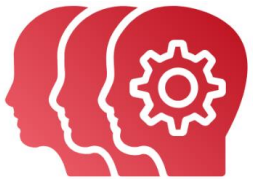


Visualizing the Hidden Activity of Artificial Neural Networks

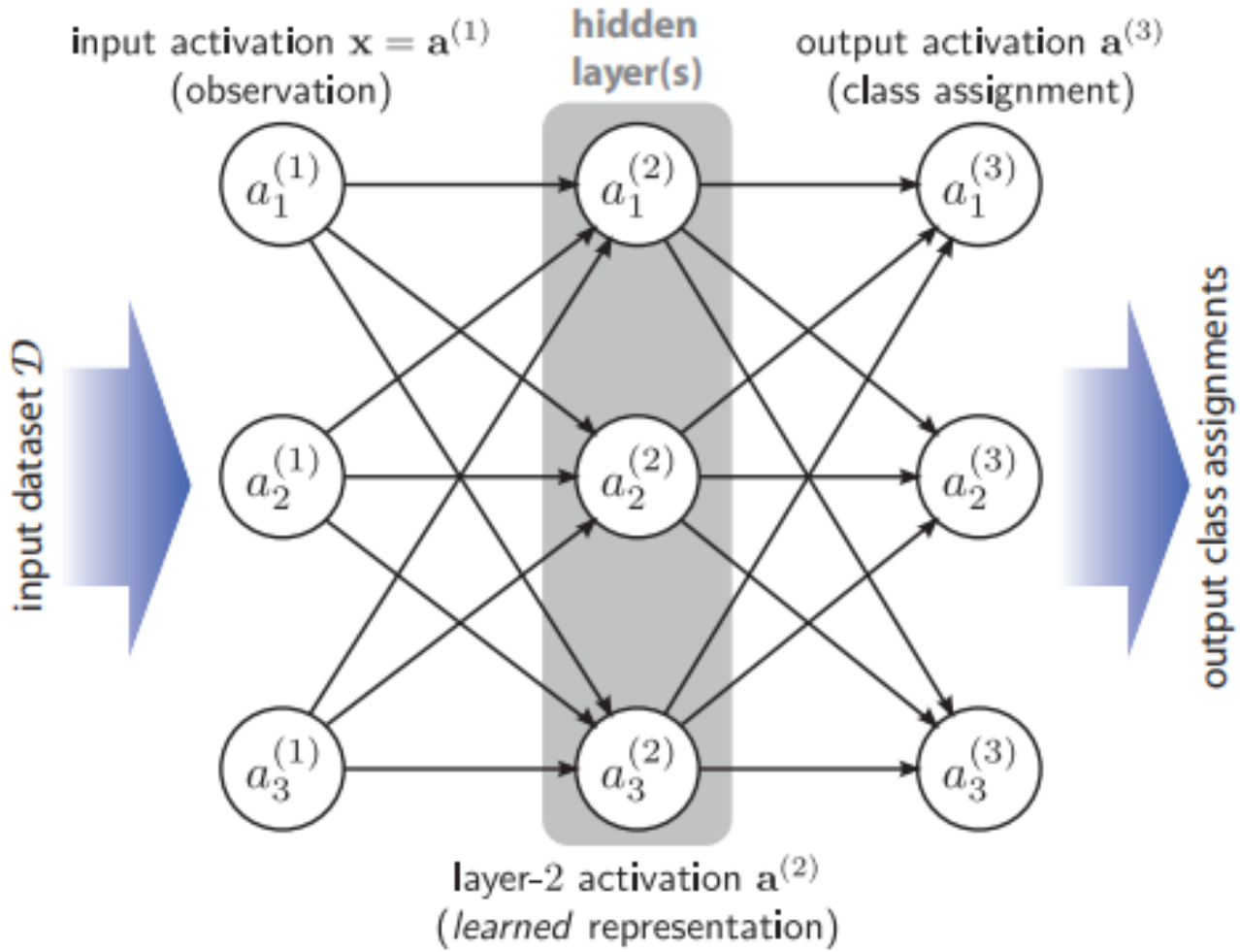
高久怡

2019/09/29

Two Task -- dimensionality reduction：

- visualizing the relationships between *learned representations* of observations

- visualizing the relationships between *artificial neurons*



alternative representation of input data

Datasets： MNIST SVHN（谷歌街景门牌号码） CIFAR-10

Neural Networks :

- MLP : four rectified linear hidden layers of 1000 neurons, The output layer is softmax with 10 neurons.

- CNN : conv(2)→maxpool→conv(2)→maxpool→FC(4096)→FC(512)→softmax(10),all relu activate

Activations：CNN中只提取FC层的表达

Projections: t-SNE MDS

NH（neighborhood hit）：measure projection quality——class separation

For a given k (in our work, $k = 6$), the NH for a point $\mathbf{a}_p$ is the ratio of its k-nearest neighbors that belong to the same class as $\mathbf{a}_p$.

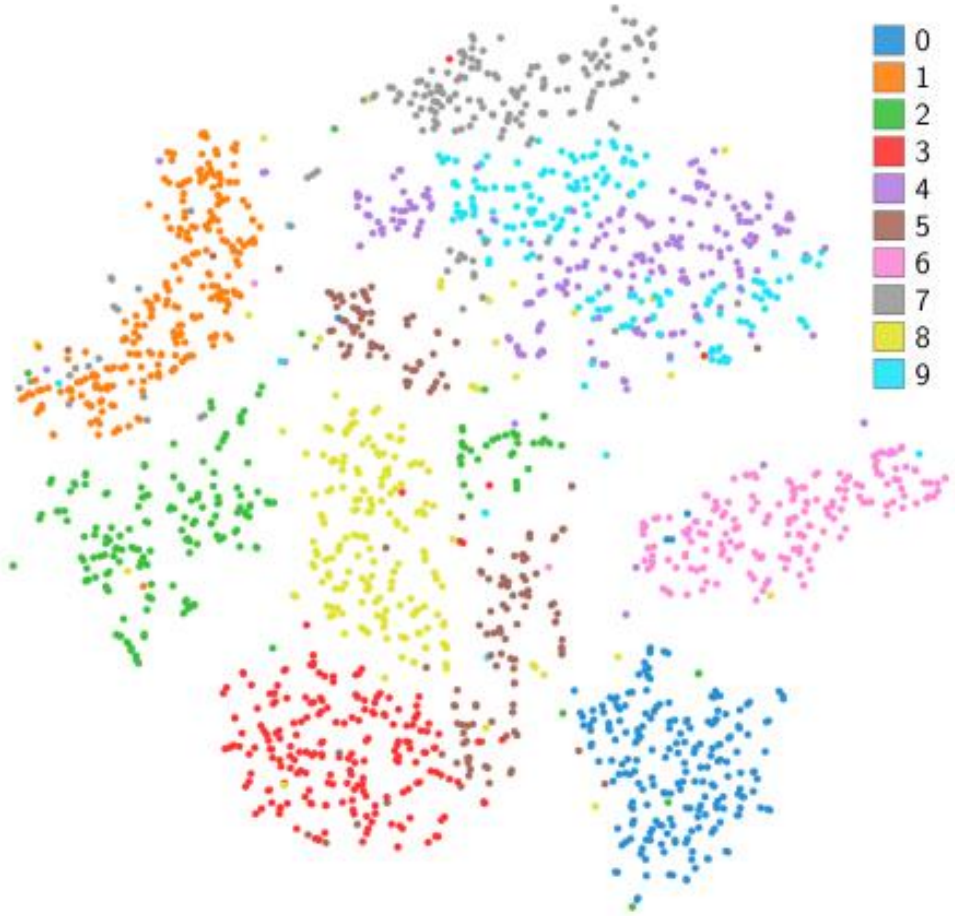
NH（neighborhood hit）： measure projection quality——class separation

