Abstract
Sparse connectivity is an important factor behind the success of convolutional neural networks and recurrent neural networks. In this paper, we consider the problem of learning sparse connectivity for feedforward neural networks (FNNs). The key idea is that a unit should be connected to a small number of units at the next level below that are strongly correlated. We use Chow-Liu’s algorithm to learn a tree-structured probabilistic model for the units at the current level, use the tree to identify subsets of units that are strongly correlated, and introduce a new unit with receptive field over the subsets. The procedure is repeated on the new units to build multiple layers of hidden units. The resulting model is called a TRF-net. Empirical results show that, when compared to dense FNNs, TRF-net achieves better or comparable classification performance with much fewer parameters and sparser structures. They are also more interpretable.

1 Introduction
It is common knowledge that architecture matters in deep learning. In general, convolutional neural networks (CNNs) are considered suitable for spatial data and recurrent neural networks (RNNs) are considered suitable for sequential data. However, it remains an art to design an appropriate architecture for a particular application. Typically, researchers need to evaluate a long list of candidate architectures before finding a satisfactory one, a process that is sometimes called graduate student descent. For this reason, there is growing interest in architecture learning [Baker et al., 2017; Zoph and Le, 2017].

In addition to saving manual labor, architecture learning is interesting for two other reasons. First, it enables us to detect salient patterns in data and represent them using network structures. This can lead to better performing and more interpretable models. Second, architecture learning naturally would involve connectivity learning, which leads to sparse models. Model sparsity is helpful in avoiding overfitting and is desirable for applications on hand-held devices.

*Equal contribution.

Figure 1: Example of local structure in 20newsgroup dataset. Three local neighborhoods are shown: video card, os, file format.
tistical structure, in place of spatial or sequential structure, is another promising way to achieve sparse connectivity.

We propose the following method for learn sparse FNNs: (1) Run the Chow-Liu algorithm [Chow and Liu, 1968] to learn a tree-structured probabilistic graphical models over the observed features; (2) use the tree to identify subsets of correlated features; introduce a new neuron for each subset; convert the newly introduced neurons into observed variables; and repeat (1) and (2) to build multiple hidden layers. To capture global patterns, we allow a small number of neurons to be connected to all units at the level below. We call deep models built by our method Tree-Receptive-Field networks (TRF-nets). Here are the contributions that we make in this paper:

- We propose a novel algorithm for learning sparse FNNs.
- We have conducted extensive experiments to compare TRF-nets with dense FNNs. The results show that the TRF-nets achieve better or comparable classification performance with much fewer parameters and sparser structures, and they are more interpretable.
- We have also empirically compared TRF-nets with sparse FNNs obtained by pruning dense FNNs and models obtained by regularization-based methods. Overall, the TRF-nets outperform other models and have higher interpretability.

2 Related Works

The primary goal in structure learning is to find a model with optimal or close to optimal generalization performance. Bruce-force search is not feasible because the search space is large and evaluating each model is costly as it necessitates training. Early works in the 1980’s and 1990’s have focused on what we call the micro expansion approach where one starts with a small network and gradually add new neurons to the network until a stopping criterion is met [Kwok and Yeung, 1997b; 1997a; Bello, 1992; Ash, 1989; Fahlman and Lebiere, 1990]. The word “micro” is used here because at each step only one (or a few) neurons are added. This makes learning a large model computationally difficult as reaching a large model would require many steps and model evaluation is needed at each step. In addition, those early methods typically do not produce layered structures that are commonly used today. Note that, in another line of work, micro expansion methods have been developed for sum-product networks [Hsu et al., 2017].

Recent efforts have concentrated on what we call the contraction approach where one starts with a larger-than-necessary structure and reduce it to the desired size. Contraction can be done either by repeatedly pruning neurons and/or connections [LeCun et al., 1990; Hassibi et al., 1993; Han et al., 2015; Guo et al., 2016; Srivinas and Babu, 2015; Li et al., 2017], or by using regularization to force some of the weights to zero [Collins and Kohli, 2014; Wen et al., 2016]. From the perspective of structure learning, the contraction approach is not ideal because it requires a complex model as input. A key motivation for a user to consider structure learning is to avoid building models manually.

