深度学习中的正则化

2018-10-31

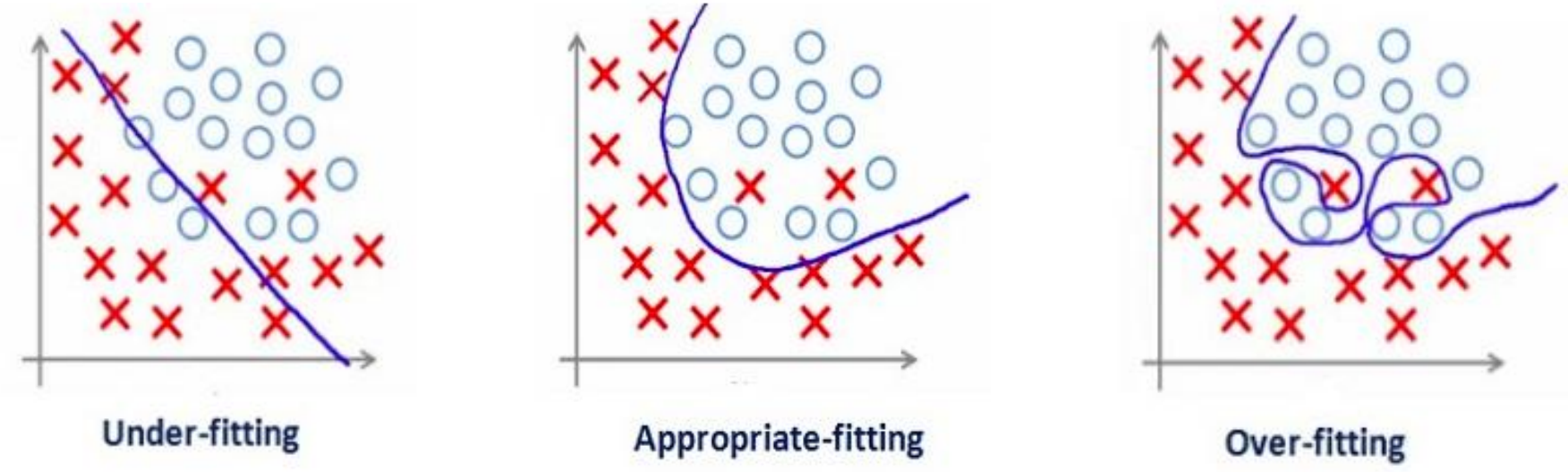
京东

过拟合

正则化

深度学习中的正则化

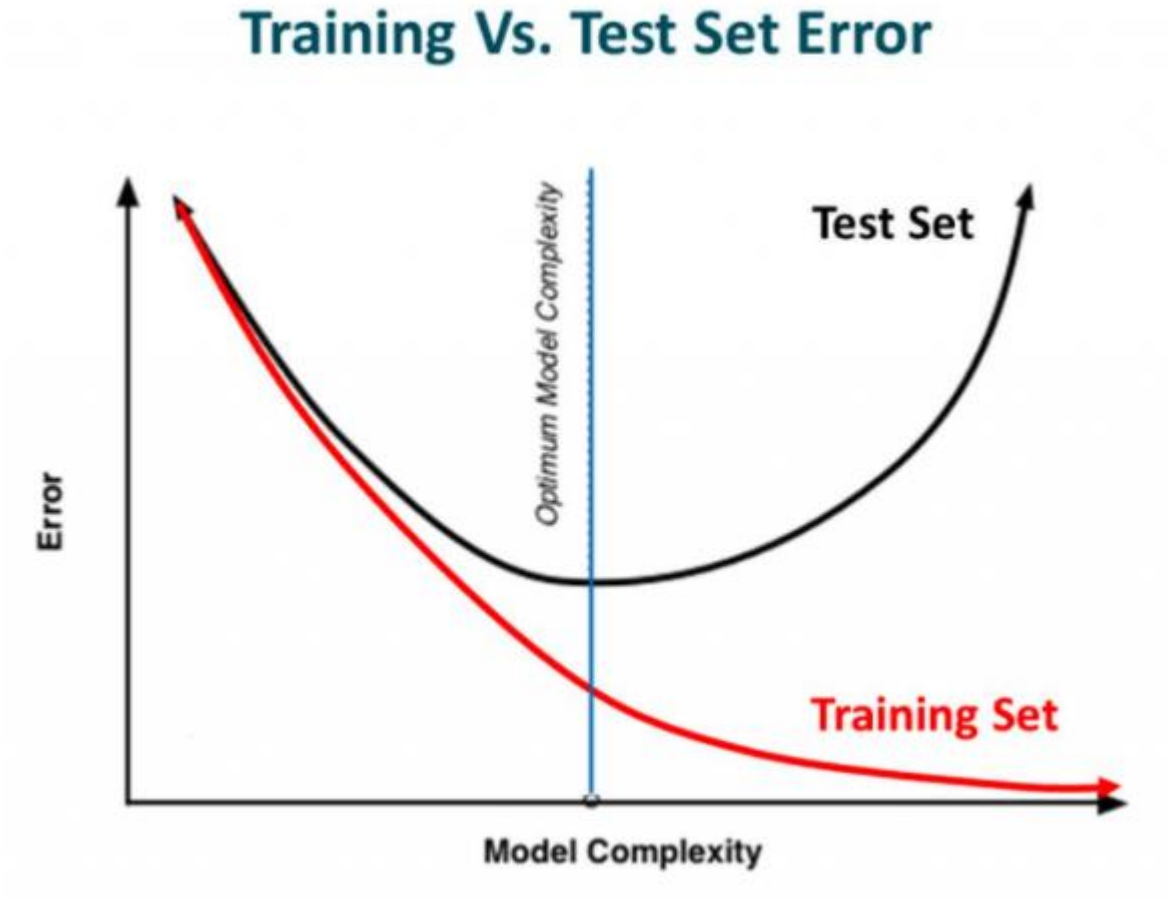
过拟合



从左往右，模型复杂度逐渐提高

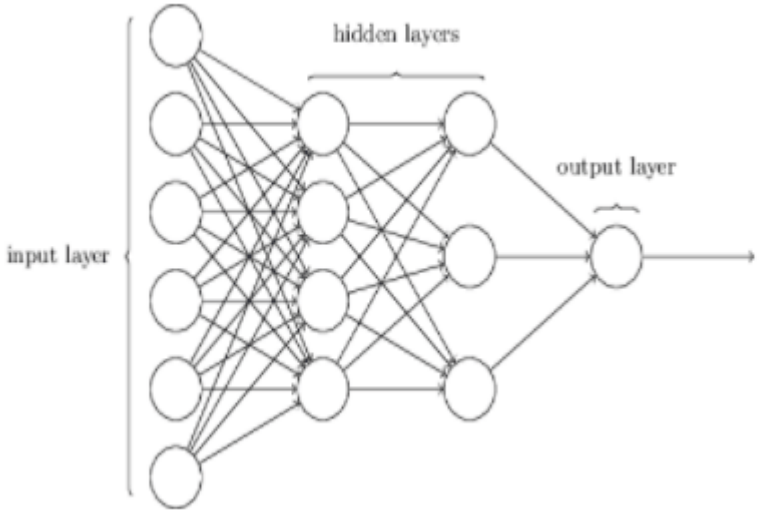
京东

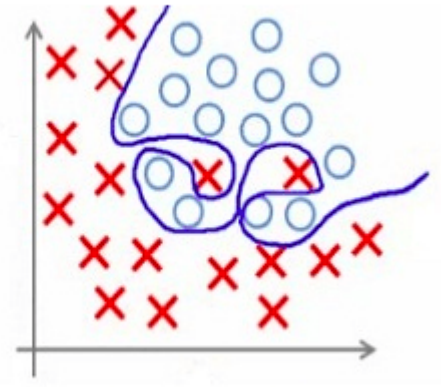
过拟合


Training Vs. Test Set Error
Test Set
Optimum Model Complexity
Error
Training Set
Model Complexity

京东

过拟合





Over-fitting

由于深度网络参数很多，深度网络非常容易出现过拟合

京东

防止过拟合

- 增加训练数据量
- 正则化

正则化

- 定义：对学习算法的修改——旨在减少泛化误差而不是训练误差

- 正则化策略：限制网络模型的神经元数量、限制模型参数（连接权重W，偏置项B等）的数目、在目标函数添加一些额外的惩罚项、集成的方法...