k=6，a target point $\mathbf{a}_p$

$$NH(\mathbf{a}_p) = \frac{\text{在}k\text{中与}\mathbf{a}_p\text{同类的点的数量}}{k}$$

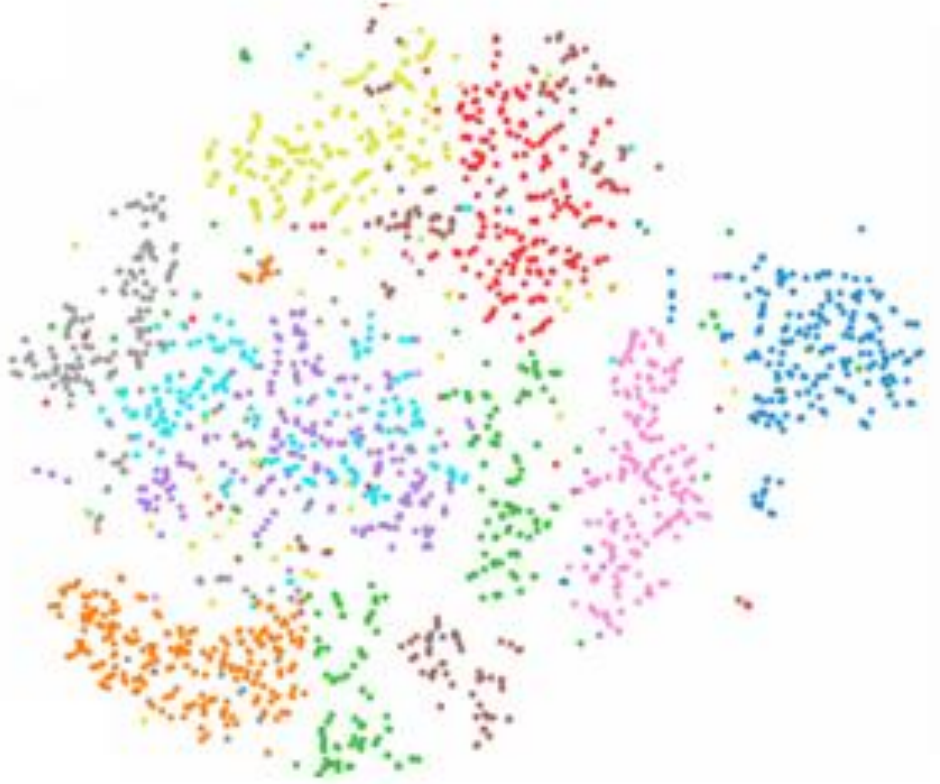
NH(a whole projection) = 所有点的NH的平均

H1：随机初始化且没训练过的网络，数据的最后一层隐藏层的表达的类别分隔度较差



Projection of observations, *MNIST* test subset (NH: 89.12%).

label、input（784）



Projection of the last MLP hidden layer activations Before training (NH: 83.78%)

