Semi-supervised deep kernel learning

Neal Jean, Michael Xie, Stefano Ermon
Department of Computer Science, Stanford University
{nealjean,xie,ermon}@stanford.edu

Abstract

Deep learning techniques have led to massive improvements in recent years, but large amounts of labeled data are typically required to learn these complex models. We present a semi-supervised approach for training deep models that combines the feature learning capabilities of neural networks with the probabilistic modeling of Gaussian processes and demonstrate that unlabeled data can significantly improve performance on real-world datasets.

Introduction

The prevailing trend in machine learning is to automatically discover good feature representations through end-to-end optimization of deep neural networks [1, 2]. However, most tasks where deep learning has been applied with great success have been characterized by large quantities of labeled data for supervised learning [3, 4, 5, 6]. Building on the deep kernel learning methods introduced by Wilson et al. [7], we propose a semi-supervised approach that combines the feature learning capabilities of deep neural networks with the ability of Gaussian processes to quantify uncertainty. By simultaneously maximizing the marginal likelihood of labeled data and minimizing the posterior variance of unlabeled data, large quantities of cheaply collected data can be used for learning.

Deep kernel learning

The deep kernel learning (DKL) model combines the adaptive feature representations of a neural network with a Gaussian process (GP) by using the learned embeddings as input to a GP kernel [7]. Given input data $x \in \mathcal{X}$, a neural network is used to extract feature vectors $h_\theta(x)$. The DKL model then models the outputs as

$$f(x) \sim \mathcal{GP}(\mu(h_\theta(x)), k_\phi(h_\theta(x), h_\theta(x')))$$

for some mean function $\mu(\cdot)$ and covariance kernel $k_\phi(\cdot, \cdot)$, where θ and ϕ represent the neural network and GP parameters respectively. For labeled data (X_L, y), the model is jointly learned by maximizing the log marginal likelihood, $\log p(y \mid X_L, \theta, \phi)$ [8].

Semi-supervised deep kernel learning

To incorporate information from unlabeled data, we exploit the fact that the probabilistic model provides us with a predictive posterior distribution, i.e., it is able to quantify the uncertainty in its predictions. Instead of maximizing the marginal likelihood of the labeled training data in a purely supervised fashion, we train a semi-supervised model by minimizing the compound objective

$$L_{semisup}(\theta, \sigma, \lambda) = -\frac{1}{n} \log p(y \mid X_L, \theta, \sigma, \lambda) + \frac{\alpha}{m} \sum_{j : x_j \in X_U} \text{cov}(X_U)_{jj}$$

where n and m are the numbers of labeled and unlabeled examples and α is a weighting constant controlling the tradeoff between maximizing the likelihood of our observations and minimizing the posterior variance on unlabeled data. This semi-supervised objective has a regularizing effect, discouraging the neural net from learning features that do not generalize well to unlabeled data.

Workshop on Bayesian Deep Learning, NIPS 2016, Barcelona, Spain.

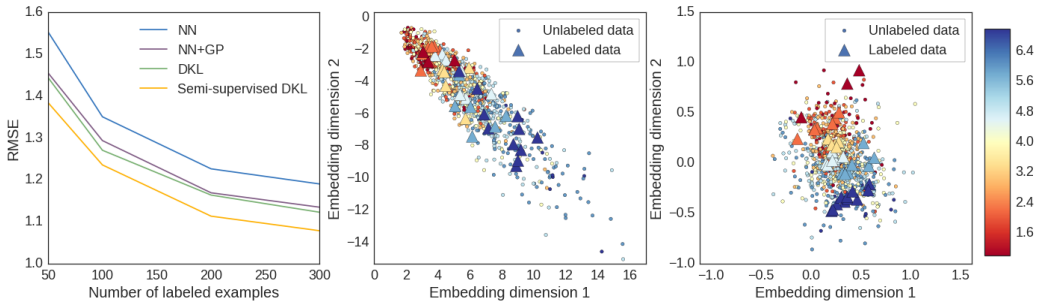

Figure 1: **A:** Average test RMSE vs. number of labeled examples for UCI Skillcraft dataset averaged over 10 trials of randomly sampled data. **B:** Two-dimensional embeddings learned by supervised DKL model using 50 labeled training examples. Large triangles represent labeled data, while small circles represent unlabeled data. The colors indicate ground truth for the output variable, which is treated as unknown for unlabeled data. **C:** Embeddings learned by semi-supervised DKL model using the same 50 labeled training examples plus 1000 unlabeled examples.

