Written Review: DanQ

BY OLIVIER MOINDROT, CHARLES LU, CHARLES BOURNHONESQUE, DAVID GOLUB
omoindrot@stanford.edu, charleslu@stanford.edu, charlesbour@gmail.com, golubd@stanford.edu

CS 273B Autumn 2016

Paper: Quang, Daniel, and Xiaohui Xie. "DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences." Nucleic acids research (2016): gkw226.

1 Overview

In DanQ, the authors create a novel neural network architecture that uses a single-layer convolutional neural network and a max pooling layer, an RNN layer, and a sequence of fully connected layers for the task of TF binding. Specifically:

1. **Input**: The model takes as input a length 1000 one-hot encoded DNA sequence.

2. **Convolution layer**: The input is fed into a CNN with 320 kernels of length 26 with ReLU activation.

3. **Max pooling**: Outputs from convolution are fed into a max pooling layer of size 13, with dropout applied with 20% probability.

4. **Bidirectional RNN**: An LSTM is used in both directions of the output from max pooling (length 75), with hidden size 320. Dropout with 50% probability at the end.

Fully connected layers: The final two layers of the model are dense layers with ReLU nonlinearity, and a multi-task sigmoid output for 919 targets (from intersecting ChIP-seq and DNase-seq peak sets).

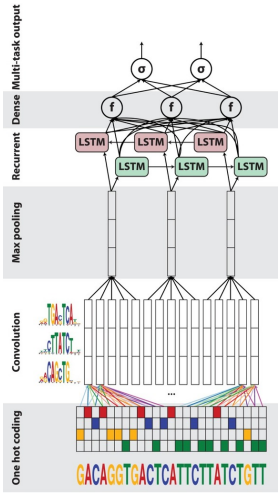

Figure 1.1. DanQ model architecture.

2 Key findings

The proposed architecture significantly improves state-of-the-art performance over the DeepSea model[2.1] and a baseline logistic regression model as measured by the PR-AUC metric. The authors also converted the kernels learned by the DanQ models to motifs and demonstrated that from 320 motifs learned by their model, over half match those already well-known in literature.

Furthermore, Quang and Xie trained a second model with an increased number of convolution kernels (1024) and initialized the weights of half of them to known motifs. This model achieved even better results.

3 Interesting ideas

The authors propose interesting ideas for both the model architecture and evaluation metrics.

For the model architecture, they only use **one convolutional layer** and introduce a **recurrent neural network** architecture on top of the convolutional layer, which can potentially model **long-term dependencies** in a much better way than stacked convolutional neural networks, as is empirically demonstrated by results.

Quang and Xie also propose the **PR-AUC** metric for model evaluation, which is less influenced by the class imbalance issue than the ROC-AUC predictor (the fact that there are many more negative samples than positive samples for binding).[3.1] The PR-AUC metric is able to pinpoint weaknesses in models that the ROC-AUC metric misses. While the logistic regression baseline model appears to have good performance in terms of its ROC-AUC, it in fact achieves less than 5% PR-AUC under the GM12878 EBF1 and H1-hESC SIX5 targets, which demonstrates how the class imbalance issue can inflate performance.

4 Weaknesses and extensions

Similar to the DeepSea model, DanQ only uses DNA sequence information for predictive modeling, which limits generalization to unseen cell types, which may have the same sequence but different binding patterns. One extension that can fix this is to include more information that depends on varying cell types such as methylation and DNase I footprinting.

Another weakness of the paper is that it is unclear with respect to how the dataset was produced. Specifically, the authors create the target labels by "intersecting 919 ChIP-seq and DNase-seq peak sets", which is not clear and should be elaborated on.

Furthermore, the current architecture only works with a fixed sequence length of 1000; potentially exploring new fully-recurrent architectures may allow the model to generalize to arbitrary sequence lengths and thus incorporate more contextual information that may contain useful signals for predicting bindings.

One possible extension to address this would be to use a different model inspired by WaveNet,[4.1] which is a new state-of-the-art model for speech synthesis. WaveNet uses a very deep fully convolutional network on raw audio input (16,000 samples/second) and manages to model the **long-term dependencies** using atrous convolutions ("à trous" meaning "with holes" in French): when we get into deeper layers, they gather information from farther away to have a more global sense of what is happening.

2.1. Zhou, Jian, and Olga G. Troyanskaya. "Predicting effects of noncoding variants with deep learning-based sequence model." Nature methods 12.10 (2015): 931-934.

3.1. https://www.kaggle.com/forums/f/15/kaggle-forum/t/7517/precision-recall-auc-vs-roc-auc-for-class-imbalance-problems/41179

4.1. Oord, Aaron van den, et al. "WaveNet: A Generative Model for Raw Audio." arXiv preprint arXiv:1609.03499 (2016). https://deepmind.com/blog/wavenet-generative-model-raw-audio/

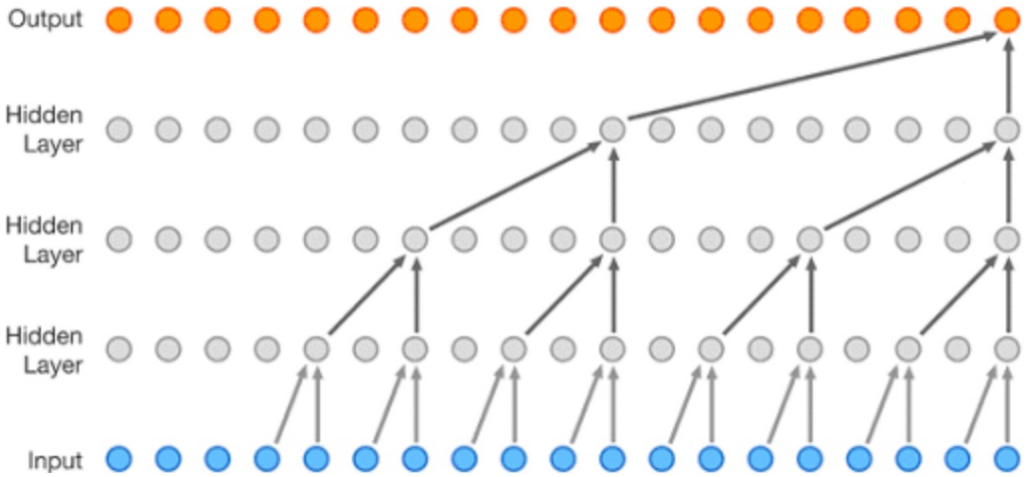

Figure 4.1. WaveNet model architecture.