Improving neural networks by preventing co-adaptation of feature detectors

Published by:

Presented by:
Melvin Laux
Outline

- Introduction
  - Model Averaging
  - Dropout
- Approach
- Experiments
- Conclusion
Model Averaging

- Model Averaging
  - Try to prevent overfitting
  - Train multiple separate neural networks
  - Apply each network on test data
  - Use average of all results

- Problem: Computationally expensive during training AND testing
  - Fast model averaging (using Dropout)
What is “dropout”? 

- Randomly drop half of the hidden units:
  - Prevents complex co-adaption on training data
  - Hidden units can no longer “rely” on others
  - Each neuron has to learn a generally helpful feature

- On every presentation of each training case:
  - Each hidden unit has 50% chance of being “dropped out” (omitted)

- On every presentation of each training case, a different network is trained (most likely) which all share the same weights
  - Allows to train a huge amount of networks in a reasonable time
Outline

- Introduction

- Approach
  - Training
  - Testing

- Experiments

- Conclusion
Training

- Stochastic gradient descent
- Mini-Batches
- Cross-entropy objective function

- Modified penalty term:
  - Set upper bound on L2-norm for the incoming weight vector of each hidden unit
  - Renormalize by division, if constraint is not met
  - Prevents weights from growing too big, even if proposed update is very large
  - Allows to start with very high learning rate which decreases during training
  - Makes a more thorough search of the weight space possible
Testing

- For testing the “mean network” is used
  - Contains ALL hidden units with halved outgoing weights
  - Compensates the fact that this network has twice as many hidden units

- Why?
  - For networks with single hidden layer and softmax output, using the mean network is equivalent to taking the mean of the probability distributions over labels predicted by all possible networks

- Assumption: Not all dropout networks make the same prediction
  - Mean network assigns a higher log probability to the correct answer than the mean of the log probabilities assigned by the dropout networks
Outline

- Introduction
- Approach
- Experiments
  - MNIST
  - TIMIT
  - CIFAR-10
  - ImageNet
  - Reuters
- Conclusion
MNIST dataset

- Popular benchmark dataset for machine learning algorithms
- 28x28 images of individual handwritten digits
- 60,000 training images and 10,000 test images
- 10 classes (obviously!)
MNIST experiments

- Training with dropout on 4 different architectures:
  - Number of hidden layers (2 and 3)
  - Number of units per hidden layer (800, 1200 and 2000)

- Finetuning with dropout of a pretrained Deep Boltzmann Machine
  - 2 hidden layers (500 and 1000 units)

- Mini batches of size 100

- Maximum length of weight vector: 15
MNIST results

- Best published result for a feed-forward NN on MNIST without using enhanced training data, wiring info about spatial transformations into a CNN or using generative pre-training is 160 errors.

- This can be reduced to 130 errors by using a 50% dropout on each hidden unit and to 110 errors by also using 20% dropout on the input layer.
MNIST results

- Results for finetuning a pretrained deep Boltzmann machine five times with standard backpropagation were 103, 97, 94, 93 and 88 errors.

- For finetuning using 50% dropout results were 83, 79, 78, 78 and 77 with a mean of 79 errors which is a record for methods without prior knowledge or enhanced training sets.
TIMIT dataset

- Popular benchmark dataset for speech recognition

- Consists of recordings of 630 speakers with 8 dialects of American English each reading 10 sentences

- Includes word- and phone-level transcriptions of the speech

- Extracted inputs: 25 ms speech windows with 10 ms strides
TIMIT experiments

- Inputs: 25 ms speech windows with 10 ms strides

- Pretrained networks with different architectures:
  - Number of hidden layers (3, 4 and 5)
  - Number of units per hidden layer (2000 and 4000)
  - Number of input frames (15 and 31)

- Standard backpropagation finetuning vs. dropout finetuning
Frame classification: Dropout of 50% of the hidden units and 20% of the input units

Frame recognition error can be reduced from 22.7% without dropout to 19.7% with dropout, a record for methods without information about the speaker identity.
CIFAR-10 dataset

- Benchmark task for object recognition
- Subset of the Tiny Images dataset (50,000 training images and 10,000 test images)
- Downsampled 32x32 color images of 10 different classes
CIFAR-10 experiments

- Best previously published error rate, without transformed data, was 18.5%

- Using a CNN with 3 convolutional layers and 3 “max-pooling” layers an error rate of 16.6% could be achieved

- When using 50% dropout on the last hidden layer this could be further reduced to 15.6%
ImageNet dataset

- Very challenging object recognition dataset
- Millions of labeled high-resolution images
- Subset of 1000 classes with ca. 1000 examples each
- All images were resized to 256x256 for the experiments
ImageNet experiments

- State-of-the-art result on this dataset is an error rate of 47.7%

- CNN without dropout
  - 5 convolutional layers interleaved with "max-pooling" layers (after 1, 2 and 5)
  - "softmax output" layer
  - Achieves an error rate of 48.6%

- CNN with dropout
  - 2 additional, globally connected hidden layers before the output layer using a 50% dropout rate
  - Achieves a record error rate of 42.4%
ImageNet results

- State-of-the-art result on this dataset is an error rate of 47.7%
- CNN without dropout achieves an error rate of 48.6%
- CNN with dropout a record error rate of 42.4%
Reuter dataset

- Archive of 804,414 text documents categorized into 103 different topics
- Subset of 50 classes and 402,738 documents
- Randomly split into equal-sized training and test sets
- Documents are represented by the 2000 most frequent non-stopwords of the dataset in the experiments

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" NEWID="798">
<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>
<DATELINE>CHICAGO, March 2 - </DATELINE><BODY>The American Pork Congress kicks off tomorrow, March 3, in Indianapolis with 160 of the nation's pork producers from 44 member states determining industry positions on a number of issues, according to the National Pork Producers Council, NPPC. Delegates to the three day Congress will be considering 26 resolutions concerning various issues, including the future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national FRV (pseudorabies virus) control and eradication program, the NPPC said. A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter</BODY></TEXT></REUTERS>
Reuters experiments

- Dropout backpropagation vs. standard backpropagation

  - “softmax” output layer

- Training done for 500 epochs
The 31.05% error rate of the standard-backpropagation neural network can be reduced to 29.63% by using a 50% dropout.
Outline

- Introduction
- Approach
- Experiments
  - MNIST
  - TIMIT
  - CIFAR-10
  - ImageNet
  - Reuters
- Conclusion
Conclusion

- Random dropout allows to train many networks “at once”
- Good way to prevent overfitting
- Can be easily implemented
- Parameters are strongly regularized by being shared by all models
- “Naive Bayes” is an extreme, yet familiar case of Dropout
- Can be further improved (Maxout Networks or DropConnect)
Questions

Questions? Ask!
References

2. Wan et al., Regularization of Neural Networks using DropConnect, Proceedings of International Conference on Machine Learning (ICML), 2013