AMC
AI ML Club

Visualizing the Hidden Activity of Artificial Neural Networks

高久怡

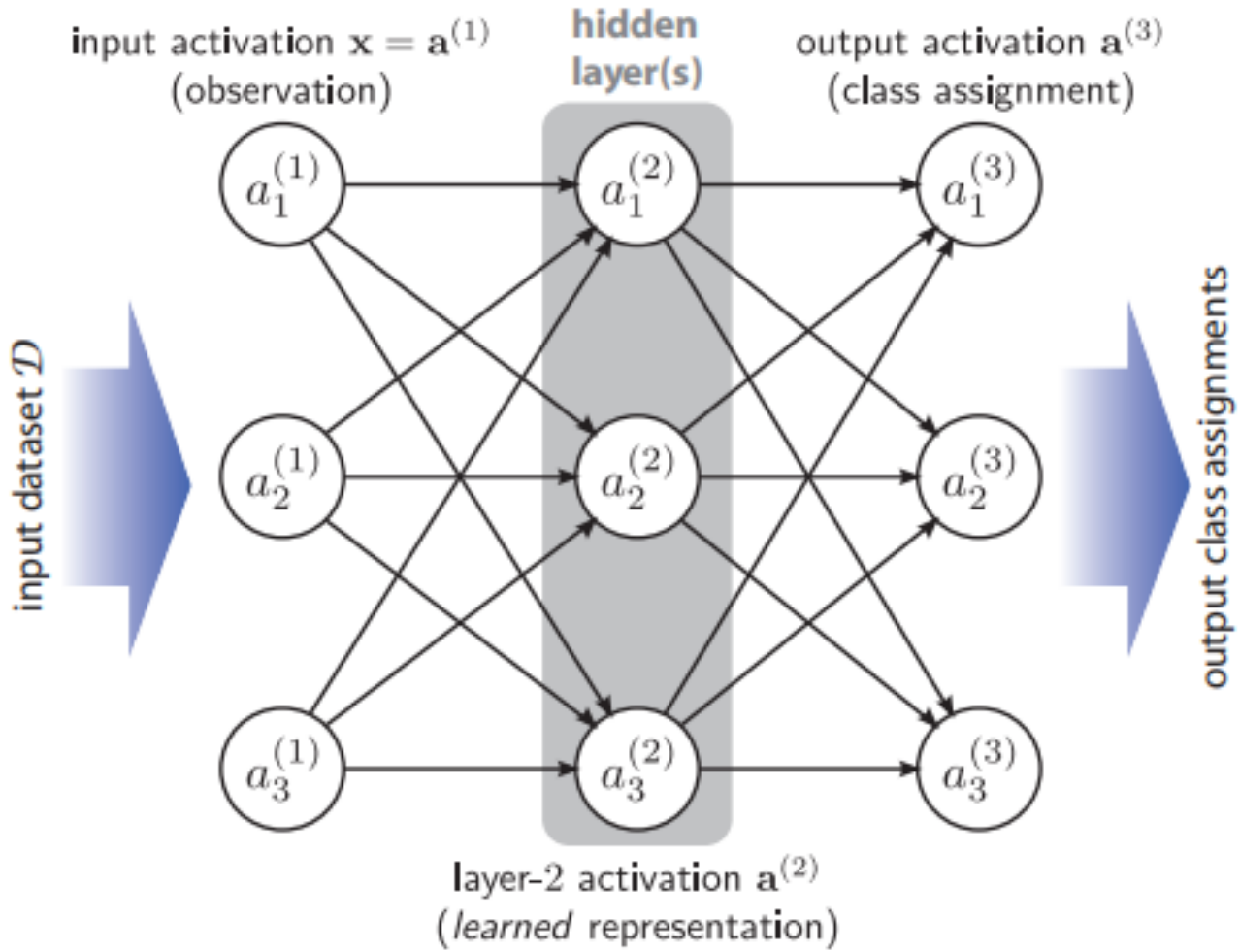
2019/09/29

AMC
AI ML Club

Two Task -- dimensionality reduction：

- visualizing the relationships between *learned representations* of observations

- visualizing the relationships between *artificial neurons*



alternative representation of input data

AMC
AI ML Club

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 $\mathbf{a}_p$.

