure 1)
2. Traditional perceptual image evaluation metrics과 어떻게 비교할까? (figure 1, 2)
3. Network architecture가 중요할까? (figure 1, 2)
4. 꼭 ImageNet classification task에서만 적용될까?
5. 네트워크가 훈련이 되어야 할까? (figure 2)