Supplementary Material - Why Are Deep Representations Good Perceptual Quality Features?

Taimoor Tariq[1], Okan Tarhan Tursun[1], Munchurl Kim[2], and Piotr Didyk[1]

[1] Università della Svizzera italiana, Switzerland.
[2] KAIST, South Korea.

1 Input Generation

One of the most important perceptual aspects of Human Visual System (HVS) is the relation between its contrast sensitivity and spatial frequency. This relation is explained by the Contrast Sensitivity Function (CSF) which is estimated by psychophysical experiments involving spatial sine-wave gratings. Our input generation routine is closely related to those input stimuli which are used in CSF experiments.

In order to make CSF function independent from a specific display device, the spatial frequency is defined in *cycles per degree* (c/deg) units. When a display device is used in experiments, it requires a conversion to device dependent units, which is *cycles per pixel*. The conversion between those two types of units can be expressed as a series of simple mathematical formulas using viewing distance and physical dimensions of the screen.

In our experiments, we generate input gratings using *cycles per pixel* units on a hypothetical display. In order to convert the spatial frequencies from CSF into *cycles per pixel*, we fix viewing at a retinal resolution of 60 pixels per degree (PPD) [5].

The relation between *cycles per pixel*, *cycles per degree* is and *pixels per degree* is:

$$\frac{cycles}{pixel} = \frac{cycles}{degree} \times (\frac{pixels}{degree})^{-1}, \tag{1}$$

To measure orientation selectivity, the spatial frequency of the input gratings is set to the frequency where Human CSF peak sensitivity occurs. We set this particular frequency to 8 *cpd* according to Mannos and Sakrison's CSF model [4].

2 Independent Attributes

The PE that we have defined combines both μ_1 and μ_2 with equal contribution. Fig. 1 shows that both of them play their role in the ability of deep representations as perceptual features. When used independently, both the μ_1 and μ_2

perform well as a measure of the perceptual representation power of CNN channels. We observe that even if selected on the basis either μ_1 or μ_2, a small group of good channels (e.g H-10) are better perceptual quality features compared to a much larger group of other channels in the layer (e.g L-50). As future work, an interesting direction might be to introduce additional parameters which are dependent on the relative importance of attributes.

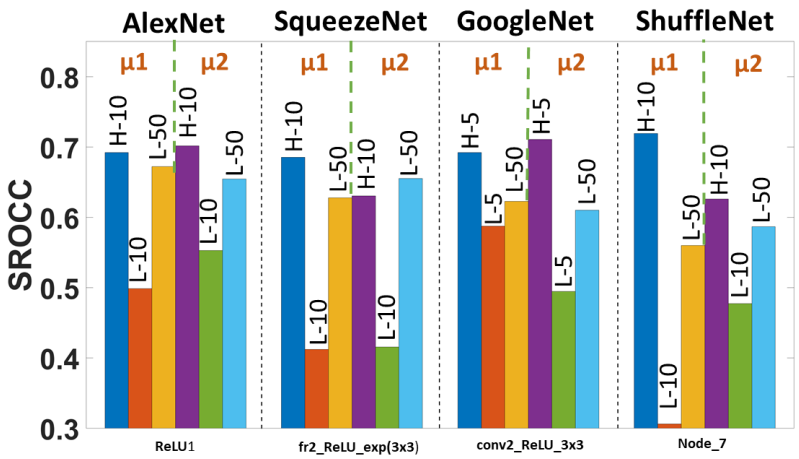

Fig. 1: Independent Attributes. The correlation of feature space distances with human MOS from the LIVE multi-distortion set. For each layer, the set of channels are selected on the basis of μ_1 and μ_2 separately.

3 Distribution of Good Channels

Fig. 2 shows how the PE of channels is distributed in different layers of the VGG-16. It can be seen that only a small percentage of channels in each layer can be characterized as effective perceptual quality features as per PE. Considering our analysis, this result further reinforces the importance of feature selection.

