WHY DOES PRUNING WORK

INSIGHTS FROM AI MODELING: HU, H., PENG, R., TAI, Y. W., & TANG, C. K. (2016). NETWORK TRIMMING: A DATA-DRIVEN NEURON PRUNING APPROACH TOWARDS EFFICIENT DEEP ARCHITECTURES. ARXIV PREPRINT ARXIV:1607.03250.

SOME DNN NODES PROVIDE LITTLE INFORMATION

Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures

Hengyuan Hu *
HKUST
hhuaa@ust.hk

Rui Peng *
HKUST
rpeng@ust.hk

Yu-Wing Tai
SenseTime Group Limited
yuwing@sensetime.com

Chi-Keung Tang
HKUST
cktang@cse.ust.hk

Although CNNs with elegant network architectures are easy to deploy in real-world tasks, designing one can be hard and labor-intensive, which involves significant amount of effort in empirical experiments. In terms of designing the network architecture, one crucial part is to determine the number of neurons in each layer. There is no way to directly arrive at an optimal number of neurons for each layer and thus even the most successful network architectures use empirical numbers like 128, 512, 4096. Experienced scientists often arrive at the numbers once they deem the network have enough representation power for the specific task. However, the extremely sparse matrices produced by top layers of neural networks have caught our attention, indicating that empirically designed networks are heavily oversized. After some simple statistics, we find that many neurons in a CNN have very low activiations no matter what data is presented. Such weak neurons are highly likely to be redundant and can be excluded without damaging the overall performance. Their existence can only increase the chance of overfitting and optimization difficulty, both of which are harmful to the network.

PRINCIPLE STAT: AVERAGE PERCENTAGE OF '0'S

- They define Average Percentage of Zeros (APoZ) of a single neuron as percentage of zero activations of that neuron after the ReLU mapping.
- *Layer* is the unit of analysis rather than single neuron, so the analysis is collapsed across all neurons in a layer
- For a given dataset with N images APoZ just is the percentage of 0s after RELU (computed as proportion of 0s/N)
  - (if a neuron contributes to more than on feature map (M>1), the computer number of 0s out of NxM)

PRINCIPLE STAT: AVERAGE PERCENTAGE OF '0'S

$APoZ_c^{(i)}$ of the c-th neuron in i-th layer is defined as:

$$APoZ_c^{(i)} = APoZ(O_c^{(i)}) = \frac{\sum_k^N \sum_j^M f(O_{c,j}^{(i)}(k) = 0)}{N \times M} \quad (1)$$

where $f(\cdot) = 1$ if true, and $f(\cdot) = 0$ if false, M denotes the dimension of output feature map of $O_c^{(i)}$, and N denotes the total number of validation examples. The larger number of validation examples, the more accurate is the measurement of APoZ. In our experiment, we use the validation set ($N = 50,000$) of ImageNet classification task to measure APoZ.

RESULTS

- Redundancy is mainly in the deeper convlayers and the fully connected layers
- They find >600 neurons with APOZ > 90%

