Application of Deep Learning in NLP and MultiModal Extraction

Kushal Arora;karora@cise.ufl.edu

March 18, 2015

Summary

Deep learning has been around for long starting from boltzman machine and multilayer preceptron. The biggest reason that was holding the progress back was the non convexity of objective function. Deeper networks always gave inferior results as compared to the deeper networks because the objective function was stuck in an inferior minima due to exploding-diminishing gradients. Only in 2006 with work of Hinton et al the field took off. The trick was pre-training of the layers, followed by the fine-tuning for actual objective. One of the core strength of deep models is the its ability to learn representation of data in an embedding space. This can be truly useful for domains where the data is discrete like NLP and vision. In these fields these models are used to project data in a continuous space and learn a supervised or unsupervised probability measure for discriminative or generative purposes.

NLP

One of the early work in this field was done by Bengio[1] by proposing Neuro Probabilistic Language Model. This model learnt both the embeddings of the words and probability of generating next word given a sequence. This model was able to beat the at that time state of the are Kneser Nayes Smooting based models. One of the drawback of this model was the context window limitation which was unable to capture the longer dependencies. The recurrent neural net model proposed by Mikolov[13, 14, 12] improved this by using Elman Networks and beating the state of the art. The Log Bilinear model proposed by Minh & Hinton[17] and Morin & Bengio[18] improved the scalability of the models by reducing the training time by $\mathcal{O}(V/\log V)$.

Collbert and Weston[6] attempted to apply the neural network approach to standard NLP tasks like Chunking, Part of Speech Tagging, Semantic Role Labelling, Name Entity Recoginition, Synonym prediction. They achieved state of the art performance in most of these tasks. They improved the performace in the follow up paper by incorporating some hand crafted features to strengthen the model on raw data.

Socher et. al. have done produced various state of the art result in language parsing and tagging[24, 25]. Their approach is based on recursive auto encoders instead of convolutional nets approach of C & W. Along with this a considerable work has been done by the same lab in sentiment analysis[27, 22, 26] and paraphrase detection[21]. The results for both are the current state of the art.

The core strength behind these improvements are continuous space word embeddings learned by these models. A lot of new and efficient approaches to learn these embedding have been proposed. Huang et al[8] have tried an approach to capture global semantics in representation by using two networks, one for local feature learning and other for global context learning. In the same paper they try to handle polysemy and homonemy using multiple representation for a word. Mikolov[15, 16] et al at Google has proposed a continuous skip-gram model to efficiently learn the word embedding. In an extension this work they learn phrases which have different meaning that their composition. GloVe[9] project by Paddington et al proposes alternate approach of learning continuous vectors from word-word co-occurrence matrix. Turian et. al.[29] have done a comparative studies of various embeddings on standard NLP tasks.

Future Work

One of the underlying theme in all these works has been to learn a better representation of the language in continuous domain. The advantage of this approach is that it is easy to extrapolate in continuous domain and coupled with a discriminative probability measure learnt on this space we can produce a model which generalizes better. My current thesis is a step in this direction where I am attempting to build a compositional model which represents both words and sequences in the continuous domain and learn a compositional operator on them. Once we learn the composition and phrase embedding in the space, we can easily embed any unseen sentence or phrase in continuous space and use the discriminative classifier on it for prediction.

MultiModal Learning

Most the application I could find for multi-modal data extraction were in the field of joint learning of text and images. The reason why deep models work well in such scenarios is that we can learn a common embedding for different modalities. Work in this domain has mostly been focused to three type of tasks 1.) unsupervised and semi-supervised image segmentation and classification[10, 7], 2.) scene description[28, 23, 30] 3.) embedding language in visual domain and its application like video and image search[10, 11, 20].