深度学习中的正则化

- 范数惩罚
- 数据增强（使用更多的数据进行训练）
- 多任务学习
- Early stopping（简单、有效）
- 参数绑定与参数共享（例如CNN）
- Dropout（集成大量深度神经网络的实用bagging方法）

范数惩罚

- 在原本的代价函数后面再加上一个正则化项
- L2正则化项（或者叫权重衰减）

$$C = C_0 + \frac{\lambda}{2n}\sum_w w^2$$

任意损失函数，比如交叉熵损失函数

L2正则化项如何防止过拟合呢?

京东

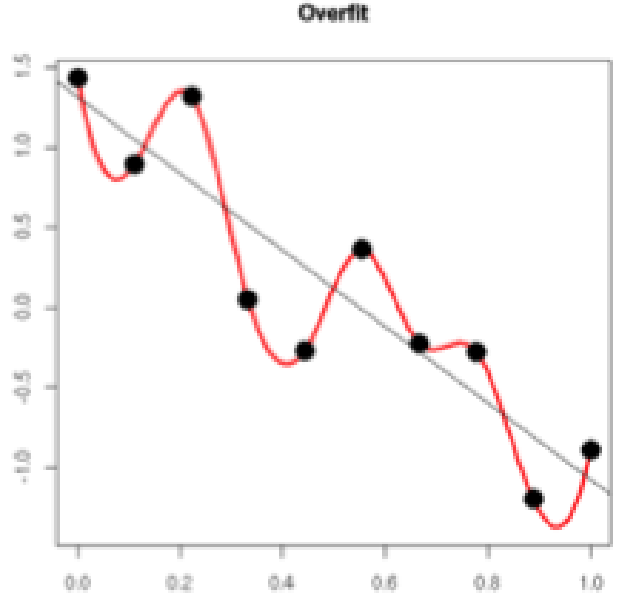
$$C = C_0 + \frac{\lambda}{2n}\sum_w w^2$$

对w,b求导

$$\frac{\partial C}{\partial w} = \frac{\partial C_0}{\partial w} + \frac{\lambda}{n} w$$
$$\frac{\partial C}{\partial b} = \frac{\partial C_0}{\partial b}.$$

- L2对b的更新没有影响，但对w的更新有影响，使得w趋于平滑

$$w \rightarrow w - \eta\frac{\partial C_0}{\partial w} - \frac{\eta\lambda}{n} w$$
$$= \left(1 - \frac{\eta\lambda}{n}\right) w - \eta\frac{\partial C_0}{\partial w}.$$



京东

范数惩罚

- L1正则化项

$$C = C_0 + \frac{\lambda}{n}\sum_w |w|. \quad \xrightarrow{\text{对w求导}} \quad \frac{\partial C}{\partial w} = \frac{\partial C_0}{\partial w} + \frac{\lambda}{n}\,\mathrm{sgn}(w),$$

$$w \to w' = w - \frac{\eta\lambda}{n}\mathrm{sgn}(w) - \eta\frac{\partial C_0}{\partial w},$$

- 当w为正时，更新后的w变小。当w为负时，更新后的w变大，使网络中的权重尽可能为0

数据增强

- 目标识别常用的方法是旋转、翻转、缩小/放大、位移等

- 语音识别中对输入数据添加随机噪声
- NLP中常用思路是进行近义词的替换
- 噪声注入，可以对输入添加噪声，也可以对隐藏层或者输出层添加噪声

多任务学习

- 多个任务通过底层的共享表示（shared representation）来互相帮助学习，提升泛化效果
- 例如：人脸面部关键点定位和属性（是否戴眼镜、是否微笑、性别等）预测