A ‘ZOOM IN’ HISTOGRAM ON APOZ DISTRIBUTION IN FC-6

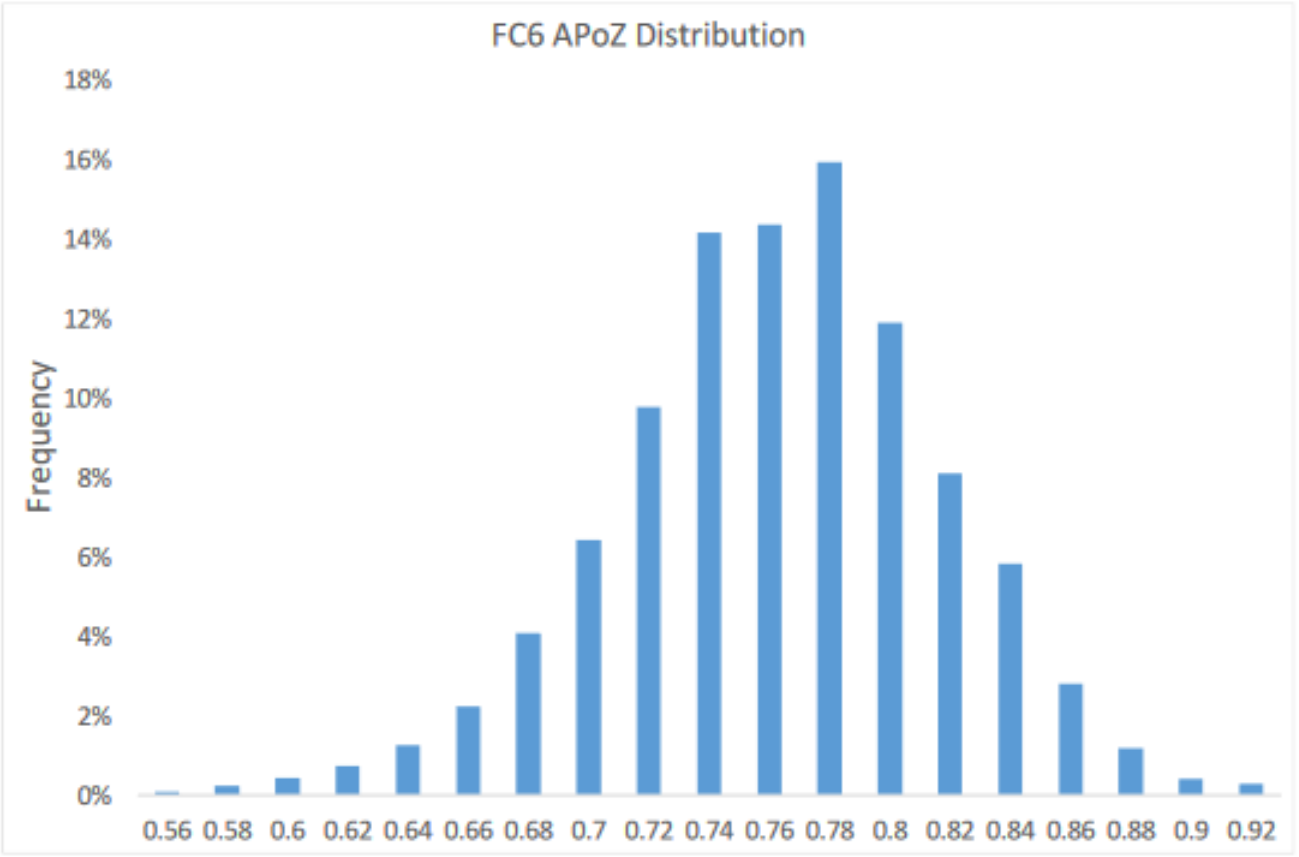


Figure 2: FC6 APoZ Distribution

ARE HIGH-APOZ NEURONS REDUNDANT?

- They implement a neuron-pruning approach
- After training to criteria, the remove all weights to and from high-APOZ nodes (i.e., they remove these nodes from the net)
- They then re-init the network with the last set of weights (prior to pruning) and retrain to criteria.

