Machine Learning Reading Notes

Brendan Duke

March 26, 2017

1 Definitions

Deep Neural Networks (DNNs) are engineered systems inspired by the biological brain [1].

The **softmax function** is a continuous differentiable version of the argmax function, where the result is represented as a one-hot vector [1, Chapter 6]. Softmax is a way of representing probability distributions over a discrete variable that can take on n possible values.

Formally, softmax is given by Equation 1.

$$\text{softmax}(\boldsymbol{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{1}$$

Mahalanobis Distance

Neighbourhood Components Analysis (NCA) is a method of learning a Mahalanobis distance metric, and can also be used in linear dimensionality reduction [2].

The **PCKh** metric, used by the MPII Human Pose Dataset, defines a joint estimate as matching the ground truth if the estimate lies within 50% of the head segment length [3]. The head segment length is defined as the diagonal across the annotated head rectangle in the MPII data, multiplied by a factor of 0.6. Details can be found by examining the MATLAB evaluation script provided with the MPII dataset.

Non-maximum suppression in object detection, in general, is a set of methods used to prune an initial set of object bounding boxes that may be uncorrelated with the actual object detections in an image, down to a subset that are [4]. In edge detection, non-maximum suppression is used to suppress any pixels (i.e. not include them in the set of detected edges) that are not the maximum response in their neighbourhood.

LSTM (Long Short Term Memory) neural networks are a type of recurrent neural network whose characteristic feature is the presence of a gated self-loop that allows retention of its "cell state", which are the pre-non-linearity activations of the previous time step [1, Chapter 10].

Cell state is updated at each time step according to Equation 2.

$$s_i^{(t)} = f_i^{(t)} s_i^{(t-1)} + g_i^{(t)} \sigma \left(b_i + \sum_j U_{i,j} x_j^{(t)} + \sum_j W_{i,j} h_j^{(t-1)} \right) \tag{2}$$

The vectors $\boldsymbol{f}^{(t)}$ and $\boldsymbol{g}^{(t)}$ in Equation 2 also take inputs from $\boldsymbol{x}^{(t)}$ and $\boldsymbol{h}^{(t-1)}$, with their own weight tensors $\boldsymbol{U}^f$ and $\boldsymbol{W}^f$, $\boldsymbol{U}^g$ and $\boldsymbol{W}^g$, respectively.

Similar gate functions exist to gate the inputs and outputs to the LSTM, as well.

2 Paper Summaries

2.1 DeepPose: Human Pose Estimation via Deep Neural Networks [5]

This paper uses DNNs as a method for human pose estimation, based on the success of [6] and [7] for object detection using DNNs.

This is in contrast to the existing work in human pose estimation at the time, which focused on explicitly designed pose models. Papers about these methods can be found in the "Related Work" section of [5].

The input to the 7-layered convolutional DNN (based on AlexNet [8]) is the full image.

2.2 Dropout: A Simple Way to Prevent Neural Networks from Overfitting [9]

Dropout is a technique used to overcome the problem of overfitting in deep neural nets with large numbers of parameters. The idea is to train using many "thinned" networks, chosen by randomly removing subsets of units and their connections. The predictions from the thinned networks are approximately averaged at test time by using a single, unthinned, network with reduced weights.

- Existing regularization methods: stopping training as soon as validation error stops improving, L1 and L2 regularization, and weight sharing [10].

2.3 End-to-end people detection in crowded scenes [11]

This paper is focused on jointly creating a set of bounding-box predictions for people in crowded scenes, in such a way that post-processing steps such as Non-maximum suppression are not necessary.

A recurrent LSTM layer is used in a model that is trained end-to-end, with a new loss function that operates on sets of bounding-box predictions. A new training set for human detection in crowded scenes, called "Brainwash", is produced.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, "Deep learning," 2016, book in preparation for MIT Press. [Online]. Available: http://www.deeplearningbook.org

[2] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in Neural Information Processing Systems 17*, L. K. Saul, Y. Weiss, and L. Bottou, Eds. MIT

Press, 2005, pp. 513–520. [Online]. Available: http://papers.nips.cc/paper/2566-neighbourhood-components-analysis.pdf

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[4] R. Rothe, M. Guillaumin, and L. J. V. Gool, "Non-maximum suppression for object detection by passing messages between windows," in *Computer Vision - ACCV 2014 - 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I*, 2014, pp. 290–306. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16865-4_19

[5] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013. [Online]. Available: http://arxiv.org/abs/1312.4659

[6] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 2553–2561. [Online]. Available: http://papers.nips.cc/paper/5207-deep-neural-networks-for-object-detection.pdf

[7] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[9] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[10] S. J. Nowlan and G. E. Hinton, "Simplifying neural networks by soft weight-sharing," *Neural Comput.*, vol. 4, no. 4, pp. 473–493, Jul. 1992. [Online]. Available: http://dx.doi.org/10.1162/neco.1992.4.4.473

[11] R. Stewart and M. Andriluka, "End-to-end people detection in crowded scenes," *CoRR*, vol. abs/1506.04878, 2015. [Online]. Available: http://arxiv.org/abs/1506.04878