A third approach is to explore the model space stochastically. One way is to place a prior over the space of all possible structures and carry out MCMC sampling to obtain a collection of models with high posterior probabilities [Adams et al., 2010]. Another way is to encode a model structure as a sequence of numbers, use a reinforcement meta model to explore the space of such sequences, learn a good meta policy from the sequences explored, and use the policy to generate model structures [Baker et al., 2017; Zoph and Le, 2017]. An obvious drawback of such stochastic exploration method is that they are computationally very expensive.

What we propose in this paper is a macro expansion method where we start from scratch and repeatedly add layers of hidden units until a threshold is met. Our method is computationally cheaper than the micro expansion approach because it evaluates one two-layer model for each layer of the final model. It also produces layered models. Note that, while parameters are trained layer by layer in deep belief network and deep Boltzmann machines, we learn model structures layer by layer. The learned model is trained as a whole via backpropagation.

Another macro expansion method has recently been proposed by Liu et [Liu et al., 2017]. It learns a new layer by solving a difficult multi-objective optimization problem. In contrast, we run Chow-Liu’s algorithm to build a tree among the units on the current layer, which gives a “spatial structure” among the units, and we build the next layer by applying the concept of receptive field from CNNs on the tree.

3 Method

In this section, we present our method for learning sparse FNNs.

3.1 Learning the Tree Structures

We start by learning a tree-structured probabilistic graphical model for the observed variables. Fig. 2(a) illustrates one such model. Each node in the model represents an observed variable. We assume the observed variables are discrete. In other words, we deal with a tree-structured Bayesian network here. The model defines a joint distribution over all the observed variables:

\[ p(x|\theta) = \prod_{x_t} p(x_t|p(a x_t)), \]

where \( T \) denote the tree structure and \( \theta \) denotes the model parameters.

Given a dataset \( D \), the log-likelihood of the parameters \( \theta \)
and introduce a latent node for each of them. A latent node is connected to all the nodes its receptive field. After latent node introduction, edges between observed nodes are removed.

A receptive field over a tree is defined by a field center $x$, which is one of the nodes, and a field radius $r$. It includes $x$ and all nodes reachable from $x$ in $r$ hops. To cover a tree with receptive fields, we need another parameter, stride $s$, which is the number of hops between the centers of two neighboring receptive fields. Given $r$ and $s$, we create the first receptive field by randomly picking a node as its center. At each step after that, we pick the center for the next receptive field by identifying a node that is $s$ hops away one of the existing centers. The process terminates when such a node no long exists.

Latent nodes introduced by the method described so far are meant to model local interactions among variables. There could be long-range interactions that they do not capture. Such interactions are indirectly modelled by multiple layers latent nodes. In this paper, we also consider modeling long-range interactions directly. Specifically, we introduce a small number of global neurons and have them connected to all observed nodes. As such, a two-layer structure built by our method looks like what is shown in Fig. 3. We use Build2LayerStructure to denote the procedure described in this subsection.

### 3.3 Training Two-Layer TRF-nets

After obtaining a two-layer structure, we next turn it into a neural network model and train it as denoising autoencoder.

Let $x = [x_i]$ and $h = [h_j]$ be the vectors of visible and hidden units respectively. And let $A = [a_{ij}]$ be the connectivity matrix where $a_{ij} = 1$ if unit $x_i$ is connected to unit $h_j$ and 0 otherwise. Let $W$ and $b$ be the weight and bias parameters of the neural network.

The model defines a conditional distribution $p(h|x)$ of the hidden units given the observed units:

$$p(h = 1|x) = s((A \circ W)x + b),$$  \hspace{1cm}(5)$$

where $s$ is the probability function, e.g. sigmoid, and $\circ$ denotes element-wise product. The model also defines a conditional distribution $p(x|h)$ of the observed units given the hidden units:

$$p(x|h) = \text{Ber}(\sigma((A \circ W)'h + b)) \quad \text{if } x \text{ is binary}$$

\[p(x|h) = N((A \circ W)'h + b, \lambda I) \quad \text{if } x \text{ is real-value},\]  \hspace{1cm}(6)$$

where $(A \circ W)'$ is the transpose of $A \circ W$.