We evaluate the semi-supervised DKL approach on the Skillcraft dataset from the UCI repository, a regression task with 18-dimensional features and a real-valued output from 1 to 8 [9]. Fig. 1A compares the semi-supervised model to several approaches that use only labeled data: stand-alone neural networks (NN), fixed neural networks with a Gaussian process on top (NN+GP), and DKL models where the neural network and Gaussian process are trained jointly. Following Wilson et al. [7], the NN+GP model is initialized with the trained NN parameters, and the DKL model is initialized from the corresponding trained NN+GP. Our semi-supervised DKL model outperforms the purely supervised methods when labeled data is limited.

To gain some intuition about how unlabeled data helps learning, we visualize the neural network embeddings learned by the DKL (Fig. 1B) and semi-supervised DKL models (Fig. 1C). The semi-supervised DKL model learns a representation in which the unlabeled examples are more closely clustered around labeled examples. When labeled data is scarce, complex models such as neural networks are prone to overfitting and learning feature representations that fail to generalize well to unseen data. By encouraging the learned features to also minimize predictive variance, the semi-supervised DKL model effectively uses unlabeled examples as additional training data.

Related work

The success of deep neural networks lies in the representative power of deep, but finite, hierarchies of parameterized basis functions [1, 2]. Conversely, non-parametric Gaussian processes can use infinitely many fixed basis functions through a covariance kernel that captures structure in the data [8, 10, 11]. Damianou and Lawrence [12] introduced deep Gaussian processes, which stack GPs by modeling the outputs of one layer with a GP in the next layer. The deep kernel learning method of Wilson et al. [7] combines neural networks with the non-parametric flexibility of Gaussian processes, training the model end-to-end in a supervised setting.

Our semi-supervised learning objective draws inspiration from transductive experimental design, which chooses informative experiments by seeking data points that are both hard to predict and informative for the unexplored test data [13]. Other methods based on prediction uncertainty have also been explored, such as minimum entropy regularization [14, 15], as well as methods that leverage unsupervised pre-training or stochastic perturbations [16, 17].

Conclusions

Many important problems are challenging in large part because of the limited availability of training data. In these settings, the ability to learn from unlabeled data is critical. We show that more powerful hierarchical feature representations can be learned when deep neural networks and Gaussian processes are jointly trained to optimize a semi-supervised objective that aims to minimize uncertainty over unlabeled data.

References

[1] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2013).

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *Nature* 521.7553 (2015), pp. 436–444.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[4] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[5] Geoffrey Hinton et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups". In: *IEEE Signal Processing Magazine* 29.6 (2012), pp. 82–97.

[6] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2013), pp. 6645–6649.

[7] Andrew Gordon Wilson et al. "Deep Kernel Learning". In: *The Journal of Machine Learning Research* (2015).

[8] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. The MIT Press, 2006.

[9] Moshe Lichman. *UCI Machine Learning Repository*. 2013. URL: `http://archive.ics.uci.edu/ml`.

[10] Zoubin Ghahramani. "Probabilistic machine learning and artificial intelligence". In: *Nature* 521.7553 (2015), pp. 452–459.

[11] Radford M. Neal. *Bayesian Learning for Neural Networks*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1996. ISBN: 0387947248.

[12] Andreas C. Damianou and Neil D. Lawrence. "Deep Gaussian Processes". In: *The Journal of Machine Learning Research* (2013).

[13] Kai Yu, Jinbo Bi, and Volker Tresp. "Active Learning via Transductive Experimental Design". In: *The International Conference on Machine Learning (ICML)* (2006).

[14] Yves Grandvalet and Yoshua Bengio. "Semi-supervised learning by entropy minimization". In: *Advances in neural information processing systems*. 2004, pp. 529–536.

[15] Chenyang Zhao and Shaodan Zhai. "Minimum variance semi-supervised boosting for multi-label classification". In: *2015 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2015, pp. 1342–1346.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Using deep belief nets to learn covariance kernels for Gaussian processes". In: *Advances in neural information processing systems*. 2008, pp. 1249–1256.

[17] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. "Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning". In: *arXiv preprint arXiv:1606.04586* (2016).