Supplemental Material for Exploring explicit coarse-grained structure in artificial neural networks

Xi-Ci Yang,[1] Z. Y. Xie,[2,*] and Xiao-Tao Yang[1,†]

[1] *College of Power and Energy Engineering,*
Harbin Engineering University, Harbin 150001, China

[2] *Department of Physics, Renmin University of China, Beijing 100872, China*

A. COMPLETE EXPRESSION OF z

As illustrated in Fig.4(c) in Main Text, if each local mapping is expanded to the second order, then the complete expression of z in terms of x_i, with $i = 1, 2, 3, 4$, can be written as

$$z = \mathbf{a} W \mathbf{b}^T, \tag{S1}$$

where the two vectors $\mathbf{a}$ and $\mathbf{b}$ are defined as

$$\begin{aligned} \mathbf{a} &= \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_1x_2 & x_1x_2^2 & x_1^2x_2 & x_1^2x_2^2 & x_1x_2^3 & x_1^3x_2 \end{bmatrix} \\ \mathbf{b} &= \begin{bmatrix} 1 & x_3 & x_3^2 & x_3^3 & x_3^4 & x_4 & x_4^2 & x_4^3 & x_4^4 & x_3x_4 & x_3x_4^2 & x_3^2x_4 & x_3^2x_4^2 & x_3x_4^3 & x_3^3x_4 \end{bmatrix}. \end{aligned} \tag{S2}$$

The weight W can be represented as a 15×15 matrix, whose nonzero elements are denoted as 1 in Fig. S1, just for clarity.

B. THE ACTION OF MULTI-CHANNEL CONVOLUTIONS

For concreteness, the convolutional layer with four three-channel kernels will turn the three-channel input data into four-channel output, as illustrated in Fig. S2, where the + sign means equal-weight superposition of the three dot-product results.

C. TAYLORNET STRUCTURE USED IN THE CLASSIFICATION ON CIFAR-10 DATASET

The detailed structure of the TaylorNet used in the Main Text in the classification on CIFRA-10 dataset, is shown in Fig. S3.

$$\begin{array}{c|ccccccccccccccc}
 & 1 & x_1 & x_1^2 & x_1^3 & x_1^4 & x_2 & x_2^2 & x_2^3 & x_2^4 & x_1x_2 & x_1x_2^2 & x_1^2x_2 & x_1^2x_2^2 & x_1x_2^3 & x_1^3x_2 \\
1 & \boxed{1} & \boxed{1} & 1 & 1 & 1 & \boxed{1} & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
x_3 & \boxed{1} & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & & & & & \\
x_3^2 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & & & & & \\
x_3^3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\
x_3^4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\
x_4 & \boxed{1} & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & & & & & \\
x_4^2 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & & & & & \\
x_4^3 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & 0 & & \\
x_4^4 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & & & & \\
x_3x_4 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & & & & & \\
x_3x_4^2 & 1 & & & & & & & & & & & & & & \\
x_3^2x_4 & 1 & & & & & & & & & & & & & & \\
x_3^2x_4^2 & 1 & & & & & 0 & & & & & & & & & \\
x_3x_4^3 & 1 & & & & & & & & & & & & & & \\
x_3^3x_4 & 1 & & & & & & & & & & & & & &
\end{array}$$

FIG. S1. Weight W appearing in Eq. (S1), corresponding to the coefficients of z illustrated in Fig.4(c) in Main Text. For simplicity, the nonzero elements are denoted as 1. The terms in the red box are the ones showing up in the convolution operations.

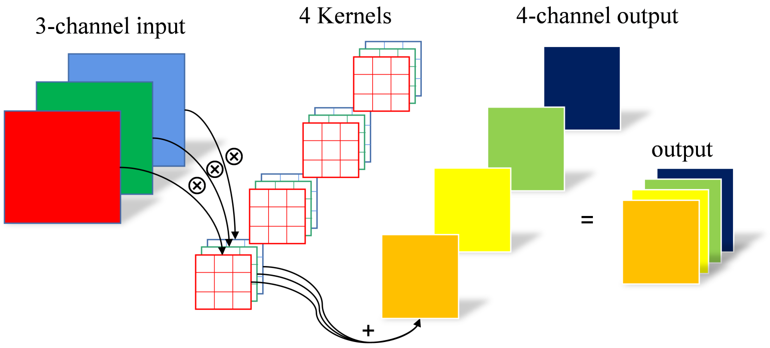


FIG. S2. The action of a convolutional layer with four three-channel kernels, as mentioned in the Main Text.

