基于Improving neural networks by preventing co-adaptation of feature detectors论文，解析来自网络

## 观点

该论文从神经网络的难题出发，一步一步引出dropout为何有效的解释。大规模的神经网络有两个缺点：

* 费时
* 容易过拟合

过拟合是很多机器学习的通病，过拟合了，得到的模型基本就废了。而为了解决过拟合问题，一般会采用ensemble方法，即训练多个模型做组合，此时，费时就成为一个大问题，不仅训练起来费时，测试起来多个模型也很费时。总之，几乎形成了一个死锁。

Dropout的出现很好的可以解决这个问题，每次做完dropout，相当于从原始的网络中找到一个更瘦的网络，如下图所示：



因而，对于一个有N个节点的神经网络，有了dropout后，就可以看做是2n个模型的集合了，但此时要训练的参数数目却是不变的，这就解脱了费时的问题。（**所以李宏毅视频中讲到DropOut是ensemble的终极版本**）

## 动机论

在自然界中，在中大型动物中，一般是有性繁殖，有性繁殖是指后代的基因从父母两方各继承一半。但是从直观上看，似乎无性繁殖更加合理，因为无性繁殖可以保留大段大段的优秀基因。而有性繁殖则将基因随机拆了又拆，破坏了大段基因的联合适应性。

但是自然选择中毕竟没有选择无性繁殖，而选择了有性繁殖，须知物竞天择，适者生存。我们先做一个假设，那就是基因的力量在于混合的能力而非单个基因的能力。不管是有性繁殖还是无性繁殖都得遵循这个假设。为了证明有性繁殖的强大，我们先看一个概率学小知识。

比如要搞一次恐怖袭击，两种方式：   
- 集中50人，让这50个人密切精准分工，搞一次大爆破。   
- 将50人分成10组，每组5人，分头行事，去随便什么地方搞点动作，成功一次就算。

哪一个成功的概率比较大？ 显然是后者。因为将一个大团队作战变成了游击战。

那么，类比过来，有性繁殖的方式不仅仅可以将优秀的基因传下来，还可以降低基因之间的联合适应性，使得复杂的大段大段基因联合适应性变成比较小的一个一个小段基因的联合适应性。

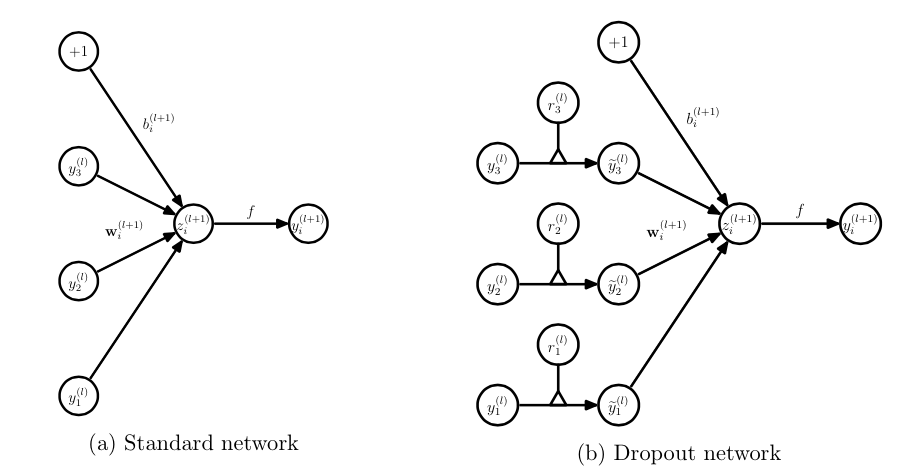
dropout也能达到同样的效果，它强迫一个神经单元，和随机挑选出来的其他神经单元共同工作，达到好的效果。消除减弱了神经元节点间的联合适应性，增强了泛化能力。

补充：植物和微生物大多采用无性繁殖，因为他们的生存环境的变化很小，因而不需要太强的适应新环境的能力，所以保留大段大段优秀的基因适应当前环境就足够了。而高等动物却不一样，要准备随时适应新的环境，因而将基因之间的联合适应性变成一个一个小的，更能提高生存的概率。

