
Deep Networks on Graph-Structured Data

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053
Anonymous Author(s)

Affiliation
Address
email

Abstract

1 Experiments

In order to measure the performance of spectral networks on real-world data and to explore the effect of the graph estimation procedure, we performed experiments on three datasets from computer vision, text categorization and computational biology.

1.1 Reuters

We used the same version of the Reuters dataset as in [?], which consists of training and test sets each containing 201,369 documents from 50 mutually exclusive classes. Each document is represented as a log-normalized bag of words for 2000 common non-stop words. As a baseline we used the fully-connected network of [?] with two hidden layers consisting of 2000 and 1000 hidden units regularized with dropout.

We based the spectral network architecture on that of a classical convolutional network, namely by interleaving graph convolution and graph pooling layers and ending with a fully connected layer. Performing pooling at the beginning of the network was especially important to reduce the dimensionality in the graph domain and alleviate the expensive Graph Fourier Transform operation. We chose hyperparameters by performing initial experiments on a validation set consisting of one-tenth of the training data. All architectures were then trained using the same hyperparameters, which can be found in the Appendix. Since the experiments were computationally expensive, we did not train all models until full convergence. This enabled us to explore more model architectures and obtain a clearer understanding of the effects of graph construction.

Graph	Architecture	Parameters	Accuracy (epoch 200)	Accuracy (epoch 200)
-	FC-2000-1000-50	$8 \cdot 10^6$	70.2	70.2
Supervised	GC4-P4-FC-1000-50	$2 \cdot 10^6$	69.41	70.0
Supervised	GC8-P8-FC-1000-50	$2 \cdot 10^6$	69.15	-
Supervised	GC16-P4-GC16-P4-FC-1000-50	$2 \cdot 10^6$	68.68	-
Supervised	GC64-P8-GC64-P8-FC-1000-50	$2 \cdot 10^6$	68.63	-
Gaussian Kernel	GC4-P4-FC-1000-50	$2 \cdot 10^6$	64.14	-
Gaussian Kernel	GC8-P8-FC-1000-50	$2 \cdot 10^6$	64.79	-
Gaussian Kernel	GC16-P4-GC16-P4-FC-1000-50	$2 \cdot 10^6$	-	-
Gaussian Kernel	GC64-P8-GC64-P8-FC-1000-50	$2 \cdot 10^6$	-	-
Gaussian Kernel (local scaling)	GC4-P4-FC-1000-50	$2 \cdot 10^6$	68.56	-