AMC
AI ML Club

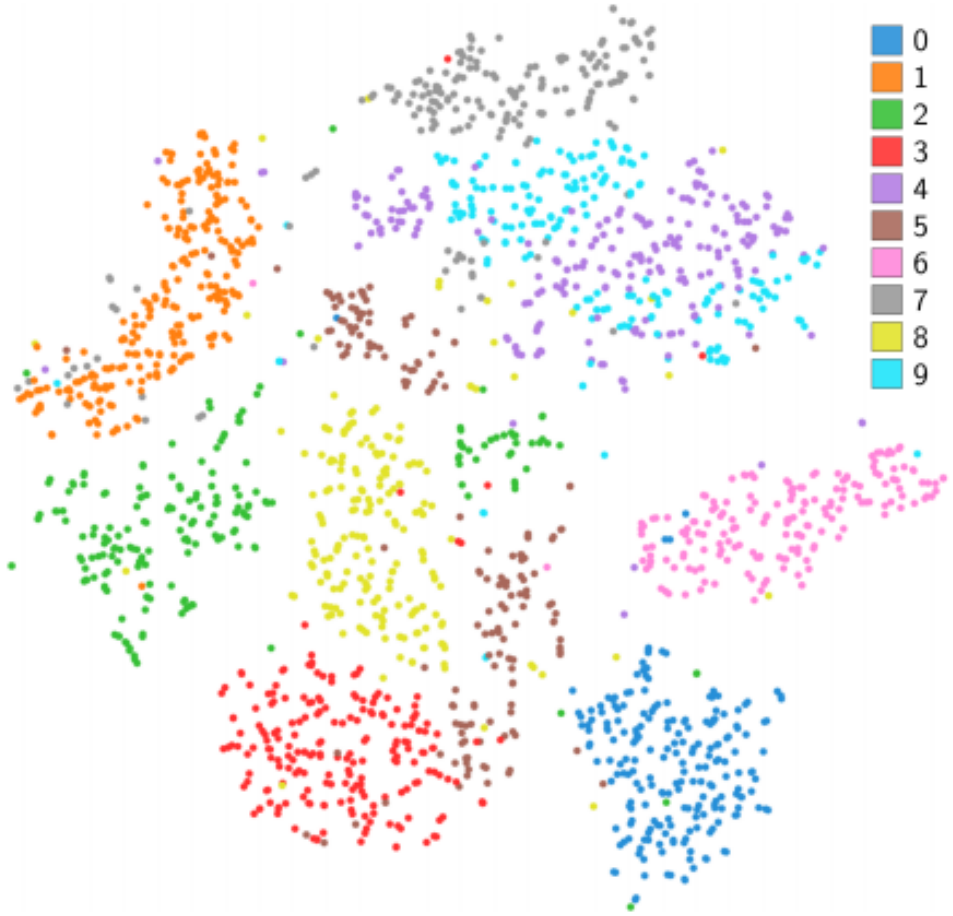
NH（neighborhood hit）： measure projection quality——class separation

