Evaluation Function for Shift Invariant Auto-encoder

Tadashi Matsuo[1] and Nobutaka Shimada[2]

***Abstract*— Essential information can be extracted by the auto-encoder method which employs an encoder and a decoder. Both elements are usually trained so that the combination restores the input as correctly as possible for a certain set of vectors. With a usual ℓ^2 measure, an image and its translated version are considered as different inputs. However, for example glyph images, there are cases where an absolute position in a window is not important, but a spatial shape itself is important. In such cases, essential information should be extracted from a spatial shape without regard to its positional difference. We propose a method to train an auto-encoder that focuses on a spatial shape and ignores spatial shifts. We demonstrate that the proposed method can unsupervisedly encode shapes without regard to its position.**

I. INTRODUCTION

Essential information can be extracted by the auto-encoder method [1], [2], which employs an encoder and a decoder. The encoder converts an input to a descriptor with lower dimensionality, and the decoder approximately restores the original input from the code. Both elements are trained so that the combination restores the input as correctly as possible for a certain set of vectors. Under this constraint, the encoder generates a numerical representation of the principal components required for input restoration. In addition, the encoder and decoder can be trained with unlabeled vectors.

As a measure of restoration, the ℓ^2 norm is often used. An image and its translated version are considered as different inputs. This means that a descriptor obtained by encoding an image includes information of its spatial shape and its absolute position in a window. However, for example glyph images, there are cases where an absolute position in a window is not important, but a spatial shape itself is important. In such cases, essential information should be extracted from a spatial shape without regard to its positional difference. We propose a method to train an auto-encoder that focuses on a spatial shape and ignores spatial shifts.

Max Jaderberg et al. have proposed "Spatial Transformer Networks"[3], which includes a module to learn a spatial transformation effective in classification. The transformation module is possible to cancel a transformation including a spatial shift, but it must be trained with a teacher label for each input image.

We propose an evaluation function to train an auto-encoder so that it extracts a typical spatial shape itself without regard to its relative position in a window. It can be applied to various auto-encoders without requiring special modules and labeled training data. By the proposed method, we can encode shapes in images even if images are difficult to be labeled, such as appearances of hand–object interactions. In this paper, we consider shift transformations, but the proposed method can be applied other transformations.

II. EVALUATION FUNCTION FOR SHIFT INVARIANT AUTO-ENCODER

Generally, an auto-encoder is trained such that the encoder–decoder combination approximately restores an input in a certain input set. It is formulated as

$$\operatorname*{argmin}_{D,E} \sum_{I \in S} \left\| I - D\left(E\left(I\right)\right) \right\|_2^2, \tag{1}$$

where I, S, $D(\cdot)$, $E(\cdot)$, and $\left\|\cdot\right\|_p$ denote the input to be reconstructed, a set of inputs, the encoder, the decoder, and the ℓ^p norm, respectively. To force an auto-encoder to capture a spatial shape, we add terms and relax the above restoration error.

To ignore spatial shifts, an encoder should be shift invariant. We force an encoder to be shift invariant by adding a shift invariance term to the evaluation function. The term is sum of differences between a restored image and an image restored from a shifted input.

To ensure restoration of a spatial shape, we relax the restoration error in (1) by evaluating the minimum of errors between the restored image and a shifted input image.

Our evaluation function is formulated as

$$\begin{aligned} \sum_{I \in S} \lambda_{\text{invariance}} & \left\{ \sum_i \left\| D\left(E\left(I\right)\right) - D\left(E\left(T_i I\right)\right) \right\|_2^2 \right\} \\ & + \lambda_{\text{restoration}} \min_i \left\| T_i I - D\left(E\left(I\right)\right) \right\|_2^2 \\ & + \lambda_{\text{sparseness}} \left(\frac{\left\| E\left(I\right)\right\|_1}{\left\| E\left(I\right)\right\|_2} \right)^2, \end{aligned} \tag{2}$$

where T_i, $\lambda_{\text{invariance}}$, $\lambda_{\text{restoration}}$ and $\lambda_{\text{sparseness}}$ denote a shift operator and scalar weights of each term. The shift operator T_i is defined as follows

$$(T_i I)(x, y) = I(x + \Delta x_i, y + \Delta y_i), \tag{3}$$

where $I(x, y)$ denote the value of the image I at the position (x, y).

We train the encoder E and the decoder D so that they minimize the proposed evaluation function (2). The shift operators $\{T_i\}$ are fixed when training the proposed auto-encoder. The first term in (2) evaluates shift invariance and the second term evaluates shape restoration. The third term evaluates spatial sparseness of descriptors[4].