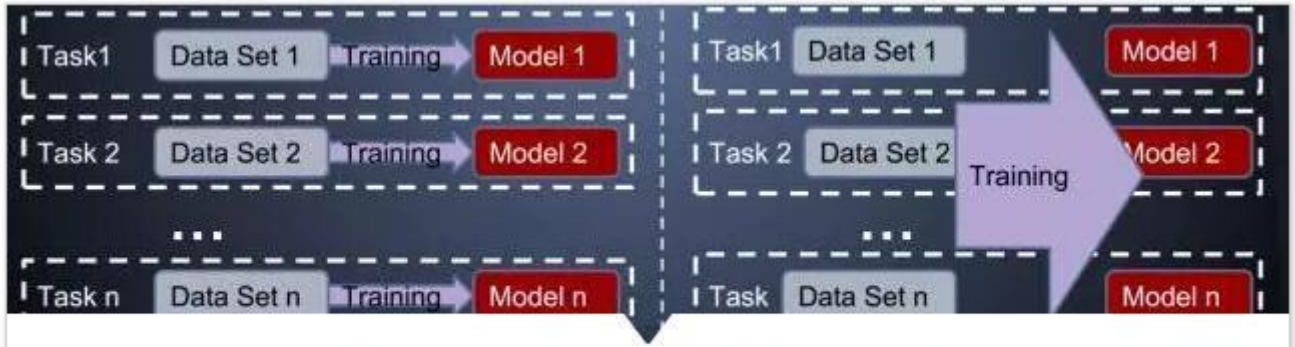


图2. 单任务学习与多任务学习的对比：1）左侧为单任务学习。2）右侧为多任务学习

（图2引自Ramtin Mehdizadeh Seraj，Multitask Learning, Jan 