The model weights are determined using gradient descent and the denoising criteria [Vincent et al., 2008] is used as the
Algorithm 1: TRF-N

\begin{algorithm}
\caption{TRF-N$(D, r, s, d)$}
\begin{algorithmic}
\State \textbf{Input}: $D$—a collection of data, \textit{k}—kernel size, \textit{s}—stride, \textit{d}—depth
\State \textbf{Output}: \textit{m}—TRF-net
\State $D_1 \leftarrow D, m \leftarrow \text{null}$
\Repeat
\State $T \leftarrow \text{CHOWLIU}(D_1)$
\State $S \leftarrow \text{BUILD2LAYERSTRUCTURE}(T, r, s)$
\State $m_1 \leftarrow \text{TRAINDENOISING}(S, D_1)$
\State $D_1 \leftarrow \text{PROJECDATA}(m_1, D_1)$
\State $m \leftarrow \text{STACKMODELS}(m, m_1)$
\Until \text{the number of layers in } \textit{m} \text{ reaches } \textit{d}
\State \Return $m$
\end{algorithmic}
\end{algorithm}

The overall algorithm of learning TRF-net is shown in Algorithm 1. The resulting sparse FNN is then trained using backpropagation for upstream tasks.

4 Experiments

We have carried out experiments to determine whether our method can produce high quality model structures. The TRF-nets produced by our method are compared with FNNs obtained by manual grid search, and sparse FNNs obtained by weight pruning and weight regularization [Han et al., 2015; Collins and Kohli, 2014]. The micro expansion and stochastic exploration methods are not included in the comparisons because they do not produce layered FNNs and they are extremely expensive computationally. The macro expansion method by [Liu et al. 2017] would be a good baseline. Unfortunately, we were able to obtain its implementation.

4.1 Datasets

Six publically available datasets were used in the experiments. One of them is about chemical compounds, and the other five are text data.

- **Tox21 challenge dataset.** There are about 12,000 environmental chemical compounds in the dataset. The tasks are to predict 12 different toxic effects. We treat them as 12 binary classification tasks. We filter out sparse features which are present in fewer than 5% of the compounds, and rescale the remaining 1,644 features to zero mean and unit variance. We report the average AUC as the results for this dataset.

- **Text classification datasets.** The five text datasets from [Zhang et al., 2015]: AG’s News, DBPedia, Yelp Review, Yahoo! Answer and Sogou News. The number of classes ranges from 2 to 14 and the dataset sizes are around 130,000~1,400,000. Stop words were removed during preprocessing. The top 10,000 most frequent words are selected as the vocabulary, and each document is represented as a bag-of-words.

4.2 Experiment Setup

We compare TRF-nets with standard fully-connected feed-forward neural networks (FNNs), sparse neural networks obtained by weight pruning (Pruned FNNs) [Han et al., 2015], and FNNs with L1 weight regularization (Reg FNNs) [Collins and Kohli, 2014].

When running the TRF-net algorithm on the Tox21 dataset, we used receptive fields with radius $r = 5$. Two different values, 1 and 2, were used for the stride $s$. When $s = 1$, the number of units on the next layer is the same as the number of units on the current layer. When $s = 2$, the number of units on the next layer is reduced by half. For text data, there are 10,000 input units. So, we set $r = 6$ and $s = 5$ for the first layer, $r = 5$ and $s = 4$ for the second layer to quickly reduce the number of units. The number of global neurons introduced at each layer is 10% of the number of TRF neurons on the same layer.

In the case of FNNs, the number of layers and the number of hidden units were determined by manual grid search. We followed [Klambauer et al., 2017] and chose the number of neurons for each layer from \{512, 1024, 2048\}, and tested both rectangular and conic structures. We considered up to 4 hidden layers. For Pruned FNNs, we took the best FNN as the initial model and performed weight pruning as described in [Han et al., 2015]. The pruned model is then retrained to obtain the final model. For Reg FNNs, we added L1 norm term to the training loss with the regularization strength found through validation (around 1e-5).

In all cases, we used ReLUs [Nair and Hinton, 2010] as the non-linear activation functions, and Adam [Kingma and Ba, 2014].
Table 1: Comparison between TRF-nets, FNNs, Pruned FNNs and Reg FNNs on 6 classification datasets. The structures of FNNs are chosen by using validation data. Each experiment is run for three times.