*This work was supported by JSPS KAKENHI Grant Number 24500224, 15H02764 and MEXT-Supported Program for the Strategic Research Foundation at Private Universities, S1311039, 2013-2016.

[1,2]Tadashi Matsuo and Nobutaka Shimada are with College of Information Science and Engineering, Ritsumeikan University, 1-1-1 Noji-higashi, Kusatsu, Shiga, JAPAN [1]matsuo@i.ci.ritsumei.ac.jp

III. EXPERIMENTS

Since an output of a convolutional neural network (CNN)[5] is calculated by convolution between a kernel and an input, a CNN is suitable for extracting local features without regard to their positions. To make an encoder nearly shift invariant, we construct an encoder consisting of single CNN and following three layer fully connected neural network (NN). Jonathan Masci et al. has proposed a convolutional auto-encoder (CAE)[6], where both an encoder and a decoder includes convolution. However, in our problem, a decoder does not have to be shift invariant as long as it can accurately restore a shape from a low dimensional descriptor. Therefore, we construct a decoder as three layer fully connected NN. In addition, we use a hyperbolic tangent as an activation function, which is placed between each pair of layers.

The encoder and the decoder are trained as two auto-encoders with the same set of images. One is trained with ordinary evaluation function (1). The other is trained with our proposed function (2), where shift parameters are defined as

$$\{(\Delta x_i, \Delta y_i)\} = \{-2, -1, 0, +1, +2\}^2 . \tag{4}$$

We extract training input images from an image scanned from the document [3] in 300 DPI resolution. The training images are $(32 \times 32)[\text{pixel}]$ and they are randomly extracted in any position of any pages. In the experiments, a descriptor is represented by a 30-dimensional vector. After training the ordinary auto-encoder and the proposed one, we apply mean shift clustering to descriptors calculated from $(32 \times 32)[\text{pixel}]$ images for each auto-encoder. On the clustering, we use images extracted from p.1–2 of the document [3].

Fig. 1 shows a part of the document and an extracted region. The cluster to which the region belongs is shown in Fig. 2(a) and (b). The former is generated by the ordinary auto-encoder, the latter is generated by the proposed auto-encoder. The images are arranged from left to right and top to bottom in ascending order of ℓ^2 distance to the center of the corresponding cluster in the descriptor space. For example, a descriptor encoded from the leftmost and topmost image in Fig. 1 (a) is the closest vector to the center of the cluster that the descriptor belongs to, where distance is measured by the ℓ^2 norm. A descriptor encoded from the right most and bottommost image in Fig. 1 (a) is the 48-th farthest vector.

The images in Fig. 2(a) include the glyph “o” (or a similar glyph) placed at the same position. Since a descriptor by the ordinary auto-encoder includes positional information, a cluster consists of images with glyphs placed at a common position. In contrast to Fig. 2(a), the glyph “o” in Fig. 2(b) are placed at various positions. Since a descriptor by the proposed auto-encoder is mainly determined by a spatial shape, clusters of images with similar shapes can be generated. In Fig. 2(c), we show the images restored from 2(a) by the ordinary auto-encoder. In Fig. 2(d), we show the images restored from 2(b) by the proposed auto-encoder. In Fig. 2(b), the glyphs of “o” are placed at different position for each image, but the patterns in the restored images in Fig. 2(d) are placed at similar positions. This means that the proposed auto-encoder is invariant to shifts of input. The proposed auto-encoder successfully generates the group of glyphs “o” without regard to its position unsupervisedly.

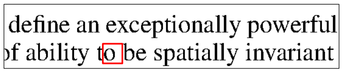
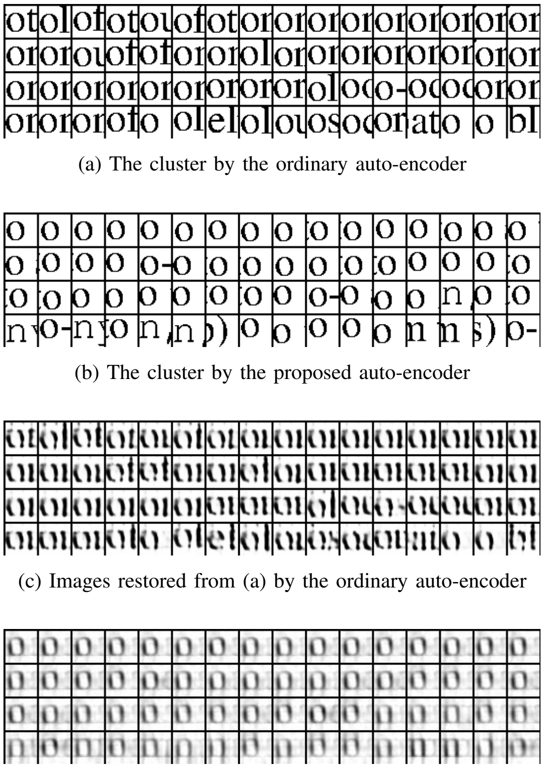