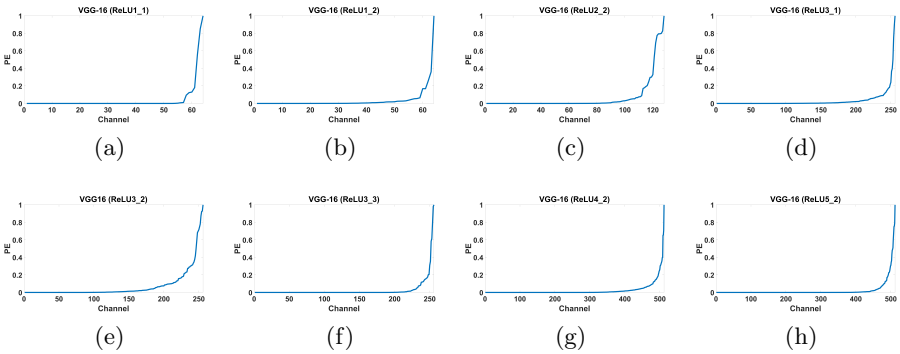

Fig. 2: The distribution of PE for channels in different layers of the VGG-16. It can be seen that good perceptual quality features constitute a relatively small proportion of channels in each layer.

4 What about networks trained on other (not Image classification) tasks?

It is well known that features delivered by pre-trained Image classification CNNs are good perceptual quality features. However, there has been no analysis of CNNs trained on other tasks such as super-resolution, object detection and emotion recognition etc. In Fig. 3, we report results of the correlation of feature space distance with human MOS for images in the LIVE multi-distortion data-set. We use channels from a pre-trained super-resolution network (VDSR [3]), a pre-trained object detection network (Tiny-YOLOv2 [6]) and an emotion recognition network (FER+ [1]). It can be seen that channels of the object detection network are better perceptual quality features as compared to the other networks. As object detection is somewhat related to image classification, the result is expected. Furthermore, our hypothesis extends to networks trained on other tasks as well. Tasks like image classification and object recognition encapsulate the most basic and major function of the human visual system, therefore, the learned representations are better samples of how the HVS extracts features from images. Super-resolution network representations are centered around similarity of images, but Fig. 3 shows that they do not deliver effective perceptual quality features, probably because SR is not a natural task for the HVS. Considering that object detection correlates more to the functionality of the HVS, it is a worthwhile future direction to investigate the features of object recognition CNNs as an alternative to image classification representations as perceptual quality features.

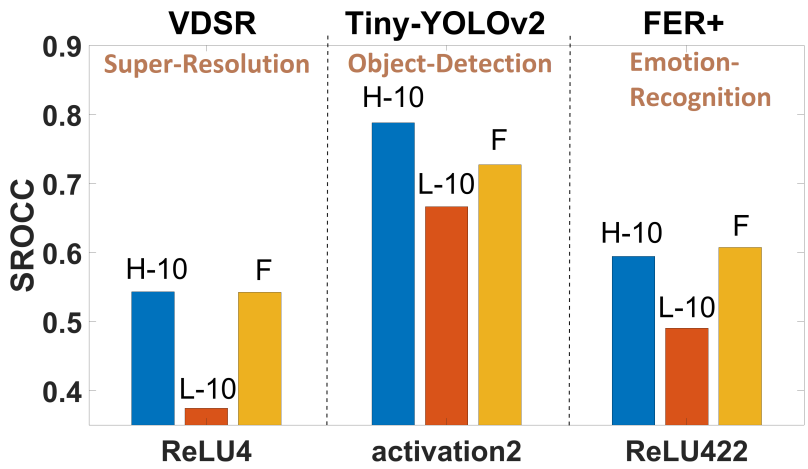

Fig. 3: Non-image classification networks. The correlation of feature space distances with human MOS for the LIVE multi distortion data-set.

5 Additional OQA Results

We present additional results from our OQA experiment in Table 1 and Table 2. The results present the correlation of feature space distances with human MOS for the images from both the LIVE subjective quality dataset [7] and the LIVE multi-distortion dataset [2]. The correlation is quantified with RMSE (lower means better), LCC (higher means better) and SROCC (higher means better) after fitting. The results in Table 1 and 2 supplement the OQA results in the main paper, demonstrating the validity of our hypothesis and analysis. It can be seen that smaller sets of channels which are selected based on PE scores provide a much better set of perceptual quality features compared to even larger sets of other channels in the layer.

Table 1: Objective Quality Assessment Test. The correlation of feature space distances (for different feature subsets) with human subjective assessment of perceptual quality, quantified by DMOS.