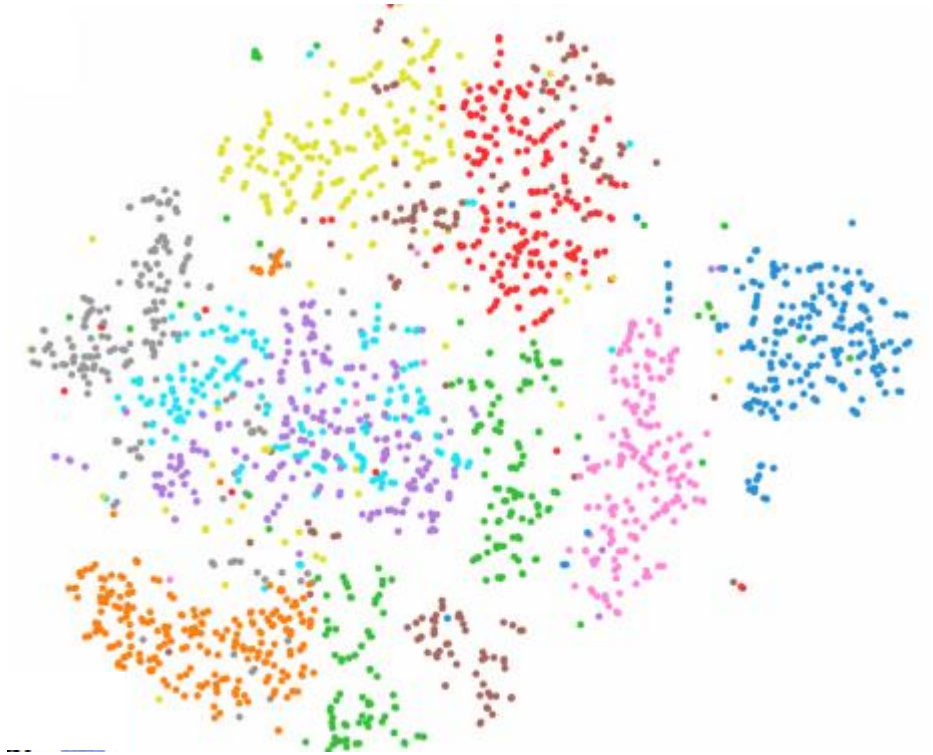
pred、activation（1000）

H1：随机初始化且没训练过的网络，数据的最后一层隐藏层的表达的类别分隔度较差

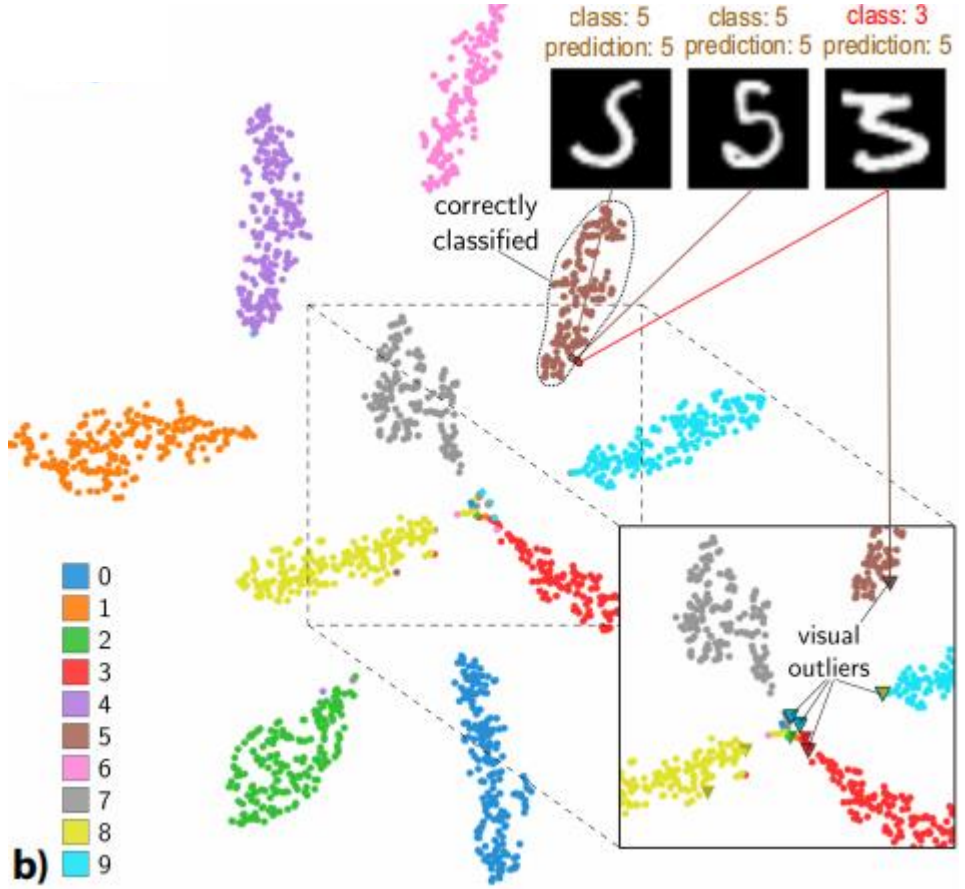
R1：与假设矛盾，有相似的class separation

C1：在网络训练之前，最后一层的隐藏层的表达就有与类别相关的比较清晰的结构

H2：经过训练之后，深层（last）表达的class separation会提升



training



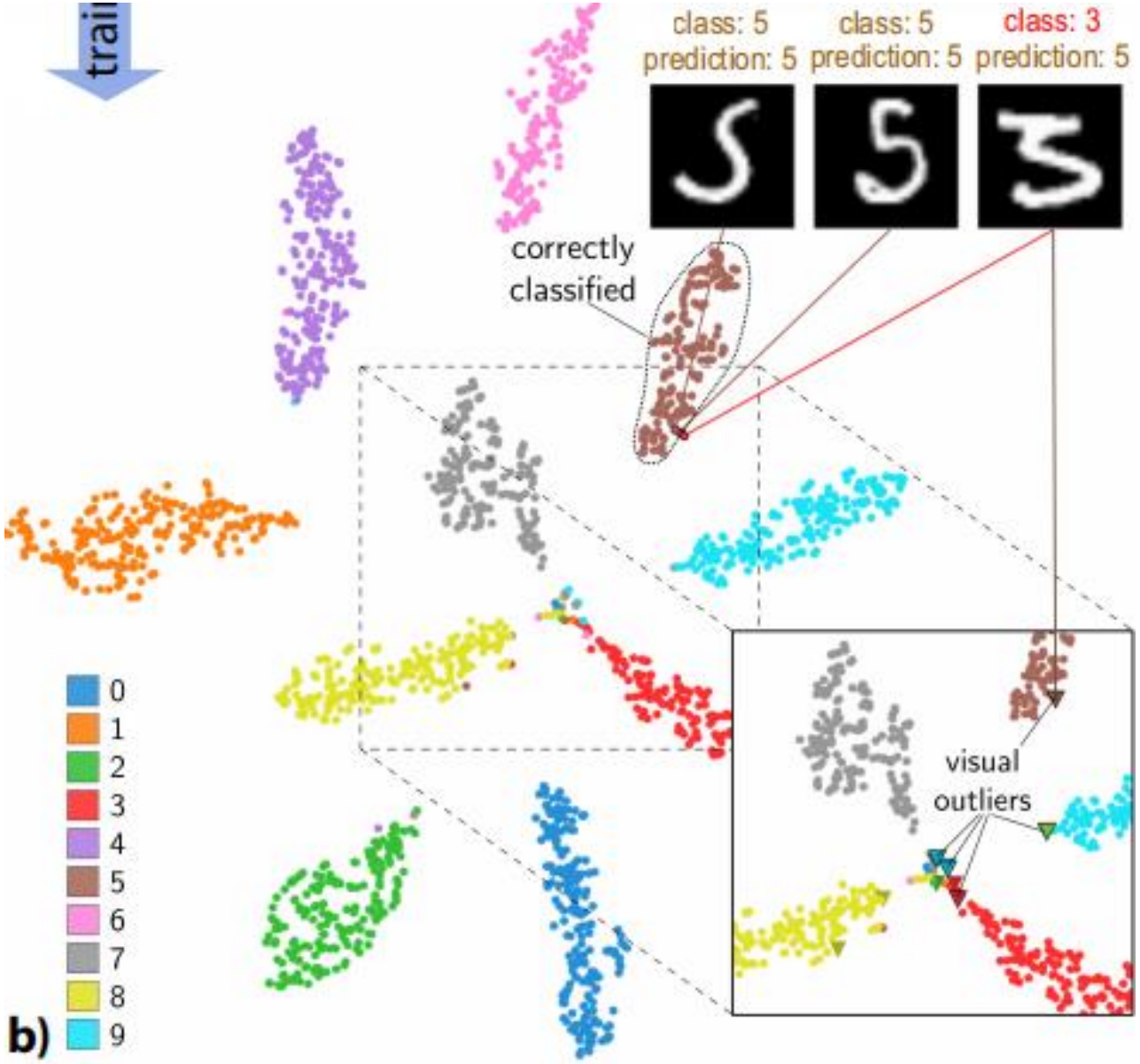
Before training (NH: 83.78%).

After training (NH: 98.36%,AC: 99.15%)