define an exceptionally powerful
of ability to be spatially invariant

Fig. 1. Extracted region for “o ” region


(a) The cluster by the ordinary auto-encoder

(b) The cluster by the proposed auto-encoder

(c) Images restored from (a) by the ordinary auto-encoder

(d) Images restored from (b) by the proposed auto-encoder

Fig. 2. The extracted clusters for “o ” region

Fig. 3 shows other example of regions. The clusters to which the region “io” belongs are shown in Fig. 4(a) and (b). The clusters to which the region “ar” belongs are shown in Fig. 5(a) and (b). The figures show that clusters by the ordinary auto-encoder include various shapes, but clusters by the proposed auto-encoder include shapes more similar to each other.

IV. CONCLUSIONS

We proposed a method to train an auto-encoder that focuses on a spatial shape and ignores spatial shifts. By some experiments, we showed that the proposed auto-encoder can encode a shape without regard to its position in a window.

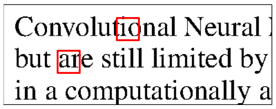
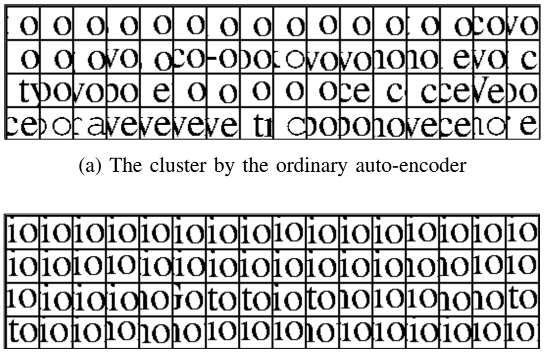

Convolutional Neural
but are still limited by
in a computationally a

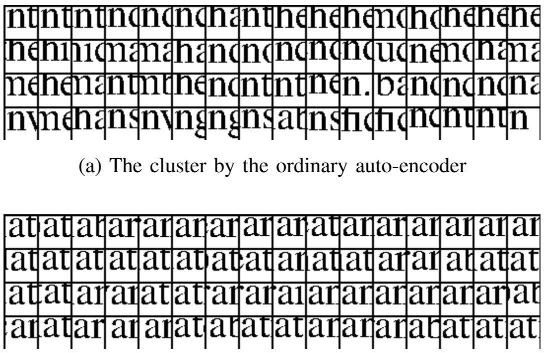
Fig. 3. Extracted regions of “io” and “ar”



(a) The cluster by the ordinary auto-encoder

(b) The cluster by the proposed auto-encoder

Fig. 4. The extracted clusters for “io” region



(a) The cluster by the ordinary auto-encoder

(b) The cluster by the proposed auto-encoder

Fig. 5. The extracted clusters for “ar” region

It can unsupervisedly encode shapes in images which are difficult to be labeled. As a future work, we are trying to encode hand–object interactions by their appearances, which are difficult to be labeled, as a base of modeling relation between an object and possible interactions.

REFERENCES

[1] P. Baldi and K. Hornik, “Neural networks and principal component analysis: Learning from examples without local minima,” *Neural Networks*, vol. 2, no. 1, pp. 53 – 58, 1989. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0893608089900142

[2] A. Makhzani and B. J. Frey, “k-sparse autoencoders,” *CoRR*, vol. abs/1312.5663, 2013. [Online]. Available: http://arxiv.org/abs/1312.5663

[3] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, “Spatial transformer networks,” *CoRR*, vol. abs/1506.02025, 2015. [Online]. Available: http://arxiv.org/abs/1506.02025

[4] T. Matsuo and N. Shimada, “Construction of latent descriptor space of hand-object interaction,” in *Proceedings of 22nd Japan-Korea Joint Workshop on Frontiers*, Feb. 2016, pp. 117–122.

[5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.

[6] J. Masci, U. Meier, D. Cirean, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” in *Artificial Neural Networks and Machine Learning ICANN 2011*, ser. Lecture Notes in Computer Science, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds. Springer Berlin Heidelberg, 2011, vol. 6791, pp. 52–59. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-21735-7_7