There are plenty of approaches to these tasks as this is still a very active area of research. I couldn't understand many of the approaches due to complexity and lack off time. Approach that was easiest for me to understand and is closest to what I have been reading is again by Socher et al. In [20] Socher et al tries to embed the images into the same domain as words with same words and images sharing embeddings. The drawback of this model is that it cannot describe

multi-object scenes. In a follow up work [11] this drawback is overcome by representing phrases too. The task of generating sentences from the language has been attempted by the group in following paper[23] based on recursive autoencoder used for sentiment classification and parsing. A lot of interesting deep learning approaches in each category exist and some of the promising ones are referred above, but due to time constraint I was not able to go through all of them. I can focus on a specific portion depending upon the task at hand.

Knowledge Base- Extraction and Inference

Almost all the applications of Deep Learning in knowledge bases has been focused around embedding the entities in continuous space and learn relations as tensors. Most of the work in done by Antonie Bordes from CNRS, France in collaborations with Weston and Collobert. One of the first work from Bordes in this domain [4] embeds entities in continuous space and learn relation matrices(two for each relation). Evaluation is done by ranking the object for subject relation pair, evaluating embedding by examining neighbors for entities. In [2] Bordes et. al. tackles the problem of open domain meaning representation. This is achieved by semantic role labeling using Senna Toolkit[6] and then uses energy based approach to minimize the matching energy of triplet(lemmas) to sysnets(Wordnet entities). Semantic Matching Energy function used was earlier discussed in this paper[3] by same authors. Socher et.al.[19] applied similar approach using neural tensors to reason over knowledge bases. Their system returns probability of a fact(triplet) which might be explicitly present or can be inferred. Training function used in this is similar to one used in [6]. The author in a follow up paper[5] model is used for learning new facts in knowledge based on entities not present in KB. This is done by embedding words using[6] and entities in the same latent space.

References

[1] Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. Neural probabilistic language models. In *Innovations in Machine Learning*, pages 137–186. Springer, 2006.

[2] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. Joint learning of words and meaning representations for open-text semantic parsing. In *International Conference on Artificial Intelligence and Statistics*, pages 127–135, 2012.

[3] Antoine Bordes, Xavier Glorot, Jason Weston, and Yoshua Bengio. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, 94(2):233–259, 2014.

[4] Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. Learning structured embeddings of knowledge bases. In *Conference on Artificial Intelligence*, number EPFL-CONF-192344, 2011.

[5] Danqi Chen, Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. *arXiv preprint arXiv:1301.3618*, 2013.

[6] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[7] Stephen Gould, Richard Fulton, and Daphne Koller. Decomposing a scene into geometric and semantically consistent regions. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1–8. IEEE, 2009.

[8] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 873–882. Association for Computational Linguistics, 2012.

[9] RichardSocher JeffreyPennington and ChristopherD Manning. Glove: Global vectors for word representation.

[10] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE, 2009.

[11] Miron Livny, Raghu Ramakrishnan, Kevin Beyer, Guangshun Chen, Donko Donjerkovic, Shilpa Lawande, Jussi Myllymaki, and Kent Wenger. Devise: integrated querying and visual exploration of large datasets. In *ACM SIGMOD Record*, volume 26, pages 301–312. ACM, 1997.

[12] Tomáš Mikolov. *Statistical language models based on neural networks*. PhD thesis, Ph. D. thesis, Brno University of Technology, 2012.

[13] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *INTERSPEECH*, pages 1045–1048, 2010.

[14] Tomas Mikolov, Stefan Kombrink, Lukas Burget, JH Cernocky, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[16] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751. Citeseer, 2013.

[17] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In *Advances in neural information processing systems*, pages 1081–1088, 2009.

[18] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *AISTATS*, volume 5, pages 246–252. Citeseer, 2005.

[19] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934, 2013.

[20] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 935–943. Curran Associates, Inc., 2013.

[21] Richard Socher, Eric H Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y Ng. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *Advances in Neural Information Processing Systems*, pages 801–809, 2011.

[22] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1201–1211. Association for Computational Linguistics, 2012.

[23] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[24] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 129–136, 2011.

[25] Richard Socher, Christopher D Manning, and Andrew Y Ng. Learning continuous phrase representations and syntactic parsing with recursive neural networks. In *Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, pages 1–9, 2010.

[26] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

[27] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[28] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[29] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.