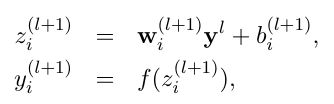
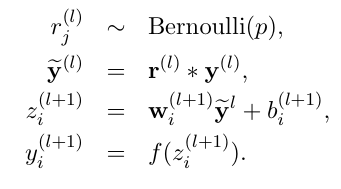
## dropout带来的模型的变化

而为了达到ensemble的特性，有了dropout后，神经网络的训练和预测就会发生一些变化。

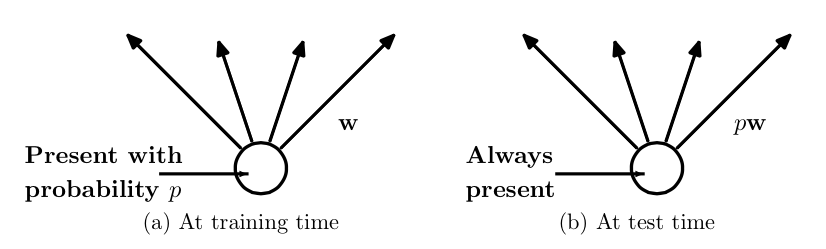
* 训练层面

无可避免的，训练网络的每个单元要添加一道概率流程。   


对应的公式变化如下如下：

* + 没有dropout的神经网络   
    
  + 有dropout的神经网络   
    
* 测试层面

预测的时候，每一个单元的参数要预乘以p，或者是输出时，乘以一个1-p。



## 论文中的其他技术点

* 防止过拟合的方法：
  + 提前终止（当验证集上的效果变差的时候）
  + L1和L2正则化加权
  + soft weight sharing
  + dropout

（李宏毅中的视频中均有提到）

* dropout率的选择
  + 经过交叉验证，隐含节点dropout率等于0.5的时候效果最好，原因是0.5的时候dropout随机生成的网络结构最多。
  + dropout也可以被用作一种添加噪声的方法，直接对input进行操作。输入层设为更接近1的数。使得输入变化不会太大（0.8）
* 训练过程
  + 对参数w的训练进行球形限制(max-normalization)，对dropout的训练非常有用。
  + 球形半径c是一个需要调整的参数。可以使用验证集进行参数调优
  + dropout自己虽然也很牛，但是dropout、max-normalization、large decaying learning rates and high momentum组合起来效果更好，比如max-norm regularization就可以防止大的learning rate导致的参数blow up。
  + 使用pretraining方法也可以帮助dropout训练参数，在使用dropout时（训练），要将所有参数都乘以1/p。

部分实验结论

该论文的实验部分很丰富，有大量的评测数据。

* maxout 神经网络中得另一种方法，Cifar-10上超越dropout
* 文本分类上，dropout效果提升有限，分析原因可能是Reuters-RCV1数据量足够大，过拟合并不是模型的主要问题
* dropout与其他standerd regularizers的对比   
  + L2 weight decay
  + lasso
  + KL-sparsity
  + max-norm regularization
  + dropout
* 特征学习   
  + 标准神经网络，节点之间的相关性使得他们可以合作去fix其他节点中得噪声，但这些合作并不能在unseen data上泛化，于是，过拟合，dropout破坏了这种相关性。在autoencoder上，有dropout的算法更能学习有意义的特征（不过只能从直观上，不能量化）。
  + 产生的向量具有稀疏性。
  + 保持隐含节点数目不变，dropout率变化；保持激活的隐节点数目不变，隐节点数目变化。
* 数据量小的时候，dropout效果不好，数据量大了，dropout效果好
* 模型均值预测
  + 使用weight-scaling来做预测的均值化
  + 使用mente-carlo方法来做预测。即对每个样本根据dropout率先sample出来k个net，然后做预测，k越大，效果越好。
* Multiplicative Gaussian Noise   
  使用高斯分布的dropout而不是伯努利模型dropout
* dropout的缺点就在于训练时间是没有dropout网络的2-3倍。

（但和ensemble的2^n倍相比减慢了很多）

进一步需要了解的知识点

* dropout RBM
* Marginalizing Dropout   
  具体来说就是将随机化的dropout变为确定性的，比如对于Logistic回归，其dropout相当于加了一个正则化项。
* Bayesian neural network对稀疏数据特别有用，比如medical diagnosis, genetics, drug discovery and other computational biology applications

[转载地址](http://blog.csdn.net/stdcoutzyx/article/details/49022443)