H2：经过训练之后，深层（last）表达的class separation会提升

R2：与假设一致

C2：网络在学习过程中肯定学到了可以抓住类别结构的数据的另一种表达



last layer activation

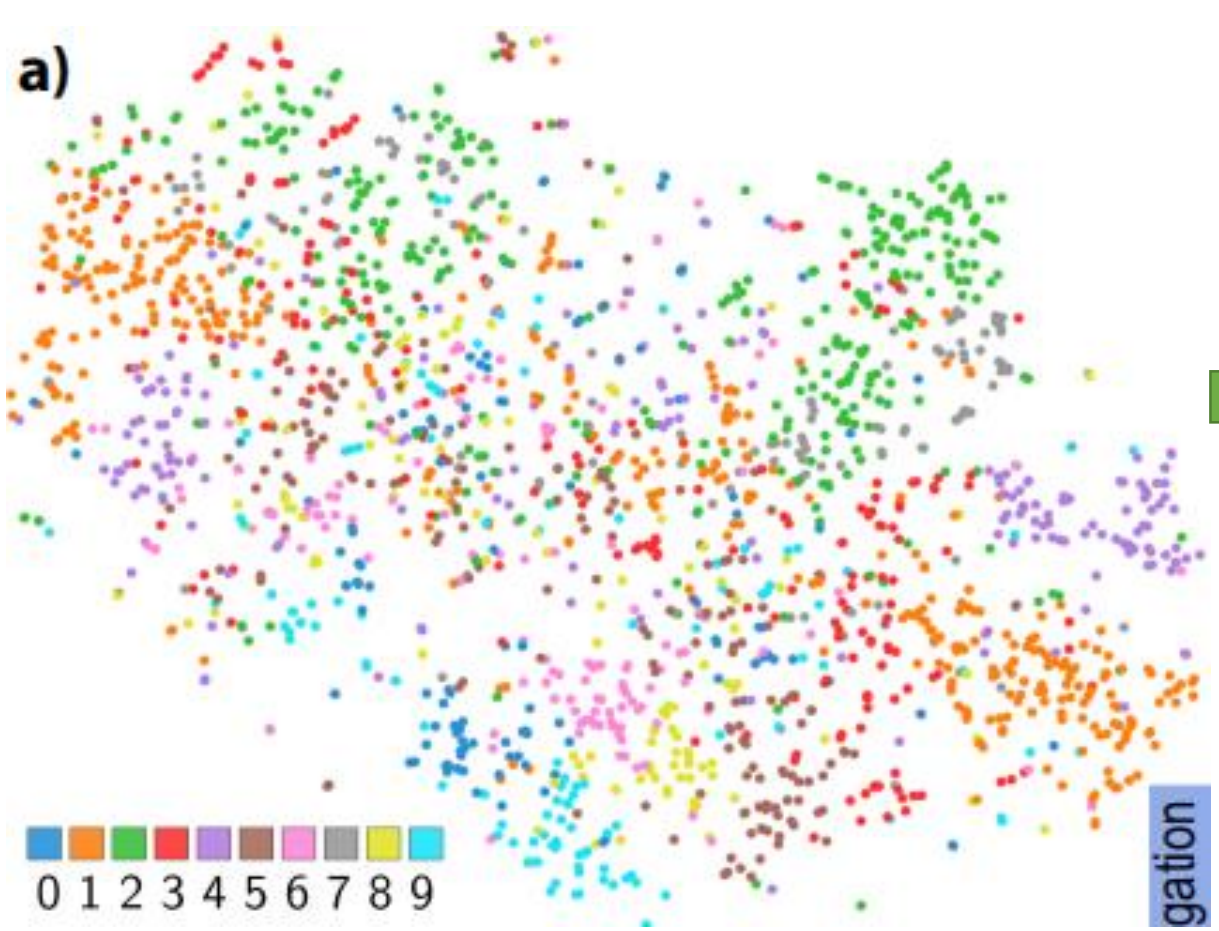
misclassification:隐藏层激活的相似性与最终的类别分配相关联

SVHN比MNIST更复杂，分类更困难

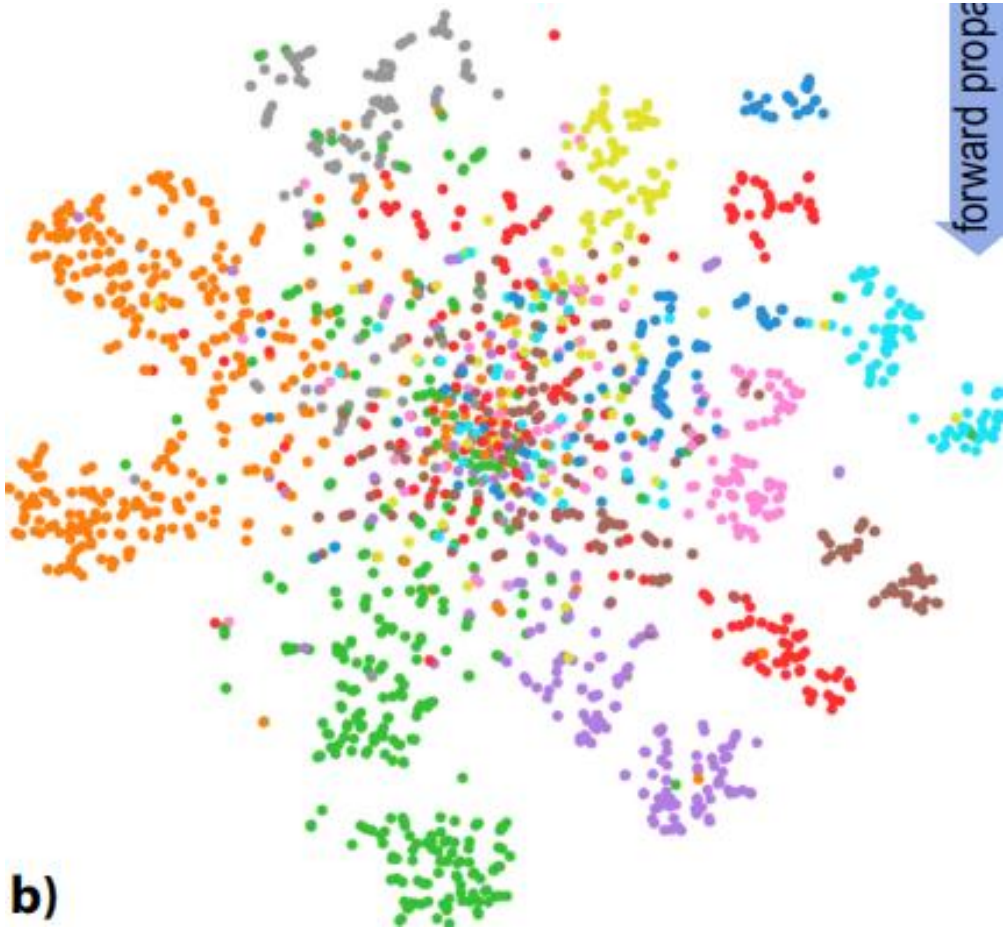
Fig. 4. Projection of the last MLP hidden layer activations before training, *SVHN* test subset (NH: 20.94%). Poor class separation is visible.

H3：训练好的NN的深层激活表达相比于浅层表达来说更具有类判别性

AMC
AI-ML.Club



FP



First hidden layer (NH: 52.78%).

Last hidden layer (NH: 67%).