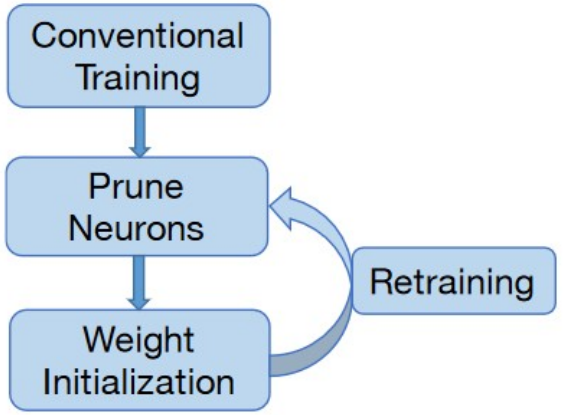


Figure 3: Three main steps for trimming

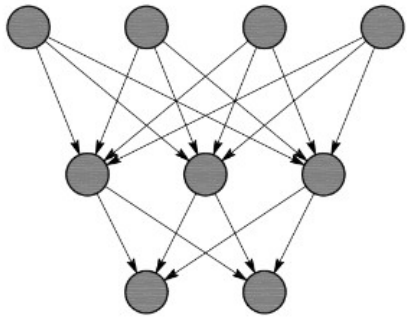


Figure 4: Before pruning

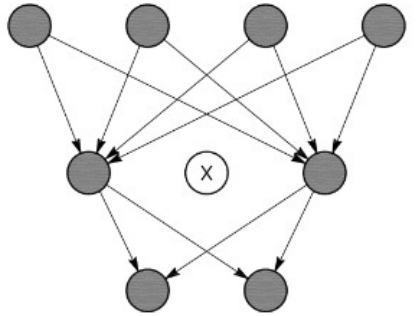


Figure 5: After pruning

Table 4: Iterative Trimming Result on VGG-16 {CONV5-3, FC6}

Network (CONV5-3, FC6)	Compression Rate	Before Fine-tuning (%)		After Fine-tuning (%)	
		Top-1 Accuracy	Top-5 Accuracy	Top-1 Accuracy	Top-5 Accuracy
(512, 4096)	1.00	68.36	88.44	68.36	88.44
(488, 3477)	1.19	64.09	85.90	71.17	90.28
(451, 2937)	1.45	66.77	87.57	71.08	90.44
(430, 2479)	1.71	68.67	89.17	71.06	90.34
(420, 2121)	1.96	69.53	89.49	71.05	90.30
(400, 1787)	2.28	68.58	88.92	70.64	89.97