2014,SFU Machine Learning Reading Group）

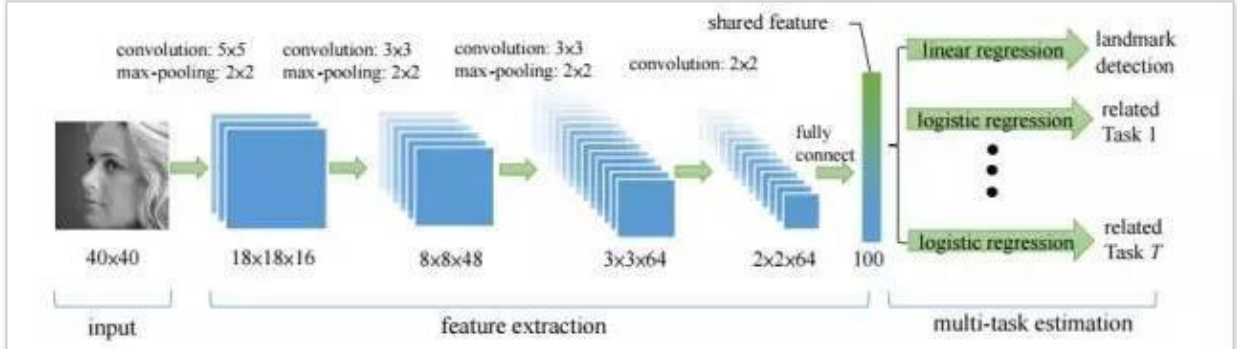
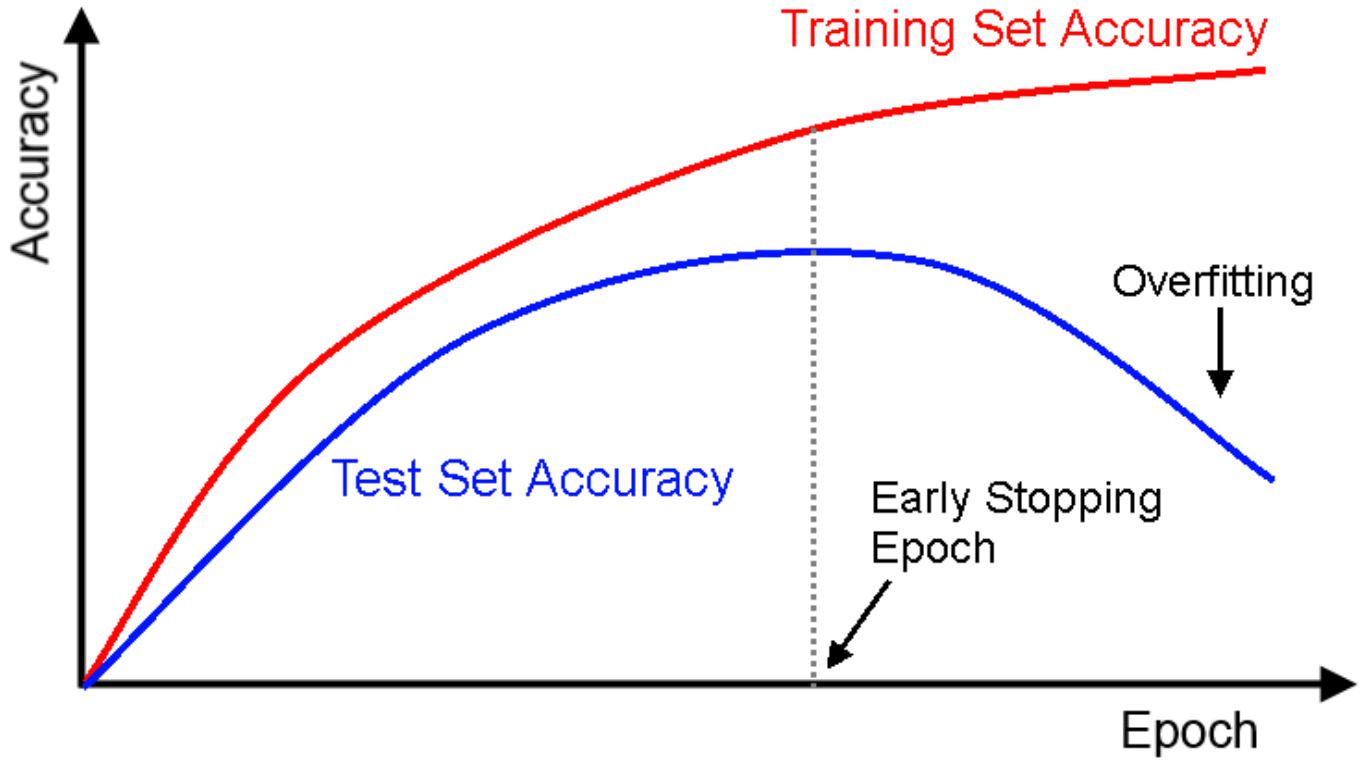


