SUTD 2022 50.039 Homework 1

James Raphael Tiovalen / 1004555

Introduction

The authors of the paper “A fast learning algorithm for deep belief nets” discussed a fast, greedy algorithm that can train densely-connected deep belief nets (DBNs) with multiple hidden layers. Evaluating the paper, we deem that while the wake-sleep and up-down algorithms can be used to fine-tune a DBN and improve both the generative model and its ability to classify correctly, they are quite slow. Furthermore, there are more modern network architectures that would perform better on benchmark datasets, such as convolutional neural networks (CNNs).

Summary

The paper first discussed the concept of “complementary priors”, which managed to eliminate the explaining away effects that would make inference difficult in DBNs. Then, the paper proceeded to illustrate that DBNs are essentially probabilistic generative neural network models that consist of a stack of Restricted Boltzmann Machines (RBMs), which in turn consist of multiple stochastic, latent variables. The latent variables, also sometimes known as hidden units or feature detectors, usually have binary values.

Then, the paper introduced the unsupervised learning algorithm that took advantage of the concept of “complementary priors”. The paper then showed that the stacking procedure improves a variational lower bound on the likelihood of the training data, and thus, the greedy algorithm would theoretically be able to achieve approximately maximum likelihood learning. The DBN weights are optimized using a layer-by-layer algorithm with a time complexity proportional to the size and depth of the networks. This approach determines how the variables in one layer depend on the variables in the layer above through the top-down, generative weights. After learning, a single bottom-up run that starts with an observed data vector in the bottom layer and applies the generative weights in the reverse direction can be used to infer the values of the latent variables in each layer.

The paper then also showed a fine-tuning method using the up-down algorithm during optimizing the weights so that the model does not suffer from the “mode-averaging” problems that could make the model learn poor recognition weights. By adding a final layer of variables that represent the intended outputs and backpropagating error derivatives, discriminative fine-tuning can be accomplished. Backpropagation works significantly better when networks with many hidden layers are applied to highly structured input data, such as photos, if the feature detectors in the hidden layers are initialized by learning a deep belief net that represents the structure in the input data (Hinton & Salakhutdinov, 2006).

An example generative model with three hidden layers was also demonstrated to have outperformed a discriminative one on the MNIST dataset after finetuning in the paper. The model had 1.25% errors on the 10000 digit official test set, which was better than the 1.5% achieved by general back-propagation nets. It was also slightly better than a support vector machine model (with 1.4% error).

Critique

CNNs have performed better than DBNs by themselves in the current literature on benchmark computer vision datasets such as MNIST. If the dataset is not a computer vision one, then DBNs can most definitely perform better. In theory, it seems that DBNs should be the best models and that they should perform better than their more modern counterparts such as CNNs. However, it is still very hard to estimate joint probabilities accurately at the moment, and thus, other CNN-based models have become more computationally efficient, at least for the moment.

Additionally, in the paper, binomial input units were used to encode pixel gray levels as if they were the probability of a binary event. In the case of handwritten character images, this approximation works well, but in other cases, it does not. Experiments showing the advantage of using Gaussian input units rather than binomial units when the inputs are continuous-valued are described in Bengio et al. (2007).

That said, it is understandable if the paper was slightly outdated and behind the current best practices since the paper was published in 2006. In fact, there have been efforts to combine both DBNs and CNNs (Lee et al., 2009) in order to get even better results.

Conclusion

In conclusion, while this paper was relatively outdated, it outlined an important piece of research work that contributed to the momentum in the field of representation learning, allowing us to further improve our understanding of machine learning even more, and thus, indirectly pushing up the adoption rate of machine learning in real-life business, societal, and government applications to better humankind.

References

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *In Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS'06).* MIT Press, Cambridge, MA, USA, 153–160. https://dl.acm.org/doi/10.5555/2976456.2976476

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, 1–8. https://doi.org/10.1145/1553374.1553453