H3：训练好的NN的深层激活表达相比于浅层表达来说更具有类判别性

R3：与假设一致

C3：网络深层更关注抽象具体的特征，网络浅层更关注泛化特征

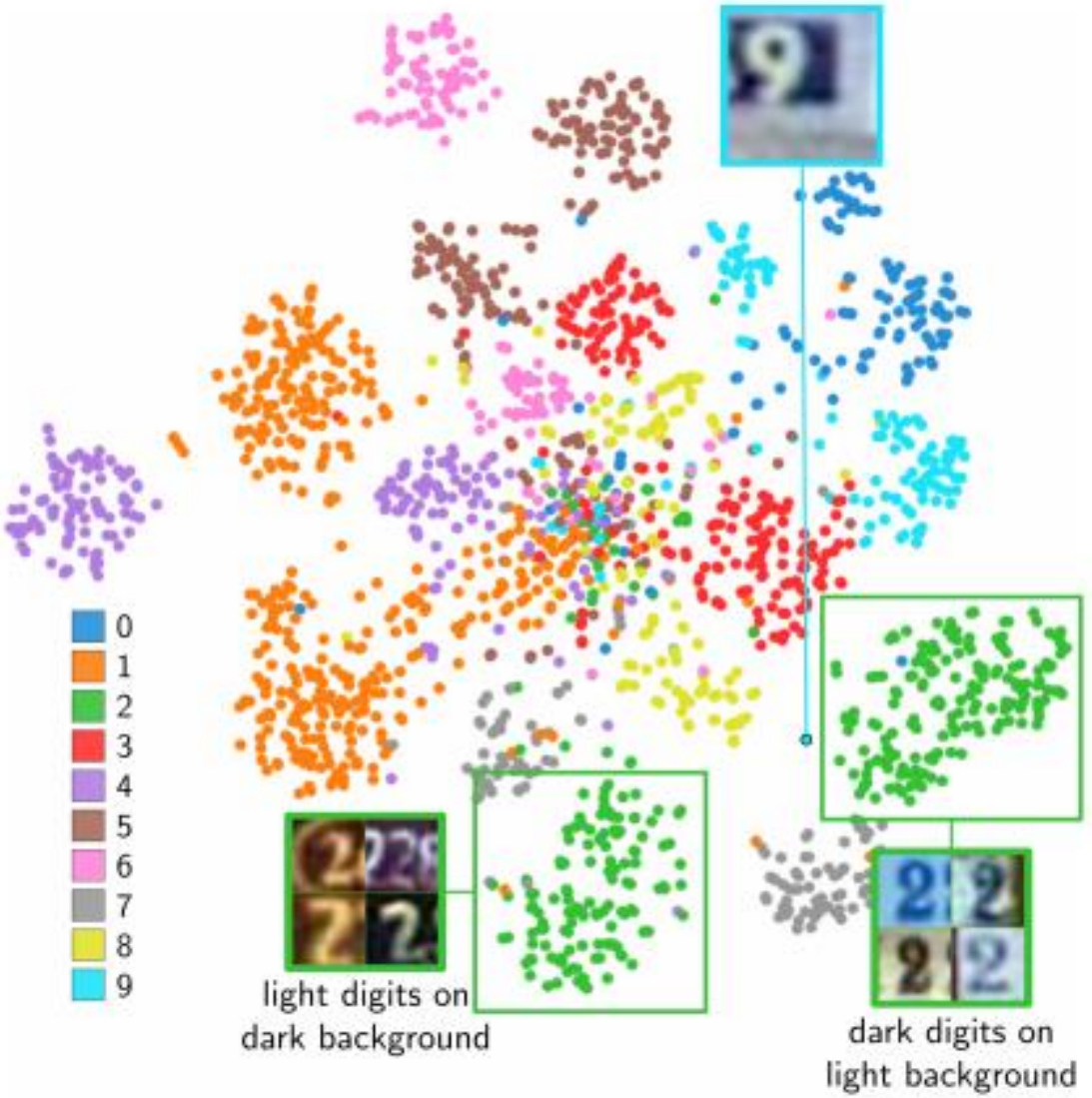


Fig. 6. Projection of the last CNN hidden layer activations after training, *SVHN* test subset (NH: 85.02%). Insets show example observations (images) from the visual clusters.

observe：每一个类别被划分为两个簇，说明对于每一个类别，都有两种不同的对于图片的内部表达

misclassification：9右边黑色的边缘可能被认为是2的变形

H4：在输入图片中移除明显不必要的变化（背景和前景）可以提升分类的准确性

propose：Sobel operator 、Gaussian blur ，产生了背景与前景边缘明亮的灰度图像，避免模型检测数字是亮或暗，对于图片的高度易变性是很重要的

R3：与假设一致，预处理图像产生的投影显示每个类别只有一个簇，说明对于每张图片都只有一种表达

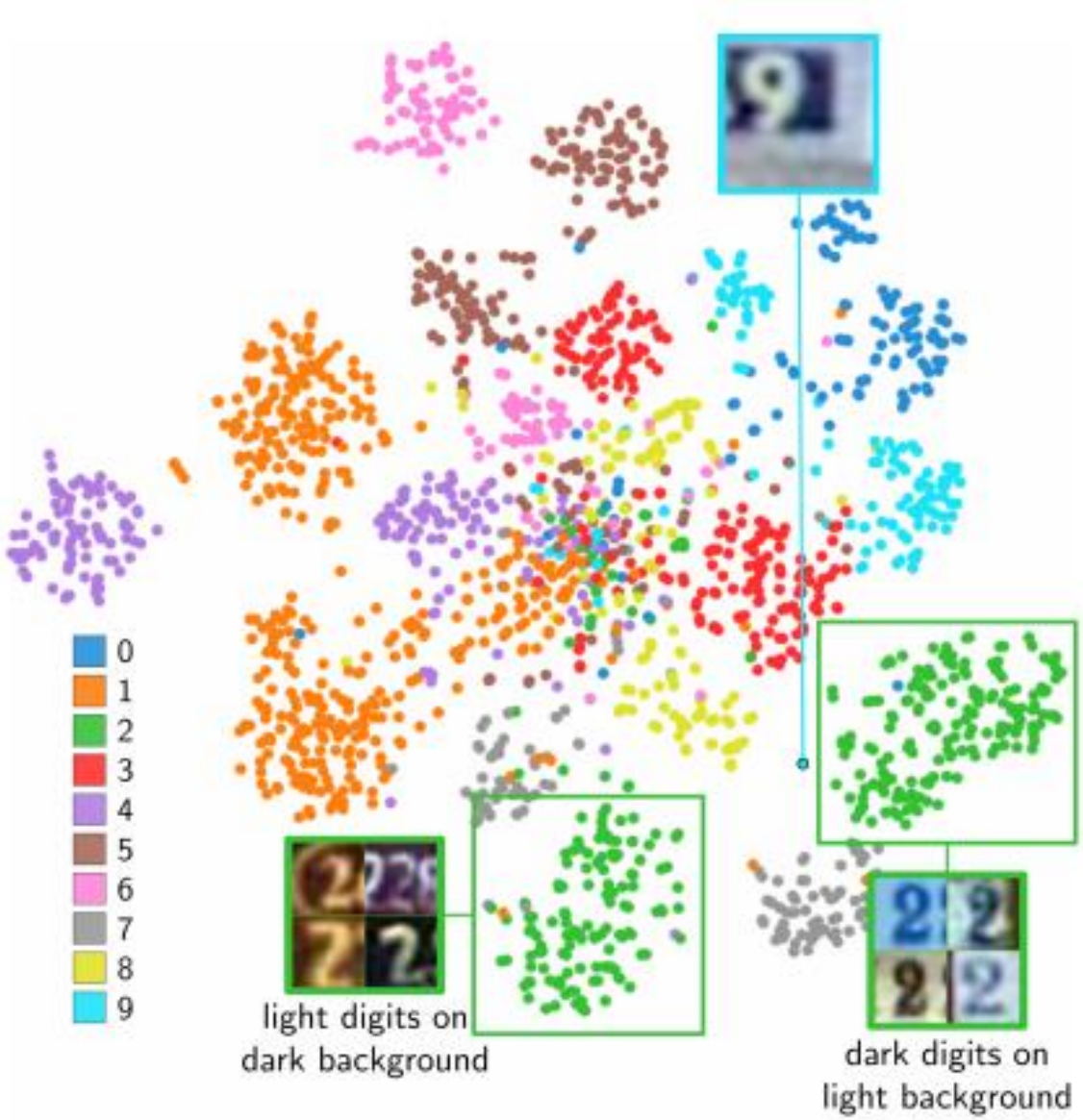


Fig. 6. Projection of the last CNN hidden layer activations after training, *SVHN* test subset (NH: 85.02%). Insets show example observations (images) from the visual clusters.