D. WEIGHT DETAILS OF THE QUADRATIC TERMS IN THE EXPERIMENT ON MNIST DATASET

As to the direct experiment on the Taylor expansion, Eq.(3) in Main Text, on 7 × 7 MNIST dataset, the trained weight of the quadratic terms are sketched in detail in Fig. S4.

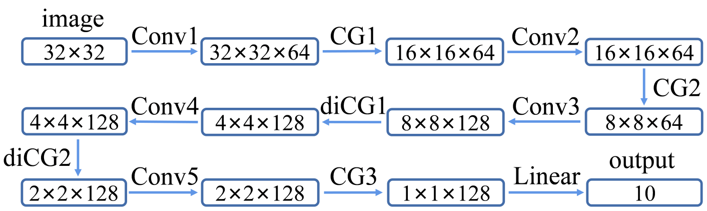


FIG. S3. The TaylorNet used in the classification on CIFAR-10 dataset, as mentioned in the Main Text. Convolutions Conv1 and Conv2 have structure Conv(3,3,64,1,1,1), Conv3, Conv4, and Conv5 have structure Conv(3,3,128,1,1,1). CG operations CG1 and CG2 have structure CG(2,2,64,2,2), CG3 have structure CG(2,2,128,1,1). Dilated CG operations diCG1 and diCG2 have structures diCG(2,2,128,1,1,4) and diCG(2,2,128,1,1,2), respectively.

The statistics are shown in Fig.3 in the Main Text.

E. SIMILARITY ANALYSIS ON THE DISTILLED DATASETS

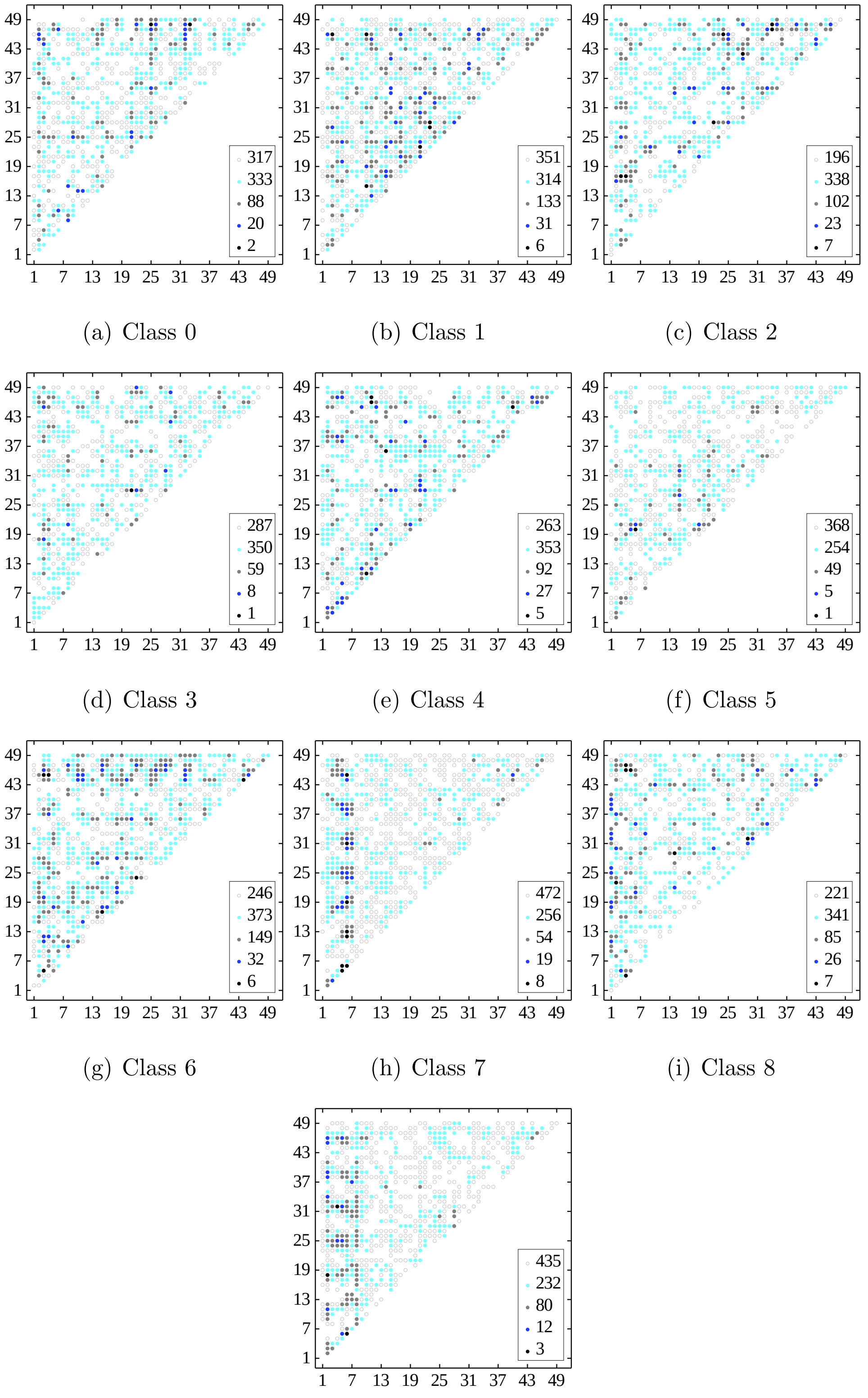
In this section, we try to characterize the similarity between the distilled images in the same class, aiming to see if it is possible to quantify how well the essential features are extracted in the level-by-level distillation process. For this purpose, we adopt two quantities, i.e., the cosine similarity s and the Euclidean distance d defined for each class i and each level α,