(390, 1513)	2.59	69.29	89.07	70.44	89.79

REDUNDANCY AND REPRESENTATIONAL GEOMETRY

TRUONG AND HASSON, IN PREPARATION

LOW-INFORMATION FEATURES AND THEIR IMPACT ON REPRESENTATIONAL GEOMETRY

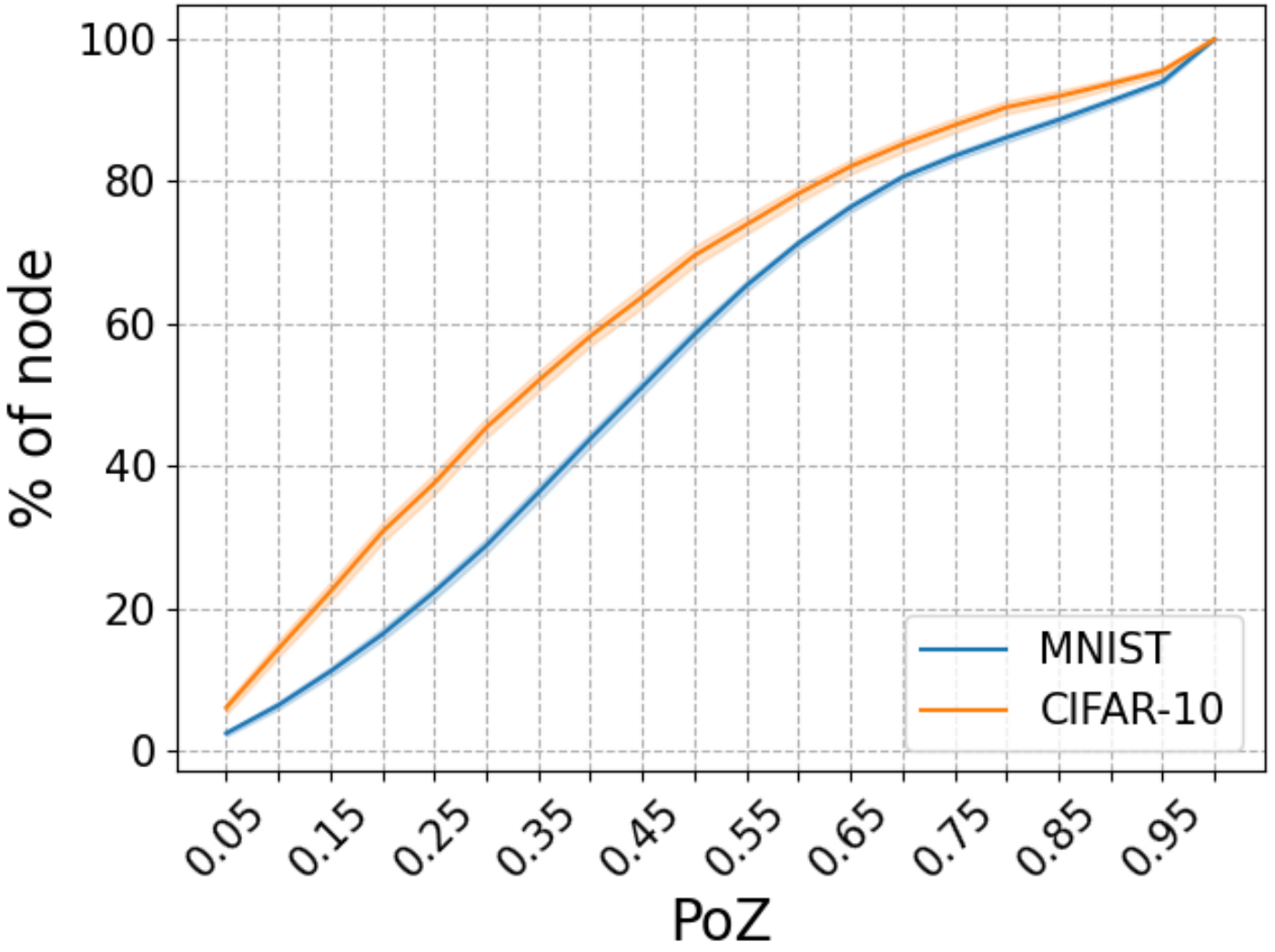
- For a given dataset, a 100% PoZ feature does not provide discriminating information between objects.
  - It may serve to separate objects in this dataset from others
- However, these features do contribute to pair-wise similarity estimations; i.e., estimation of object-similarity (cosine, Pearson)
- QUESTIONS:
  - How do these features contribute to a DNN's RDM?
  - Can their removal improve prediction of human similarity judgments

