Efficient implementation of GPU 1D convolutions for Natural Language Processing (NLP) in Theano

Darshan Hegde
Center for Data Science
New York University
e-mail: darshan.hegde@cims.nyu.edu

April 14, 2015

Supervisor	Prof. Georg Stadler
Course	MATH-GA 2012.003 / CSCI-GA 2945.003 High performance computing

Proposal

Convolution networks have been applied to visual recognition task with great success [5] [6] by many Deep Learning and Computer Vision researchers. But application of convolution networks for NLP tasks such as sentence modeling [4], document modeling [3] is very recent. Fig 1 shows convolution network used for modeling documents. Each word is represented by a vector (point in a vector space) and they are stacked together to form sentence matrix. Convolution networks provides layers of trainable transformations which converts these sentence matrices in a document to a document vector. The document vector can be used as input to a supervised classifier which can distinguish positive documents from negative. Convolution operations used in many such NLP task are 1D convolutions.

Theano [2] [1], python library used for Deep Learning supports very efficient implementation of 2D convolutions (for visual recognition tasks). The implementation stacks together many 2D images to form a mini-batch and computes convolutions for all the images in a mini-batch at the same time using GPU cores. This ensures that all cores are occupied and results in high throughput. One can implement 1D convolutions in Theano by tweaking input dimensions for 2D convolution implementation. However, this strategy will result in mini-batch size of one, and only one (or few) GPU cores are used in the computation. To ensure high occupancy of GPU cores and hence, high throughput we propose to implement 1D convolutions that support larger mini-batch sizes. The implementation may not be straight forward because sentences / documents are of variable length and load balancing between many thread blocks could an issue.

References

[1] Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[2] James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio. Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*, June 2010. Oral Presentation.

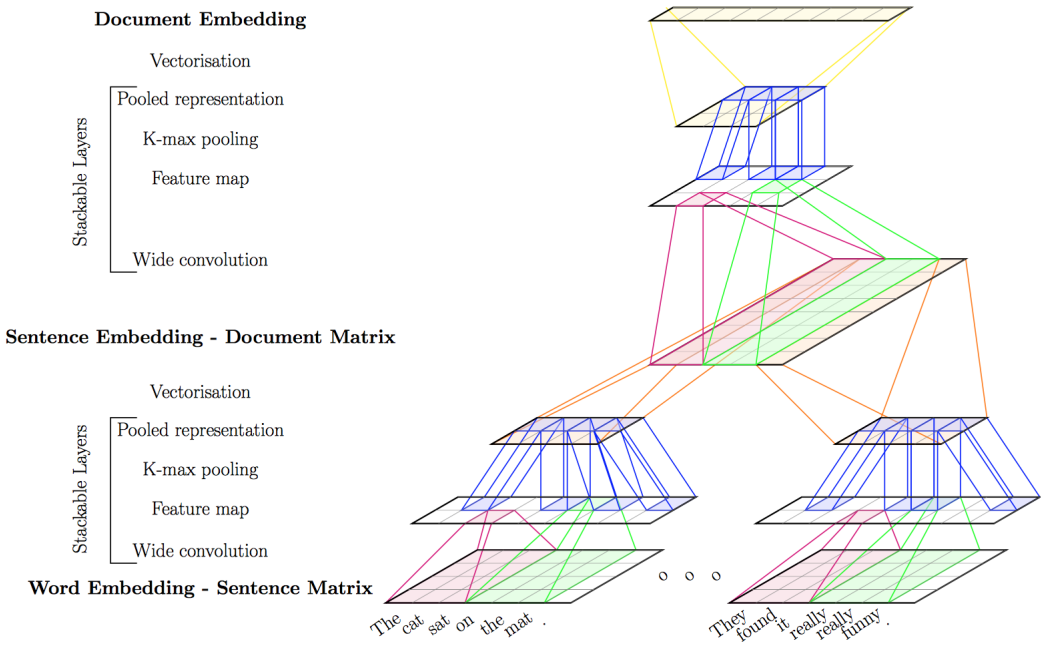

Figure 1: Convolution network for modeling documents. Diagram courtesy [3]

[3] Misha Denil, Alban Demiraj, and Nando de Freitas. Extraction of salient sentences from labelled documents. Technical report, University of Oxford, 2014.

[4] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, June 2014.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[6] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR 2014)*. CBLS, April 2014.