$$\begin{aligned} s_i^{(\alpha)} &= \frac{1}{\tau} \sum_{\substack{m,n\in D^{(\alpha;i)} \\ m\neq n}} \langle X_m | X_n \rangle \\ d_i^{(\alpha)} &= \frac{1}{\tau} \sum_{\substack{m,n\in D^{(\alpha;i)} \\ m\neq n}} \langle (X_m - X_n) \,|\, (X_m - X_n) \rangle, \end{aligned} \tag{S3}$$

where $D^{(\alpha;i)}$ denotes the set of all images that belongs to the α-th level dataset $D^{(\alpha)}$ and the i-th category. Suppose there are $n_{\alpha,i}$ images in $D^{(\alpha;i)}$ in total, then τ is defined as the binomial coefficient

$$\tau \equiv \frac{1}{n_{\alpha,i}\,(n_{\alpha,i}-1)}$$

Note in Eq. (S3), X_n denotes the n-th image as a normalized vector, i.e., the inner-product satisfies $\langle X_n | X_n \rangle = 1$. Clearly, the more similar the images are, the larger s and the smaller d should be.



(a) Class 0

(b) Class 1

(c) Class 2

(d) Class 3

(e) Class 4

(f) Class 5

(g) Class 6

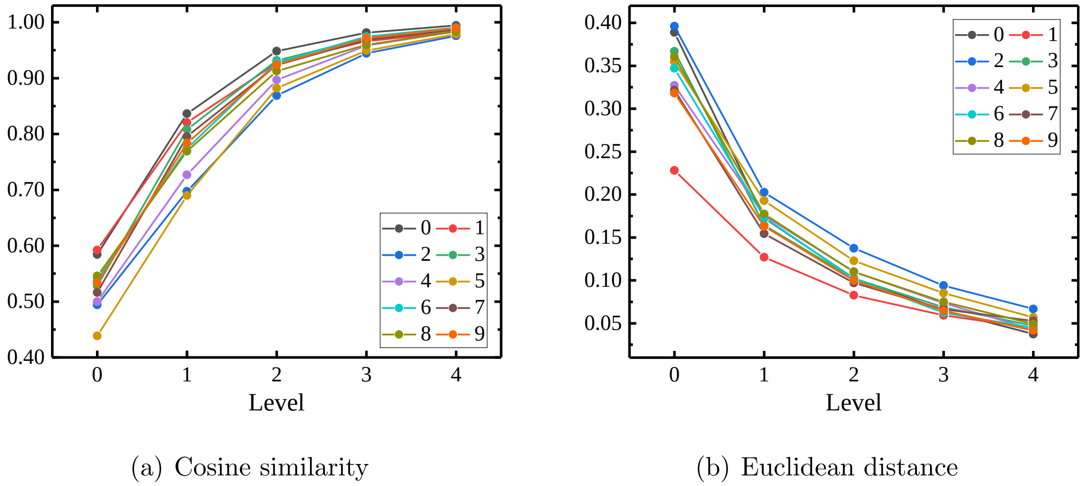
(h) Class 7

(i) Class 8

(j) Class 9

FIG. S4. The weight corresponding to quadratic terms $x_i x_j$ for each class, for experiment of Eq.(3) in Main Text on the 7×7 MNIST dataset. The abscissa and ordinate represent the linear coordinates i and j in the input image, respectively. The weight is divided into five levels according to the magnitude of the value, with darker colors representing larger weights. The inset counts the number of weights belonging to different levels.

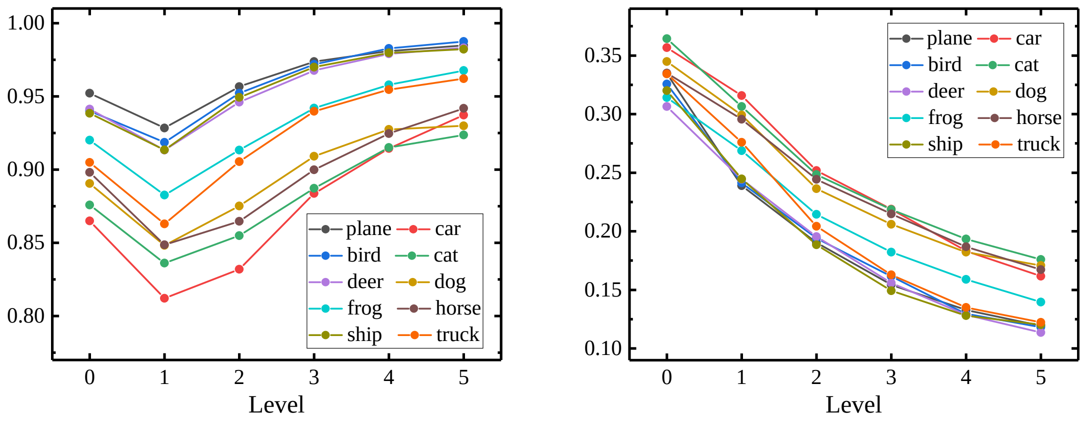