IMPACT OF ALL-ZERO FEATURE ON VECTOR CORRELATIONS

- hypothetical embeddings of three images (A, B, C); Node #6 is 0 for all images.
- When computed from the full embeddings, the pairwise Pearson similarity values between the three images (A, B; B, C; A, C) are
  0.62, 0.56, −0.18.
  When feature #6 is removed, the values become
  −0.06, 0.40, −0.94.
- Full beddings, Sim(A, B) > Sim(B,C)
  but when the 0 feature is excluded,
  Sim(A, B) < Sim(B,C)

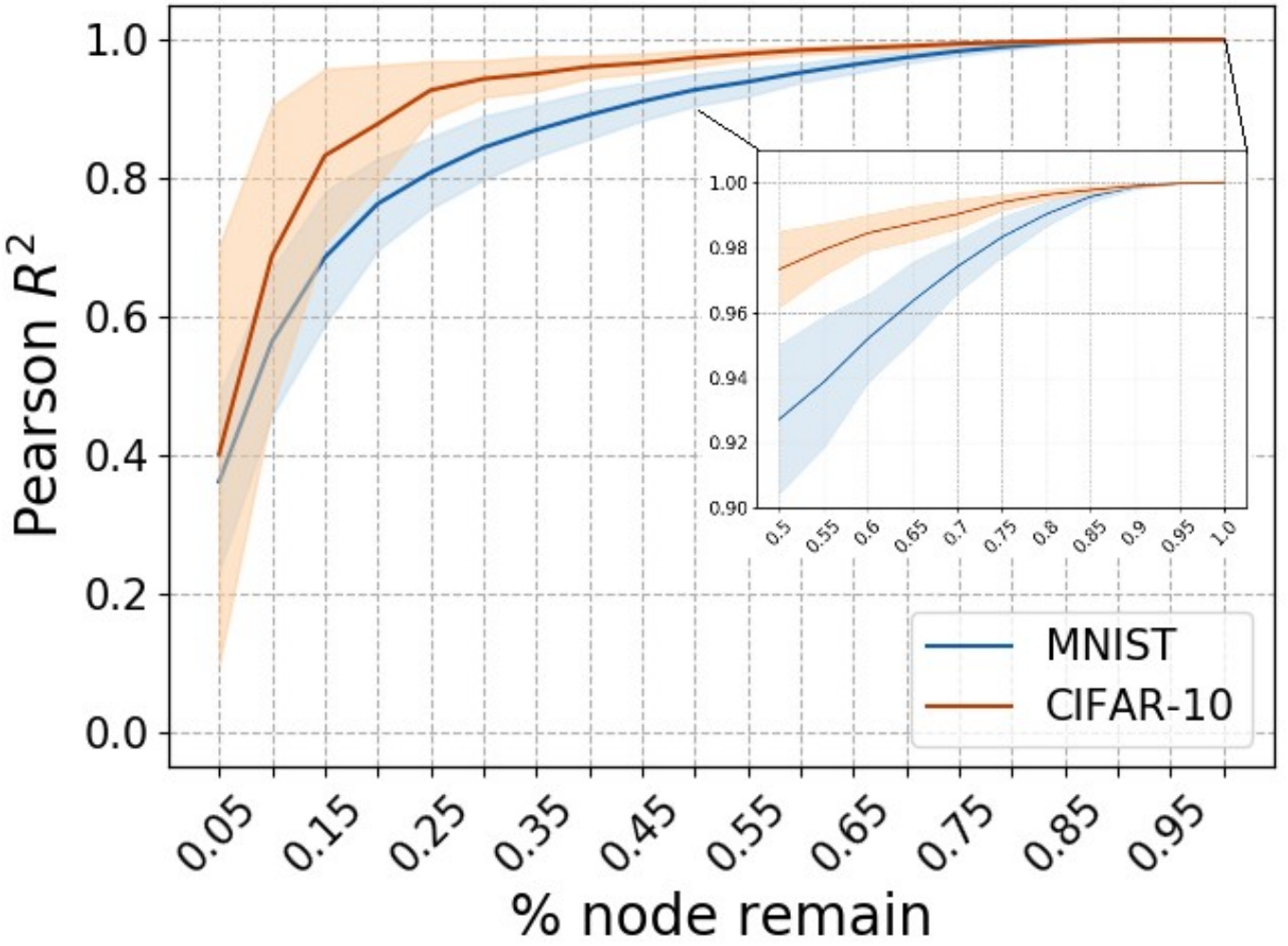
A :	0.91	0.76	0.3	0.7	0.9	0
B :	0.4	0.7	0.6	0.3	0.7	0
C :	0.02	0.4	0.9	0.2	0.2	0