<table>
<thead>
<tr>
<th>Task</th>
<th>TRF-net</th>
<th>FNN</th>
<th>Pruned FNN</th>
<th>Reg FNN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Parameter #</td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Tox21 Average</td>
<td>0.8135 ± 0.0038</td>
<td>0.22M / 10.33%</td>
<td>0.8010 ± 0.0017</td>
<td>0.7998 ± 0.0034</td>
</tr>
<tr>
<td>AG’s News</td>
<td>0.9180 ± 0.05%</td>
<td>1.62M / 6.93%</td>
<td>0.9161 ± 0.01%</td>
<td>0.9149 ± 0.09%</td>
</tr>
<tr>
<td>DBPedia</td>
<td>0.9802 ± 0.01%</td>
<td>3.50M / 13.90%</td>
<td>0.9799 ± 0.04%</td>
<td>0.9795 ± 0.02%</td>
</tr>
<tr>
<td>Yelp Review Full</td>
<td>0.5914 ± 0.06%</td>
<td>3.23M / 13.85%</td>
<td>0.5913 ± 0.14%</td>
<td>0.5883 ± 0.01%</td>
</tr>
<tr>
<td>Yahoo!Answer</td>
<td>0.7140 ± 0.01%</td>
<td>2.27M / 10.61%</td>
<td>0.7184 ± 0.07%</td>
<td>0.7174 ± 0.05%</td>
</tr>
<tr>
<td>Sogou News</td>
<td>0.9629 ± 0.07%</td>
<td>2.27M / 10.10%</td>
<td>0.9611 ± 0.06%</td>
<td>0.9620 ± 0.06%</td>
</tr>
</tbody>
</table>

Table 2: Comparison between TRF-nets with global neurons and TRF-nets with TRF neurons only

<table>
<thead>
<tr>
<th>Task</th>
<th>TRF-net</th>
<th>TRF only</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Tox21 Average</td>
<td>0.8135</td>
<td>0.8196</td>
</tr>
<tr>
<td>AG’s News</td>
<td>0.9180%</td>
<td>0.9100%</td>
</tr>
<tr>
<td>DBPedia</td>
<td>0.9802%</td>
<td>0.9775%</td>
</tr>
<tr>
<td>Yelp Review Full</td>
<td>0.5914%</td>
<td>0.5869%</td>
</tr>
<tr>
<td>Yahoo!Answer</td>
<td>0.7140%</td>
<td>0.7043%</td>
</tr>
<tr>
<td>Sogou News</td>
<td>0.9629%</td>
<td>0.9603%</td>
</tr>
</tbody>
</table>

2015] as the optimizer, and we applied dropout [Srivastava et al., 2014] with rate 0.5. We ran all the experiments for three times and report the average results.

4.3 Results

Classification Performances

Table 1 shows the classification performances of TRF-nets, FNNs, Pruned FNNs and Reg FNNs on different datasets. It is clear that TRF-nets achieved better AUC scores than FNNs on the Tox21 dataset and better or comparable accuracy on 5 text classification datasets. They did so with significantly fewer parameters and sparser structures. In comparison with FNNs, the sparsity of TRF-nets (# of connections w.r.t that of fully-connected ones) is around 10% for all datasets. This confirms that our method is able to achieve high generalization performance with much fewer parameters and sparser structures. Compared with Pruned FNNs and Reg FNNs, TRF-nets also achieved better or comparable classification accuracy. Pruned FNNs have comparable amounts of model parameters as TRF-nets. Reg FNNs are not sparse models. Their weights are pushed toward 0 by regularization, but few actually reached 0. With weights of absolute values less than 0.001 removed, their sparsity is between 20% and 30% in comparison with FNNs. Note that pruned FNNs and Reg FNNs were obtained from FNNs that took much manual search to construct. In contrast, TRF-nets were automatically and directly learned from data.

Contribution of the Tree Receptive Fields

To validate our assumption that the TRF neurons in TRF-nets capture most of the patterns in data, we performed another set of experiments with all the global neurons removed. Table 2 shows the results. As we can see, the classification performances degraded only slightly, while model sparsity is further improved significantly. The results not only show the importance of the TRF neurons in TRF-nets, but also shows that our structure learning method is effective.

Influence of the Model Depth

The results reported in Tables 1 and 2 were obtained using TRF-nets with 3 hidden layers. The decision to use 3 hidden layers was influenced by the literature. Typically authors use only a small number of hidden layers when it comes to FNNs. Nonetheless, it would be interesting to see how model depth influences the performance TRF-nets. So, we performed a set of experiments with varying model depths. The results are shown in Table 3. It can be seen that our method continues to work when the network become deeper. For example, for the AG’s News dataset, the best performance was achieved with a 5-layer TRF-net. In contrast, deep architecture might lead to severe overfitting for FNNs.