Fig. 3. Structure specification for TCDCN.

图7. TCDCN网络结构示意图

京东

Early stopping

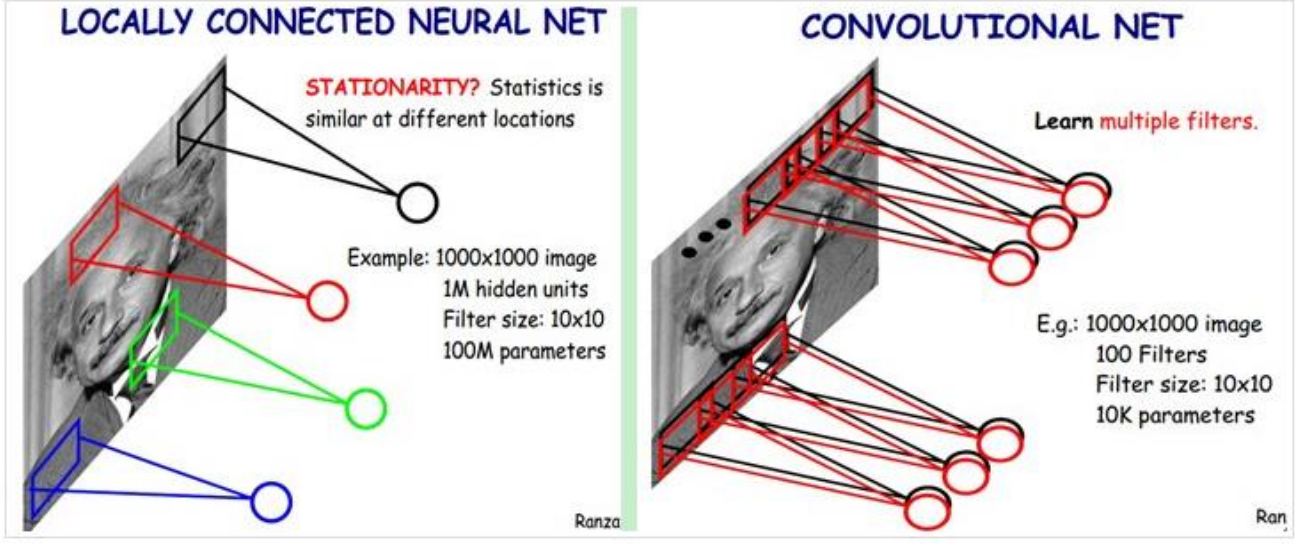


参数绑定与参数共享

- 参数绑定：任务足够相似的时候，认为模型的参数应彼此靠近

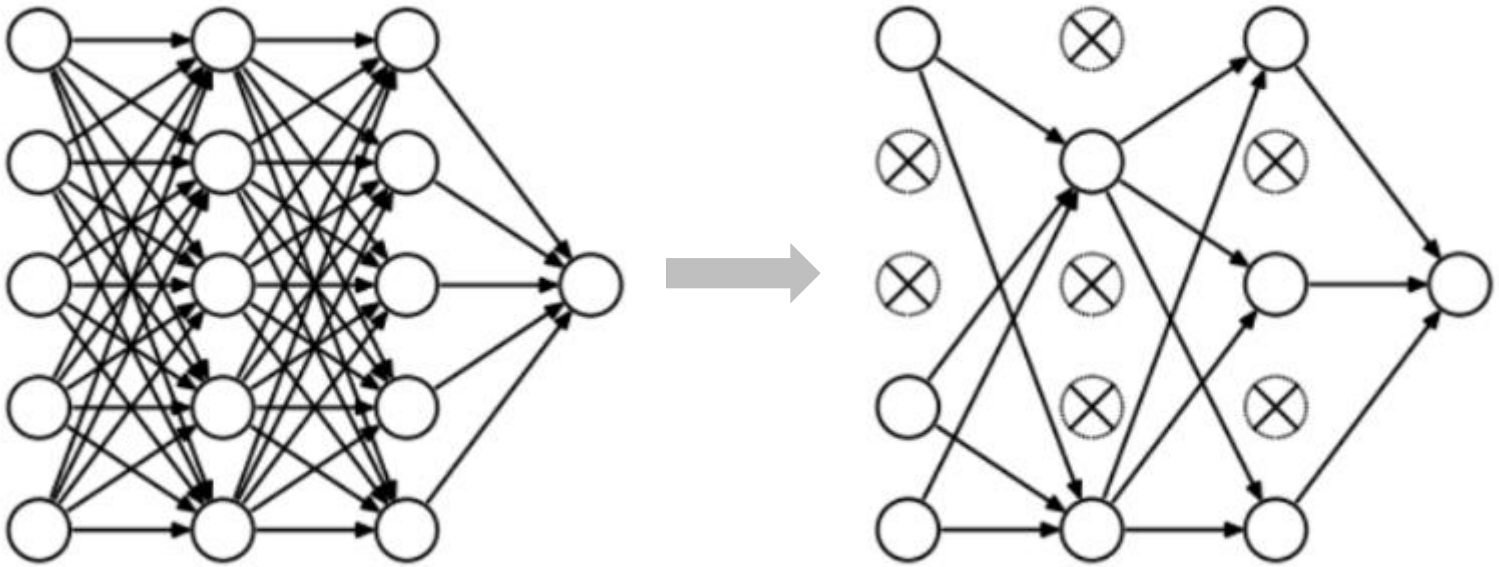
$$\Omega(w^{(A)}, w^{(B)}) = \|w^{(A)} - w^{(B)}\|_2^2$$