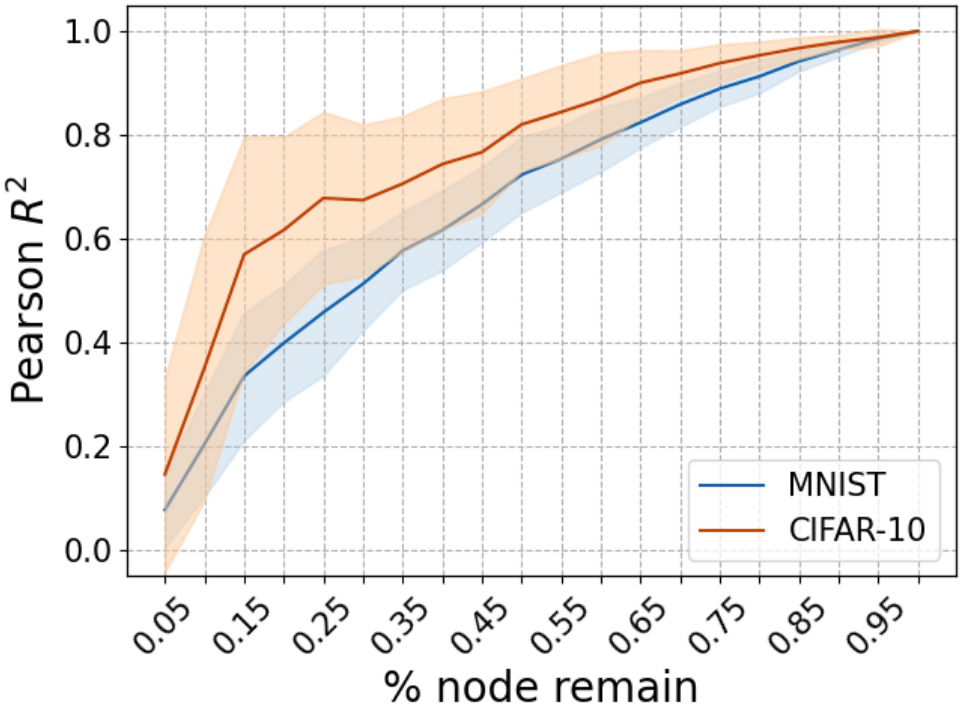
DISTRIBUTION OF POZ IN MNIST AND CIFAR

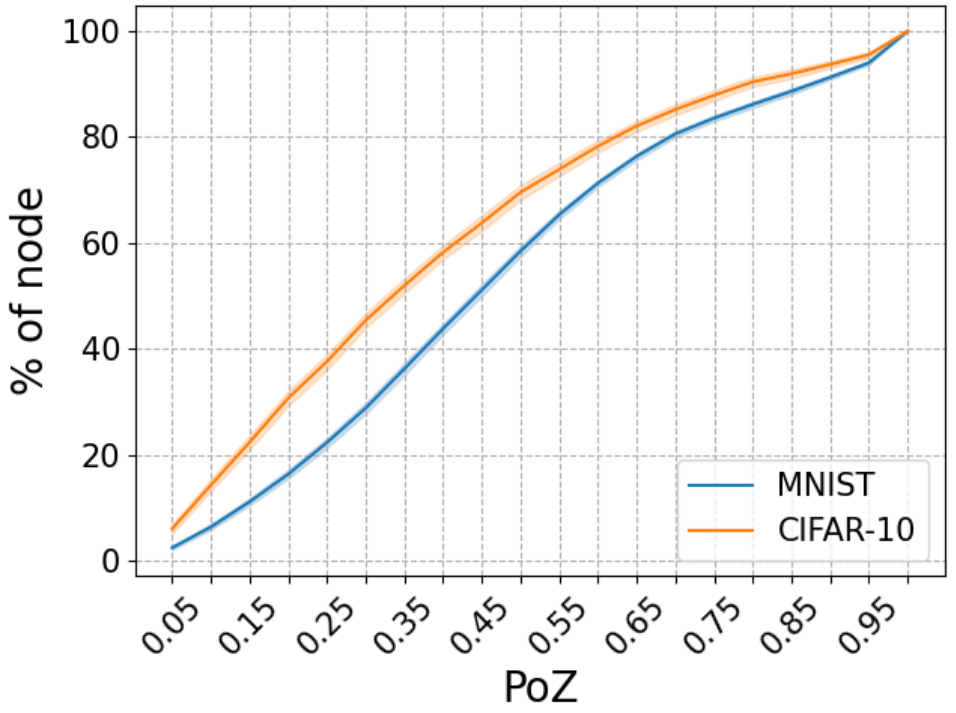


INSERTING FEATURES FROM LOW-TO-HIGH POZ AND COMPUTING 2OI VS. FULL EMBEDDING RDM

- Features inserted sequentially from low-to-high POZ.
- Each time, an RDM from partial set is computed and compared (R2) vs. the RDM of the original embeddings.
- For CIFAR-19 high R2 of 0.9 arrived with only 25% of features. R2=0.98 after 50% of features inserted.
- Do the high-PoZ features (the remaining 50%) contain any useful information at all?







INSERTING FEATURES FROM HIGH-TO-LOW POZ

- As previous slide, in reverse direction
- The lowest 20-% already produce R2>0.6 for CIFAR10. PoZ for these is 55% and above.
- Conclusion: High-Poz Contain redundant information, distributed across these features.

- Evaluation of APOZ-based pruning and impact on prediction of human similarity judgments.
- Features remove sequentially from high-to-low PoZ and 2OI computed against Human Similarity Judgments.
- Tick lines mark bins of PoZ level.