The results are shown in Fig. S5 and Fig. S6. It shows that, as expected, in all the classes in both MNIST and CIFAR-10 datasets, the cosine similarity goes up roughly as the distillation level α becomes higher, while the Euclidean distance goes down systematically. This can be regarded reasonably as direct evidence that as the distillation proceeds, the essential features which are shared by all the images in the same class are gradually extracted, and the irrelevant details which are only owned by some individual images are filtered out effectively. And this exactly reflects the essence of deep learning. Though probably there exist more advanced similarity measures which can produce a better demonstration, the effectiveness of $s_i^{(\alpha)}$ and $d_i^{(\alpha)}$ has already shown that the proposed multi-level distillation process can indeed extract some important and defining features efficiently.



(a) Cosine similarity

(b) Euclidean distance

FIG. S5. Similarity analysis on the MNIST dataset with respect to the distillation levels. The data are obtained for each class. The cosine similarity and Euclidean distance are defined in Eq. (S3). Note: Level-0 means the original MNIST dataset.



(a) Cosine similarity

(b) Euclidean distance

FIG. S6. Similarity analysis on the CIFAR-10 dataset with respect to the distillation levels. The data are obtained for each class. The cosine similarity and Euclidean distance are defined in Eq. (S3). Note: Level-0 means the original CIFAR-10 dataset.