- 参数共享：“强迫”某些参数相等，降低参数数量，例如CNN中的权重共享



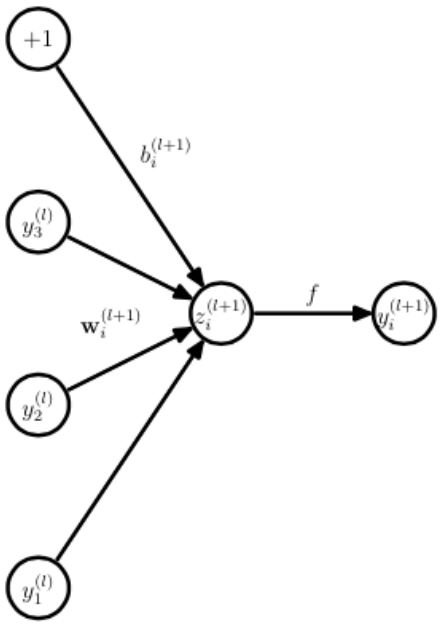
Dropout

- Dropout可以认为是集成大量深层神经网络的实用bagging方法

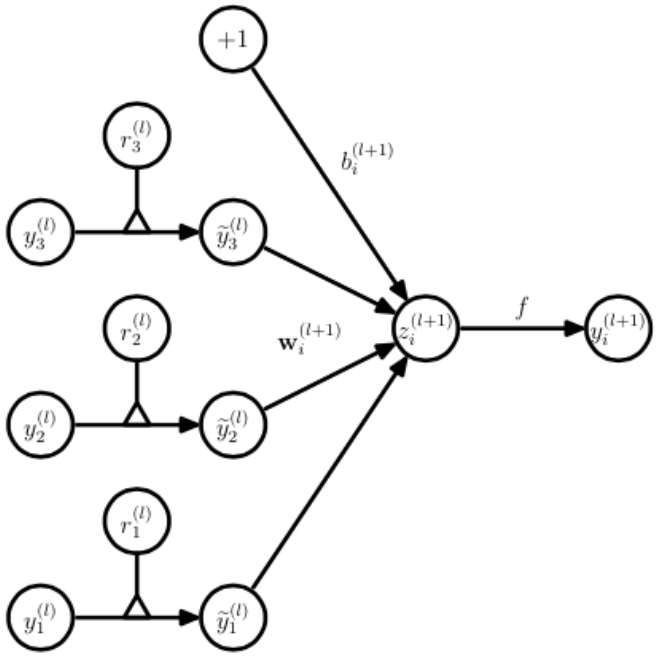


京东

➢ 在训练层面，训练网络的每个单元要添加一道概率流程



(a) Standard network



(b) Dropout network

- 没有dropout的神经网络

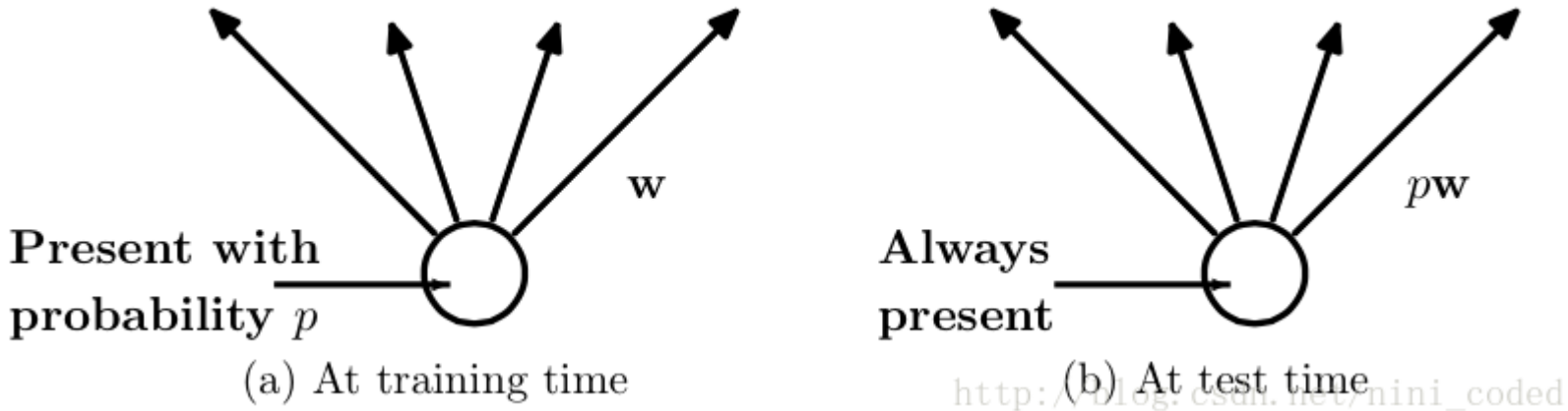
$$
\begin{aligned}
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\mathbf{y}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}),
\end{aligned}
$$

- 有dropout的神经网络

$$
\begin{aligned}
r_j^{(l)} &\sim \text{Bernoulli}(p), \\
\tilde{\mathbf{y}}^{(l)} &= \mathbf{r}^{(l)} * \mathbf{y}^{(l)}, \\
z_i^{(l+1)} &= \mathbf{w}_i^{(l+1)}\tilde{\mathbf{y}}^l + b_i^{(l+1)}, \\
y_i^{(l+1)} &= f(z_i^{(l+1)}).
\end{aligned}
$$

京东

- 在测试层面，预测的时候，每一个单元的参数要乘以p



(a) At training time (b) At test time

京东