k=6，a target point $\mathbf{a}_p$

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

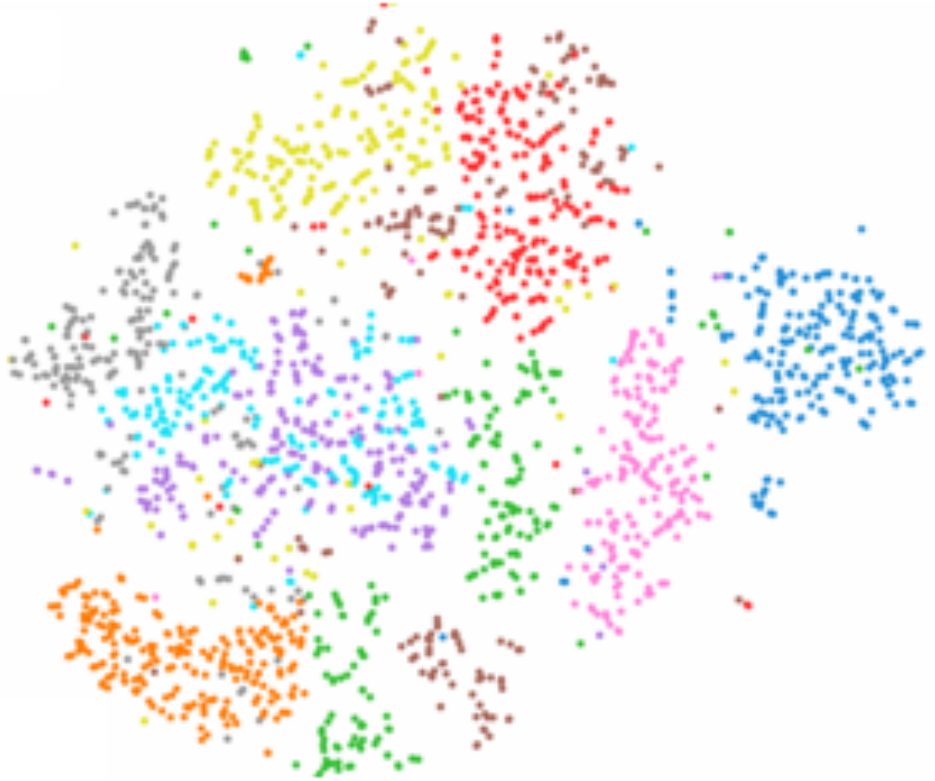
NH(a whole projection) = 所有点的NH的平均

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%)