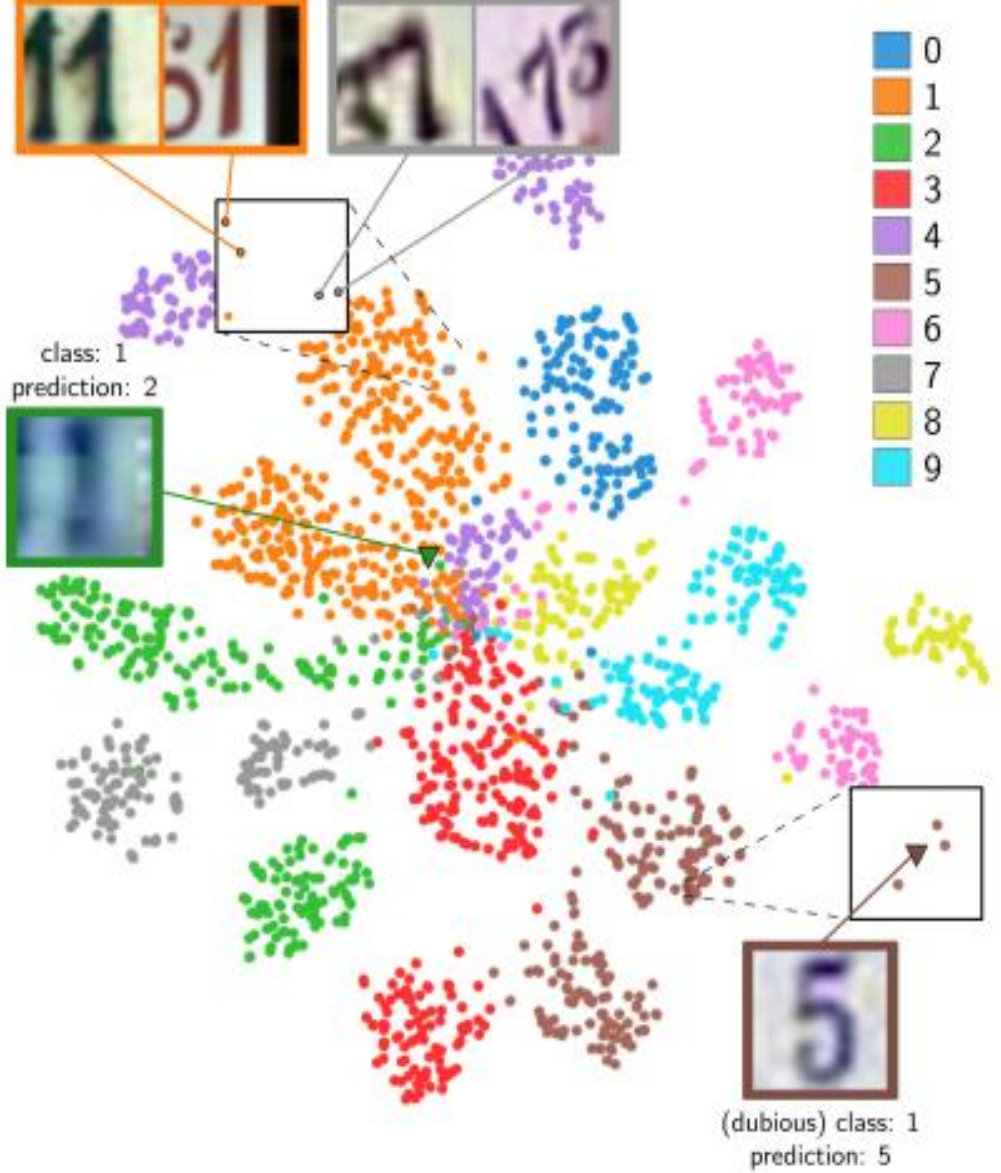


Fig. 8. Projection of last CNN hidden layer activations after training, *SVHN* training subset (NH: 93.83%, AC: 99.9%).

H4：训练集比测试集的类别分隔度更好

R3：与假设一致

potential overfitting：7预测正确，但是在1的簇的旁边，这些7与1比较类似，CNN在最后一层对于图片的内部表达overfit

Evolution of learned representation：**inter-epoch and inter-layer**

- inter-epoch：**an instance， a layer， different epochs**
- inter-layer：**an instance， an epoch， different layers**

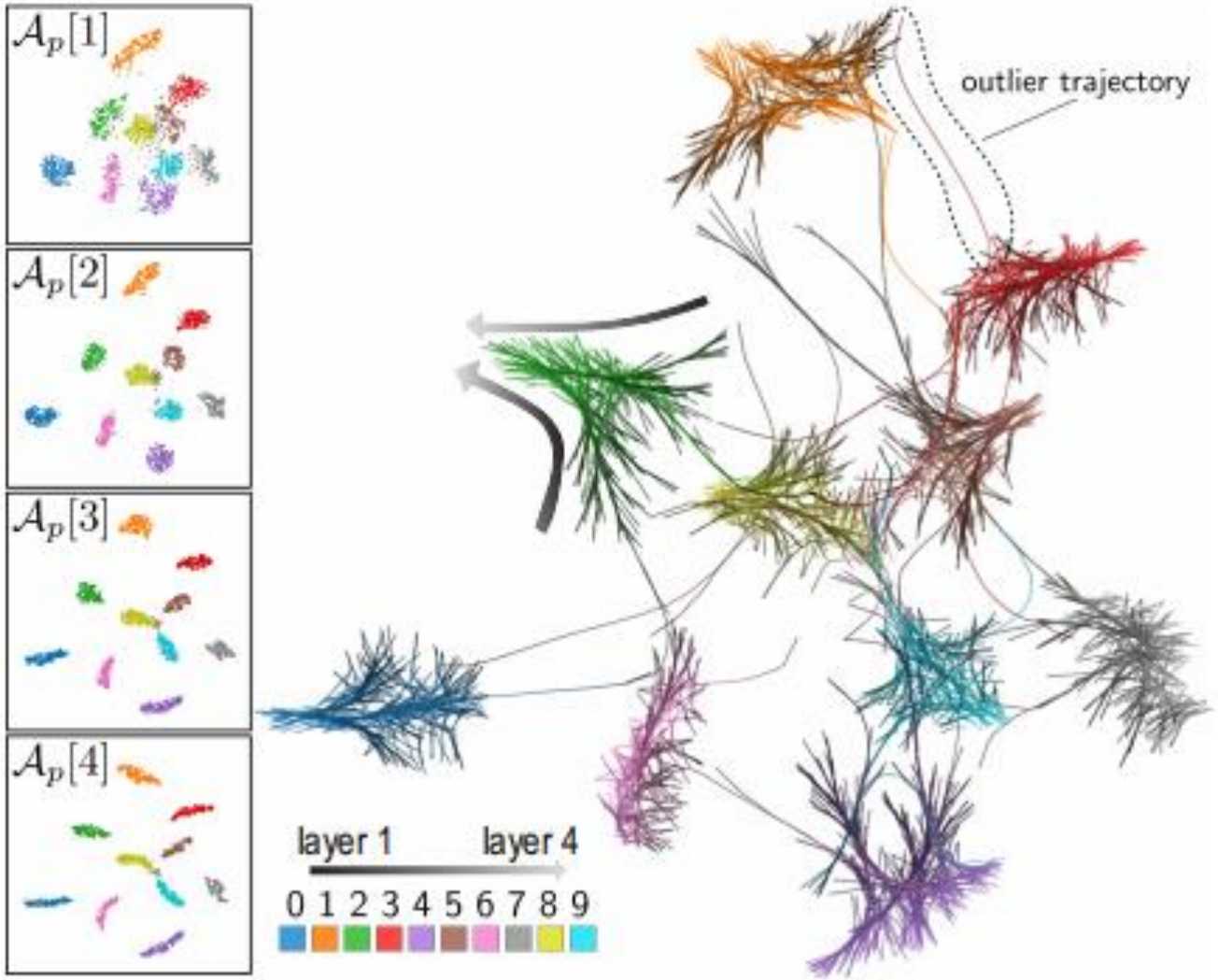


Fig. 10. Inter-layer evolution, four MLP hidden layers after training, *MNIST* test subset. Brighter trail parts show later layers.

