High-order Deep Neural Networks for Learning Multi-Modal Representations

Kyoung-Woon On
Eun-Sol Kim
Byoung-Tak Zhang
Department of Computer Science and Engineering, Seoul National University, Seoul 151-744, Korea

In multi-modal learning, data consists of multiple modalities, which need to be represented jointly to capture the real-world 'concept' that the data corresponds to (Srivastava & Salakhutdinov, 2012). However, it is not easy to obtain the joint representations reflecting the structure of multi-modal data with machine learning algorithms, especially with conventional neural networks. This is because the information which consists of multiple modalities has distinct statistical properties and each modality has a different kind of representation and correlational structure (Srivastava & Salakhutdinov, 2012). Also, noise exists in information from multi-modal input, which makes the information unreliable and inaccurate (Ernst & Di Luca, 2011).

In this paper, we develop High-order deep neural networks (HODNN) to learn the representations of multi-modal data. The HODNN connects abstract information of multiple modalities with high-order edges, which lead to a multiplicative interaction rather than just additive interaction used in conventional deep neural networks. This high-order interaction not only captures highly non-linear relationship across the modalities but also suppresses the uncorrelated noise efficiently. In addition, we apply general deep structure to each modality so as to obtain the balanced abstract information from each of it. (Ngiam et al., 2011).

Thus, the HODNN consists of two parts: modal-specific learning layers and a joint representations learning layers. The modal-specific learning layers have connections only within each modality, so the highest hidden layers of each modality represent abstract information of that modality. The joint representations learning layers is composed of higher-order interactions between the joint hidden units and multiple groups of modal-specific hidden units. The joint hidden units can learn non-linear correlations among the modal-specific hidden units. The overall architecture of HODNN is illustrated in Figure 1. In details, the modal-specific learning layers follow a general neural networks framework. The hidden representations $h_j^1$ of the specific modality is obtained by an Equation 1.

$$h_j^1 = \sigma \left( \sum_i W_{ij}^1 v_i^1 + \text{bias} \right)$$ (1)

Motivated by (Memisevic & Hinton, 2010), the multi-way factoring method is employed to maintain efficient model complexity without loosing capacity of the model heavily. As a consequence, we can obtain the joint hidden representations $h_k^{\text{joint}}$ as follows:

$$h_k^{\text{joint}} = \sigma \left( \sum_{i,j} W_{ij}^{\text{joint}} h_i^1 h_j^2 + \text{bias} \right)$$ (2)

As preliminary experiments, we focus on the joint representations learning layers to show the effect of high-order interaction. For the joint representations learning layers, the factorized version of high-order Boltzmann machine (Figure 2a) is used as a building block (Sejnowski, 1986).
High-order Deep Neural Networks for Learning Multi-Modal Representations

Figure 2. Comparative models for demonstrating effect of high-order interaction

Also, the MNIST dataset is utilized which consists of handwritten digit images and the corresponding labels. While the label vectors of the MNIST are used as targets in a discriminative task, in our experiments, the image and the label vectors are used as two different modalities, the former is for visual information and the latter is for textual information which indicate the concept 'number'. To show the competence of our model, a shallow bi-model RBM (Figure 2b) is used as a comparative model. The shallow bimodal RBM also functions as a module combining different modalities in conventional multi-modal deep networks (Ngiam et al., 2011; Srivastava & Salakhutdinov, 2012).

In order to see how the joint hidden units represent the abstract information across the modalities, first of all, 2D t-SNE embedding algorithm is applied to hidden units of both factored high-order Boltzmann machine and shallow bi-modal RBM (figure 4). The result of the factored high-order Boltzmann machine is shown to be more visually discriminative than that of the shallow bi-modal RBM. Also, it is interesting to notice that the embedding result of the joint hidden representations with factored high-order Boltzmann machine looks as if the representations of each number form their own sub-manifold structure. These interpretations imply that the representation power of suggested model is greater than the comparative model.

Second, to demonstrate how the high-order interaction efficiently cancels out the noise of either modality which is uncorrelated with other modality, it is appropriate to compare the joint hidden representations of both models when the noisy input is fed in. For this experiment, we firstly generated a corrupted dataset which consists of the 100 image inputs of number ‘1’ corrupted by other numbers and corresponding clear text inputs of number ‘1’ (Figure 3). The joint hidden representations of the corrupted dataset are shown with normal dataset by using 2D t-SNE visualization (Figure 4). As expected, in the factored high-order Boltzmann machine, the representations of the corrupted dataset are located near the representations of number ‘1’. However, in the shallow bi-modal RBM, the representations of corrupted dataset lie scattered across the numbers. It reveals the property of high-order interaction that remove the uncorrelated noise information of either modality based on the other modality.

In future works, we aim to apply our model to the study of event cognition which have emerged over the last years as a vibrant topic of scientific study. It is an important research topic because much of our behavior is guided by our understanding of events which are what happens to us, what we do, what we anticipate and what we remember in our daily life (Radvansky & Zacks, 2014). For better understanding of the study, we employ multiple wearable sensors to record the daily-life of a person. This is because the collected data from the viewpoint of the first-person plays a significant role in learning of human behavior (Zhang, 2013; Kim et al., 2016; Lee et al., 2016). Through this study, we hope to suggest an event cognition model which can perceive real-time events in real-life by using multiple wearable sensors.
Acknowledgement

This work was partly supported by the Korea government (IITP-R0126-16-1072-SW.StarLab, KEIT-10044009-HRI.MESSI, KEIT-10060086-RISF).

References