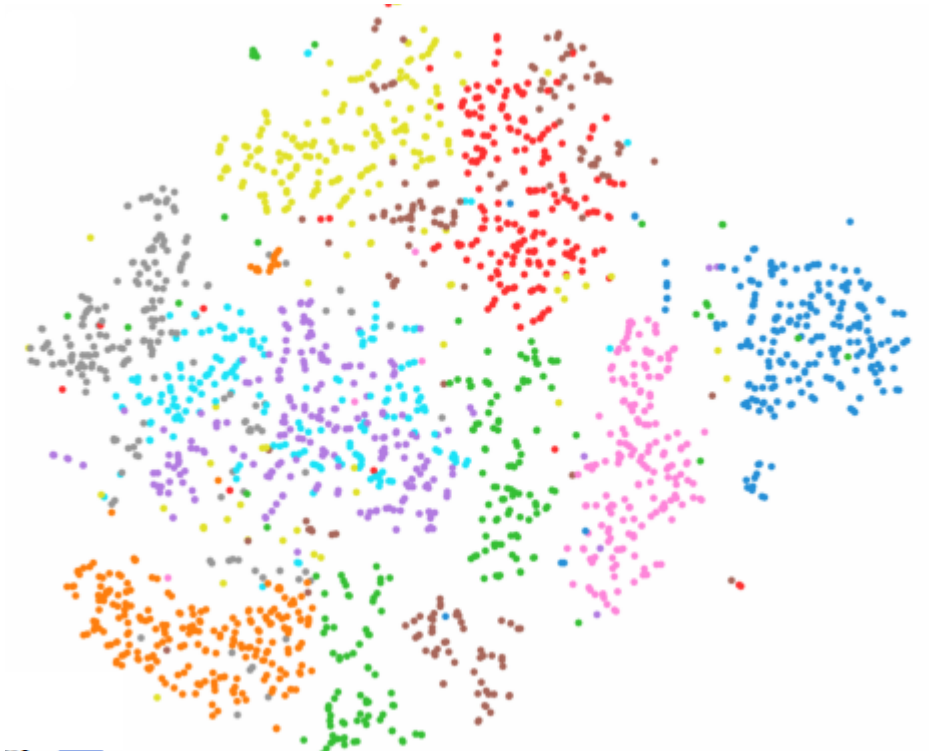
pred、activation（1000）

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

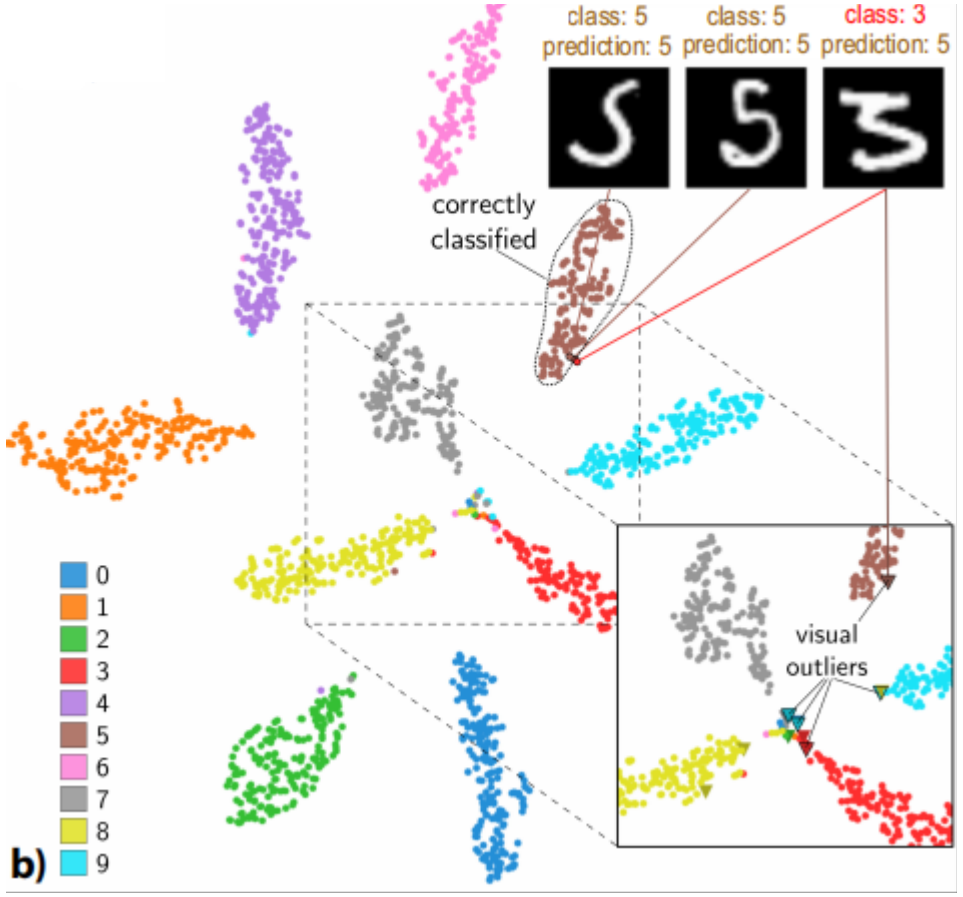
R1：与假设矛盾，有相似的class separation

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



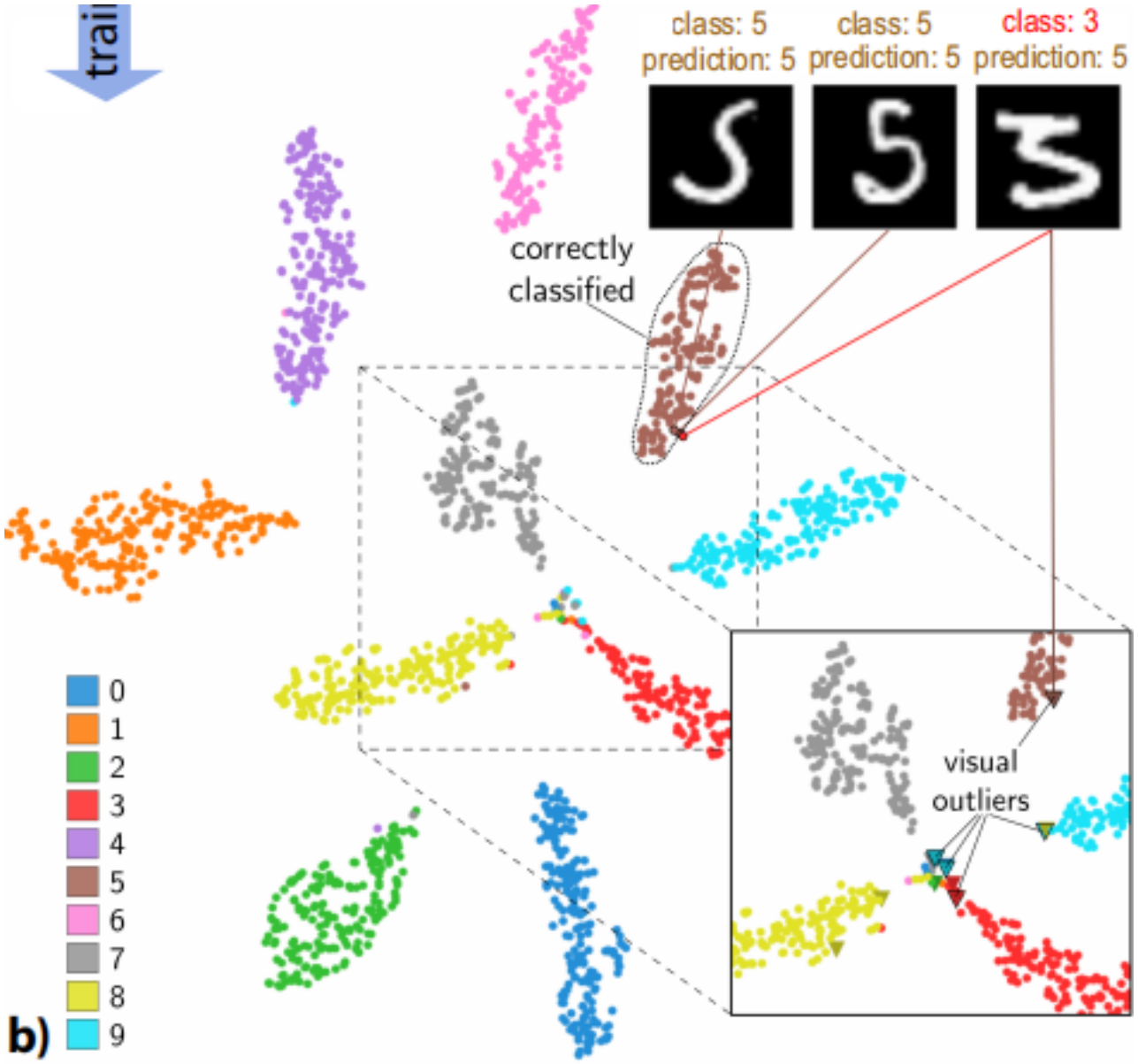
Before training (NH: 83.78%).