Computational Time

A comparison of the computational time for TRF-nets and Pruned FNNs is given in Table 4. As it can be seen, the structure learning phase of TRF-nets costs almost as much time as the finetuning phase. The overall computational time is comparable with Pruned FNNs without structure searching. Note that FNNs, Pruned FNNs and Reg FNNs need to determine the best-performing FNN structure by grid searching in the pre-defined structure space, as described in the experiment setup. Pruned FNNs further prune and finetune the best structure. Although for TRF-nets there are also hyperparameters, i.e. receptive field and stride, it is significantly easier to determine those controlling hyperparameters than to determine the actual structures. Therefore, in practice the overall computational time including grid searching over the structure space for FNNs, Pruned FNNs and Reg FNNs is significantly larger than that for TRF-nets. All neural networks are trained on a Tesla K20 GPU, while the structure learning is run on a CPU.

Interpretability

Next we compare TRF-nets with FNNs and Pruned FNNs in terms of the interpretability of their hidden unit. We characterize the “meaning” of a hidden unit using the top 10 words that have the strongest correlations with the unit. Following [Chen et al., 2017], we measure the “interpretability” of a hidden unit by considering how similar pairs of words in the top-10 list are using a word2vec model [Mikolov et al., 2013a;
5 Conclusions

Structure learning for deep neural network is a challenging and interesting research problem. We have proposed an unsupervised structure learning method, which first learns a tree-structured probabilistic graphical model and then constructs hidden neurons with local receptive field over the tree. The resulting TRF-nets have shown to achieve better or comparable classification performance in all kinds of tasks compared with standard FNNs, while containing significantly fewer parameters. In addition, TRF-nets have also shown to be more interpretable than FNNs, which is interesting because interpretability is an important issue in deep learning.

Acknowledgments

Research on this article was supported by Hong Kong Research Grants Council under grant 16212516.

References


[Bello, 1992] Martin G Bello. Enhanced training algorithms, and integrated training/architecture selection for multi-

Table 3: Precision of TRF-nets with different depths on classification datasets. For each task, best result is bolded.

<table>
<thead>
<tr>
<th>Task</th>
<th>TRF-net</th>
<th>Pruned FNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tox21 Average</td>
<td>0.8139 ± 0.0026</td>
<td>0.8081 ± 0.0063</td>
</tr>
<tr>
<td>AG’s News</td>
<td>0.8135 ± 0.0038</td>
<td>0.8025 ± 0.0091</td>
</tr>
<tr>
<td>DBPedia</td>
<td>4.082 ± 0.0009</td>
<td>4.081 ± 0.0063</td>
</tr>
<tr>
<td>Yelp Review Full</td>
<td>7.215 ± 0.0048</td>
<td>7.200 ± 0.0010</td>
</tr>
<tr>
<td>Yahoo!Answer</td>
<td>4.311 ± 0.0010</td>
<td>3.812 ± 0.0015</td>
</tr>
</tbody>
</table>

Table 4: Comparison of computational time (s) between TRF-nets and Pruned FNNs on 6 classification datasets.

<table>
<thead>
<tr>
<th>Task</th>
<th>TRF-nets</th>
<th>Pruned FNNs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tox21 Average</td>
<td>145</td>
<td>43</td>
</tr>
<tr>
<td>AG’s News</td>
<td>3,388</td>
<td>3,956</td>
</tr>
<tr>
<td>DBPedia</td>
<td>4,082</td>
<td>6,562</td>
</tr>
<tr>
<td>Yelp Review Full</td>
<td>5,072</td>
<td>12,440</td>
</tr>
<tr>
<td>Yahoo!Answer</td>
<td>7,215</td>
<td>7,340</td>
</tr>
<tr>
<td>Sogou News</td>
<td>4,311</td>
<td>4,311</td>
</tr>
</tbody>
</table>

Table 5: Interpretability scores of TRF-nets, FNNs and Pruned FNNs on different datasets.

Table 6: Qualitative interpretability results of hidden units in TRF-nets. Each line corresponds to one hidden unit.

Task | Characterization Words
--- | ---
Yelp Review Full | tasteless unseasoned flavorless paprika panko crusts unagi crumb vindaloo tortas spicy wink drapes
DBPedia | album songwriting chet saxophone hurling backstroke badminton skier journalists hardcover editors
Yahoo!Answer | xp desktop adobe peut cpu pagans atheist mormons apostle tylenol diabetic acids kidneys
AG’s News | mozilla mainframe microprocessors republicans prosecutor argument jfk noted furious harsh concessions