Results：从浅层到深层，相同类别的距离越来越近，不同类别的距离越来越远，说明深层的表达含有更具有判别性的信息

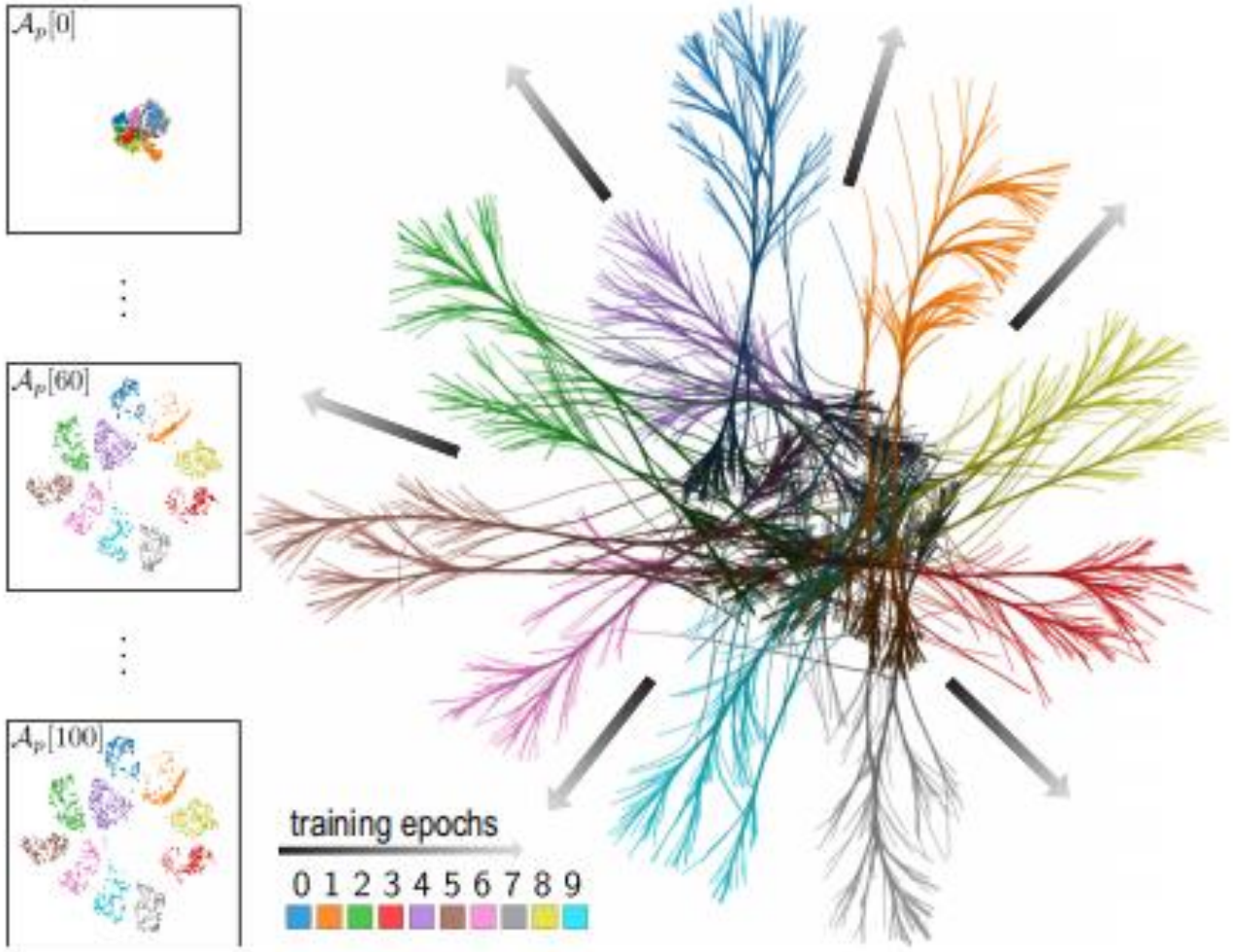


Fig. 11. Inter-epoch evolution, last CNN hidden layer, epochs 0-100, in steps of 20, *MNIST* test subset. Brighter trail parts show later epochs.

Results： 随着迭代次数的增加（网络在训练过程中学习到的参数的变化），网络的最后一层对于类别信息的抽取能力越来越强

Relations between neurons

- a given layer， 每个点代表一个神经元
- **MDS：尽可能的保留全局的两两神经元之间的距离关系**

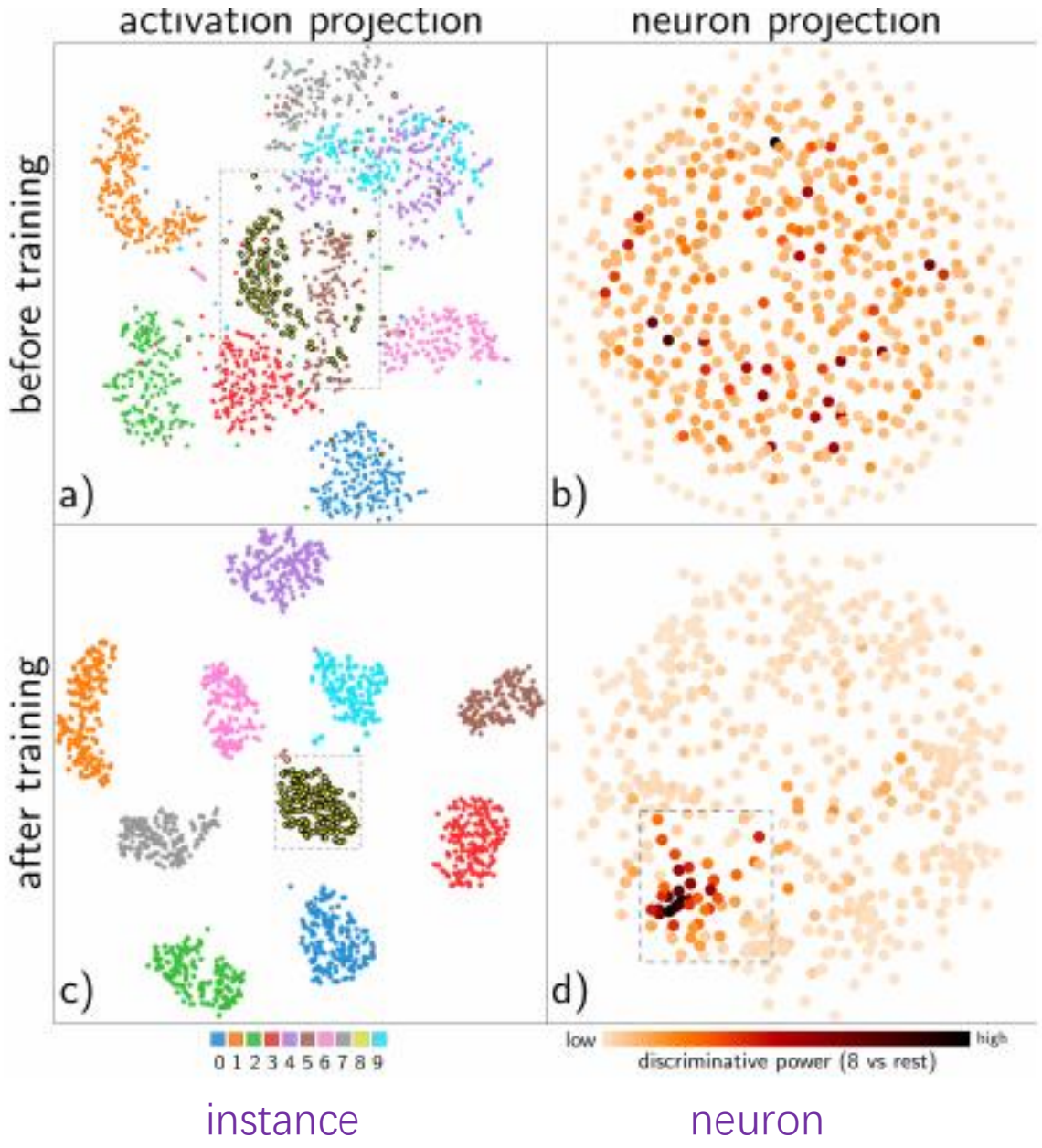


Fig. 12. Activation and neuron projections of last CNN hidden layer activations before and after training, *MNIST* test subset. Neuron projection colors show the neurons' power to discriminate class 8 *vs* rest.