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

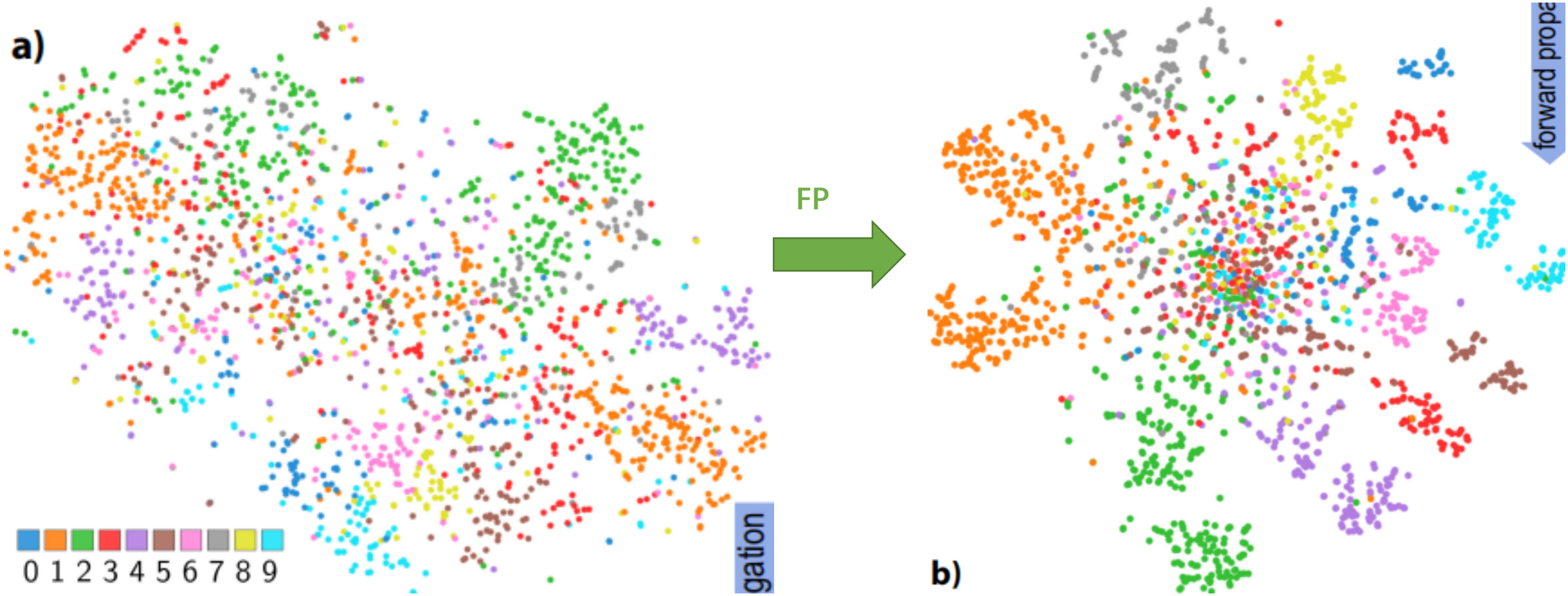
misclassification:隐藏层激活的相似性与最终的类别分配相关联

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



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

to be continued...