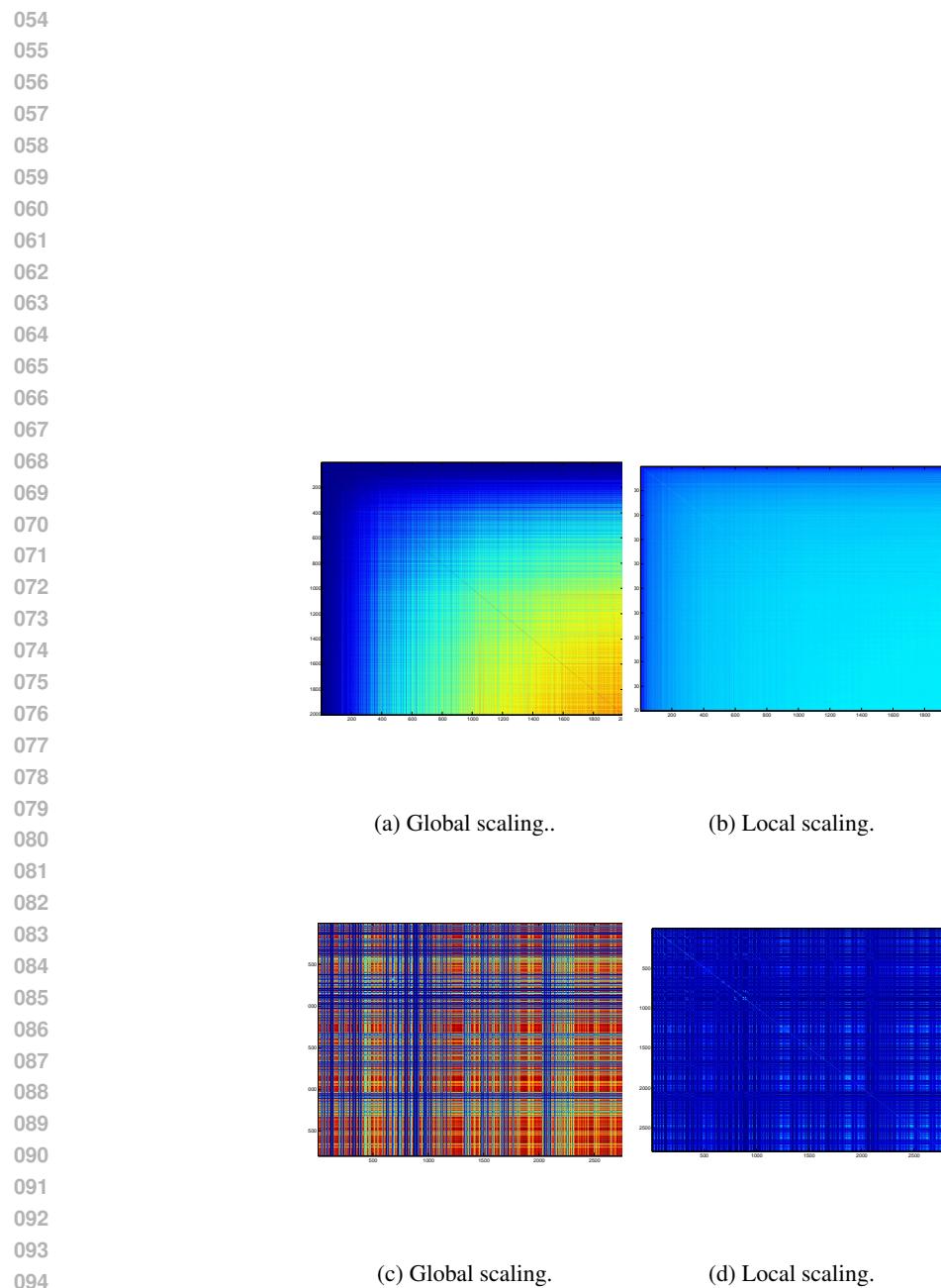


Figure 1: Similarity graphs for the Reuters (top) and Merck DPP4 (bottom) datasets.

108
109
110 1.2 Merck Molecular Activity Challenge
111

112 The Merck Molecular Activity Challenge is a computational biology benchmark where the task is to
113 predict activity levels for various molecules based on the distances in bonds between different atoms.
114 For our experiments we used the DPP4 dataset which has 8193 samples and 2796 features. We chose
115 this dataset because it was one of the more challenging and was of relatively low dimensionality
116 which made the spectral networks tractable. As a baseline architecture, we used the state-of-the-art
117 network of [?] which has 4 hidden layers and is regularized using dropout and weight decay.

118 As before, we used one-tenth of the training data to find good hyperparameters and explore archi-
119 tectures.

Graph	Architecture	Parameters	Test R^2
-	FC-4000-2000-1000-1000-50	$22.1 \cdot 10^6$	0.2728
Supervised	GC64-P8-GC64-P8-1000-1000-1	$3.8 \cdot 10^6$	0.2629
Gaussian Kernel, global scaling)	GC64-P8-GC64-P8-1000-1000-1	$3.8 \cdot 10^6$	0.1992
Gaussian Kernel, local scaling)	GC64-P8-GC64-P8-1000-1000-1	$3.8 \cdot 10^6$	-

124
125 **126 1.3 Imagenet**
127

128 In the experiments above our graph construction relied on an approximate estimation from the data.
129 To measure the influence of the graph construction compared to the filter learning in the frequency
130 domain, we performed the same experiments on the ImageNet dataset for which the graph is already
131 known, namely it is the 2-D grid. The spectral network was thus a convolutional network whose
132 weights were defined in the frequency domain. Training was performed exactly as in Figure 1,
133 except that the linear transformation was a Fast Fourier Transform.

134 Our network consisted of 4 convolution/ReLU/max pooling layers with 48, 128, 256 and 256 feature
135 maps, followed by 3 fully-connected layers with 4096 hidden units each regularized with dropout.
136 We trained two versions of the network: one classical convolutional network and one as a spectral
137 network where the weights were defined in the frequency domain only and were interpolated using
138 a spline kernel. Both networks were trained for 40 epochs over the ImageNet dataset where input
139 images were scaled down to 128×128 to accelerate training.

Graph	Architecture	Parameters	Test Accuracy (Top 5)	Test Accuracy (Top 1)
2-D Grid	Convolutional Network	$3.8 \cdot 10^6$	71.854	46.24
2-D Grid	Spectral Network	$3.8 \cdot 10^6$	71.998	46.71

145 We see that both models yield nearly identical performance. Interestingly, the spectral network learns
146 faster than during the first part of training, although both networks converge around the same time.
147 This requires further investigation.

148
149 **References**
150

151
152
153
154
155
156
157
158
159
160
161

```
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174  
175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215
```

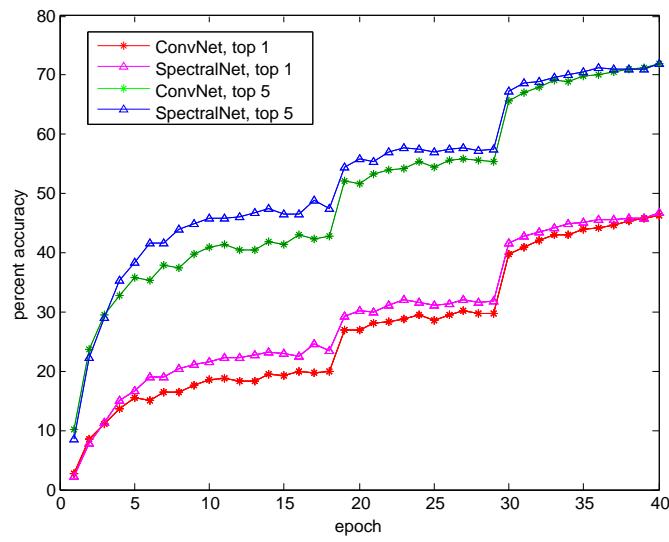


Figure 2: ConvNet vs. SpectralNet on ImageNet.