AMC
AI-ML.Club

Results：经过训练之后，最后一层中对类别8判别性较强的神经元聚集在一起，表明训练创造了对类别高度相关的神经元集合

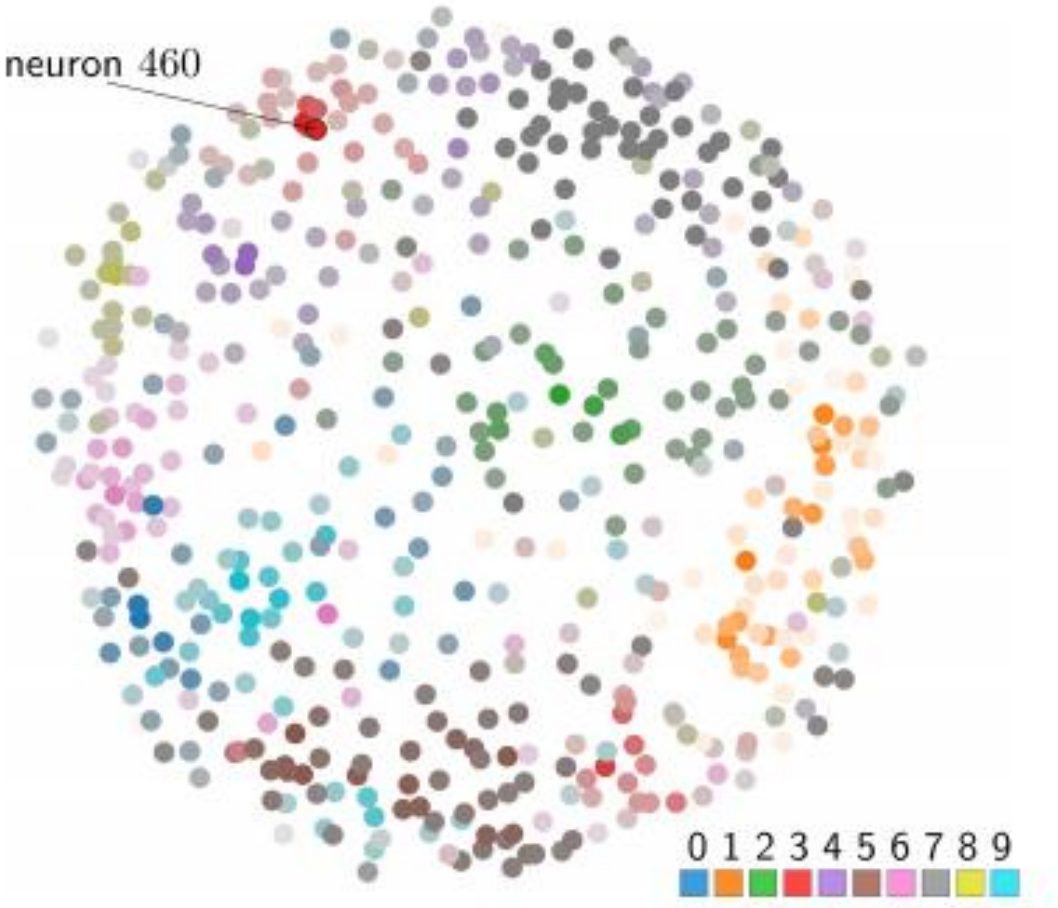


Fig. 13. Discriminative neuron map of last CNN hidden layer activations after training, *SVHN* test subset.

Discriminative neuron map：extremely randomized trees 生成判别分数 $S_{c,j}$

取 $s_{c^*,j} = \max s_{c,j}$ 作为每个神经元的类判别性分数，根据分数值大小决定颜色深浅

neuron 460与类别3高度相关

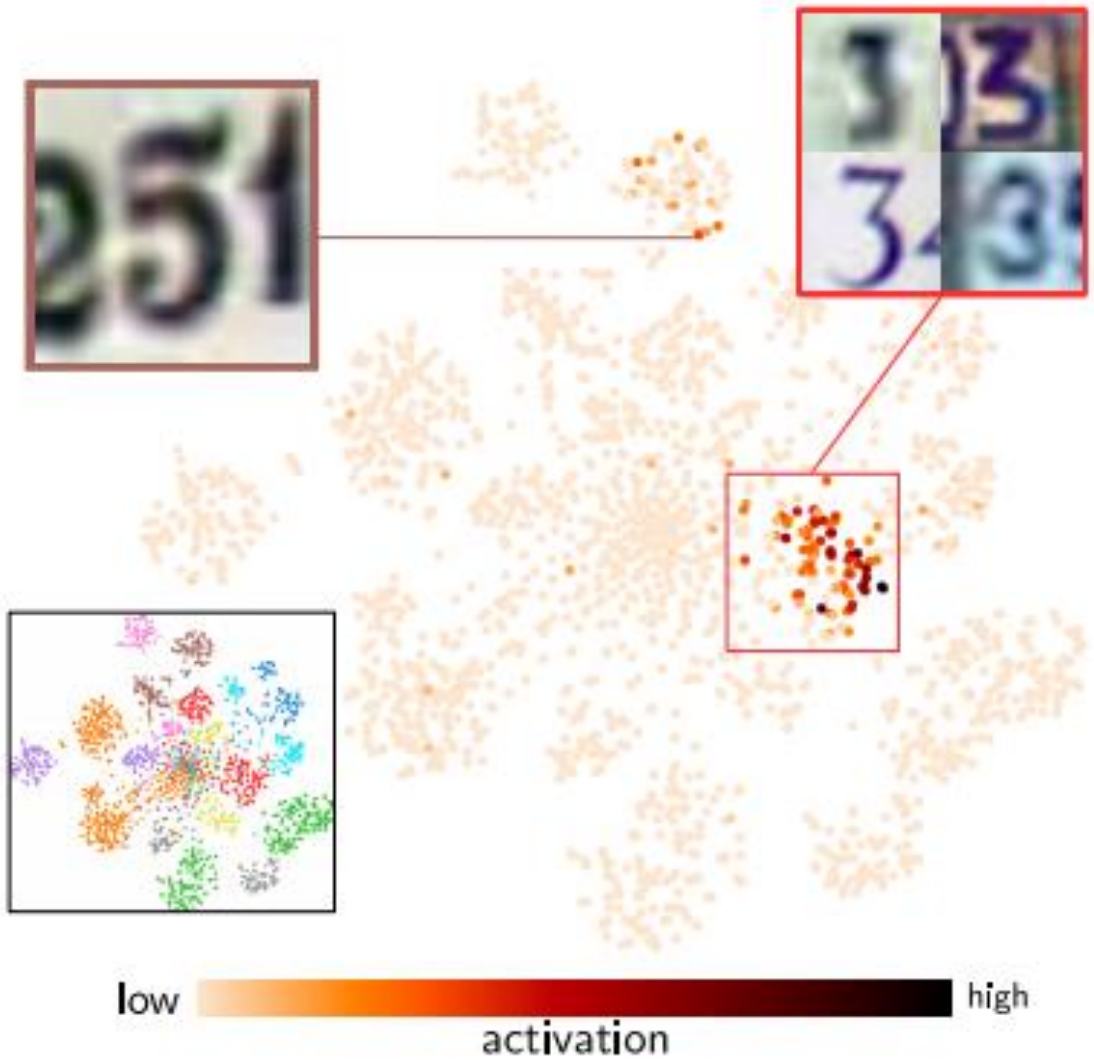


Fig. 14. Activation projection of the last CNN hidden layer after training, *SVHN* test subset. Color shows the activation of neuron 460, highly associated to class 3 (see also Fig. 13).

role of particular neurons：

所有instance中最后一层的neuron 460的激活值；方框区域是类别3的簇，此neuron 460的激活值都很高，说明neuron 460主要提取与类别3相关的信息

Thanks !