Network	Layer	Feature Set	RMSE	LCC	SROCC
VGG-16	ReLU2_2	F	9.8366	0.8146	0.8028
		H-10	8.8286	0.8538	0.8486
		L-10	12.3114	0.6878	0.6806
		L-90	10.5863	0.7813	0.7739
	ReLU4_1	F	9.8149	0.8155	0.8076
		H-2	8.8183	0.8542	0.8476
		L-2	10.2338	0.7874	0.7863
		L-80	9.8485	0.8141	0.8070
AlexNet	ReLU1	F	9.7580	0.8179	0.8155
		H-10	9.1514	0.8419	0.8368
		L-10	12.8110	0.6553	0.6562
		L-70	10.3186	0.7936	0.7931
	ReLU4	F	8.8015	0.8548	0.8605
		H-5	8.5467	0.8637	0.8651
		L-5	9.8927	0.8122	0.8197
		L-50	9.0697	0.8450	0.8507
SqueezeNet	fire2_ ReLU_ exp_3x3	F	11.2791	0.7468	0.7397
		H-10	10.8632	0.7679	0.7625
		L-10	12.6927	0.6632	0.6614
		L-50	11.6555	0.7264	0.7199
	fire6_ ReLU_ exp_3x3	F	11.4191	0.7394	0.7314
		H-5	11.8710	0.7142	0.7017
		L-5	12.6857	0.6637	0.6540
		L-50	12.0600	0.7063	0.6988
ShuffleNet	node7	F	11.0810	0.7570	0.7519
		H-10	9.9055	0.8117	0.8002
		L-10	14.2481	0.5424	0.5583
		L-70	11.6409	0.7272	0.7232
	node17	F	9.1354	0.8425	0.8421
		H-10	8.8577	0.8528	0.8477
		L-10	11.5070	0.7346	0.7407
		L-70	9.2306	0.8389	0.8414

Table 2: Objective Quality Assessment Test. The correlation of feature space distances (for different feature subsets) with human subjective assessment of perceptual quality, quantified by DMOS.

Network	Layer	Feature Set	RMSE	LCC	SROCC
GoogleNet	conv2_ ReLU_ 3x3	F	9.2730	0.8370	0.8351
		H-5	9.1360	0.8425	0.8364
		L-5	12.6595	0.6654	0.6674
		L-80	9.6636	0.8218	0.8203
	inception_ 4a-ReLU_ 3x3	F	10.2264	0.7977	0.8061
		H-5	9.8592	0.8137	0.8201
		L-5	10.8882	0.7667	0.7750
		L-45	10.0326	0.8063	0.8163
MobileNet-v2	block1_ expand_ ReLU	F	11.9441	0.7099	0.7017
		H-10	11.6059	0.7292	0.7256
		L-10	13.7130	0.5884	0.5825
		L-70	12.7912	0.6566	0.6505
	block3_ expand_ ReLU	F	10.1957	0.7991	0.8063
		H-10	9.2423	0.8385	0.8459
		L-10	13.2810	0.6219	0.6223
		L-70	10.7877	0.7716	0.7804
ResNet-18	Res2a_ ReLU	F	10.8622	0.7680	0.7702
		H-10	10.0841	0.8040	0.7898
		L-10	11.6195	0.7284	0.7339
		L-75	11.2807	0.7467	0.7549
	Res4a_ ReLU	F	9.1073	0.8436	0.8611
		H-5	9.2559	0.8379	0.8509
		L-5	10.1132	0.8028	0.8072
		L-75	9.3484	0.8344	0.8518

References

1. Barsoum, E., Zhang, C., Canton-Ferrer, C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. ArXiv **abs/1608.01041** (2016)
2. Jayaraman, D., Mittal, A., Moorthy, A.K., Bovik, A.C.: Objective quality assessment of multiply distorted images. 2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR) pp. 1693–1697 (2012)
3. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 1646–1654 (2016)
4. Mannos, J., Sakrison, D.: The effects of a visual fidelity criterion of the encoding of images. IEEE Transactions on Information Theory **20**(4), 525–536 (July 1974). https://doi.org/10.1109/TIT.1974.1055250
5. Milford, P.: Eye tracked lens for increased screen resolution. In: US Patent no. 20200132990 (2020)
6. Pedoeem, J., Huang, R.: Yolo-lite: A real-time object detection algorithm optimized for non-gpu computers. 2018 IEEE International Conference on Big Data (Big Data) pp. 2503–2510 (2018)
7. Sheikh, H.R., Sabir, M.F., Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Transactions on Image Processing **15**, 3440–3451 (2006)