Методы “регуляризации” нейронных сетей

Яворская Мария

МГУ имени М. В. Ломоносова, факультет ВМК, кафедра ММП

22 марта 2018 г.

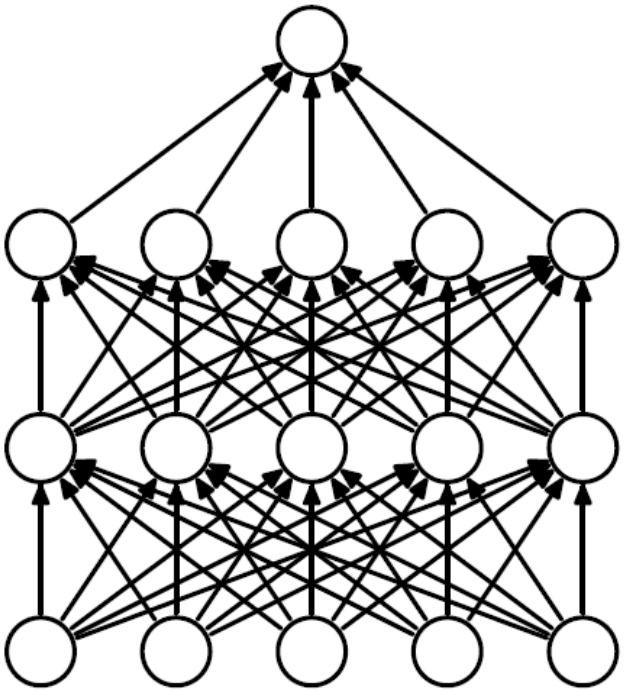
Рассматриваемые методы

- Dropout, inverted dropout;
- Батч-нормализация;
- Методы инициализации.

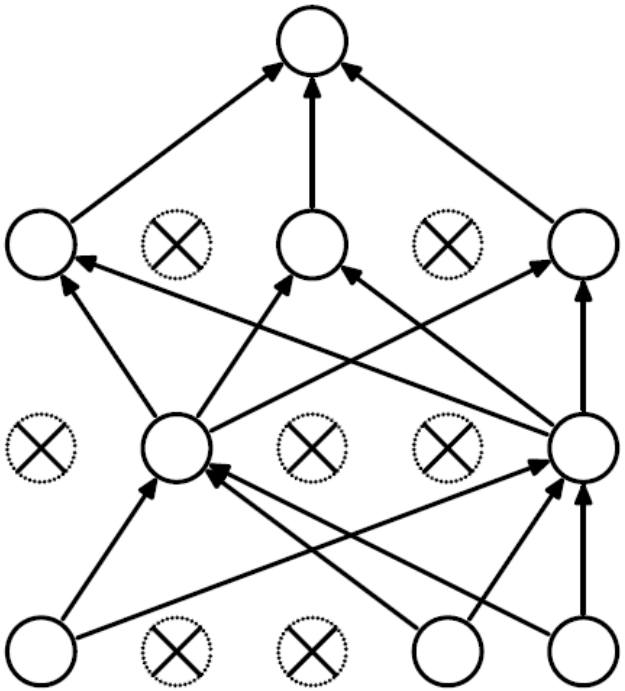
Dropout

- Проблемы:
  - переобучение — проблема, когда нейронная сеть теряет способность к обобщению;
  - сложность комбинирования ответов больших нейронных сетей во время тестирования;
- Решение: dropout;
- Идея: сети для обучения получаются с помощью исключения из сети (dropping out) нейронов. “Исключение” нейрона означает, что при любых входных данных или параметрах он возвращает 0.

Dropout



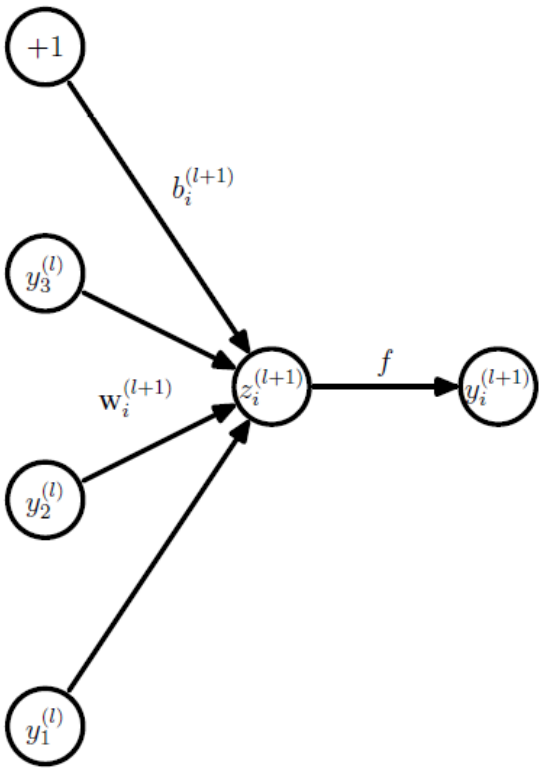
(a) Standard Neural Net



(b) After applying dropout.

Dropout

- $l \in \{1, \ldots, L\}$ — индексы скрытых слоев сети;
- $z^{(l)}$ — входной вектор слоя l;
- $y^{(l)}$ — выходной вектор слоя l ($y^{(0)} = x$ — исходные данные);
- $w^{(l)}$ и $b^{(l)}$ — вектор весов и сдвигов, соответствующих слою l;
- $f()$ — функция активации.



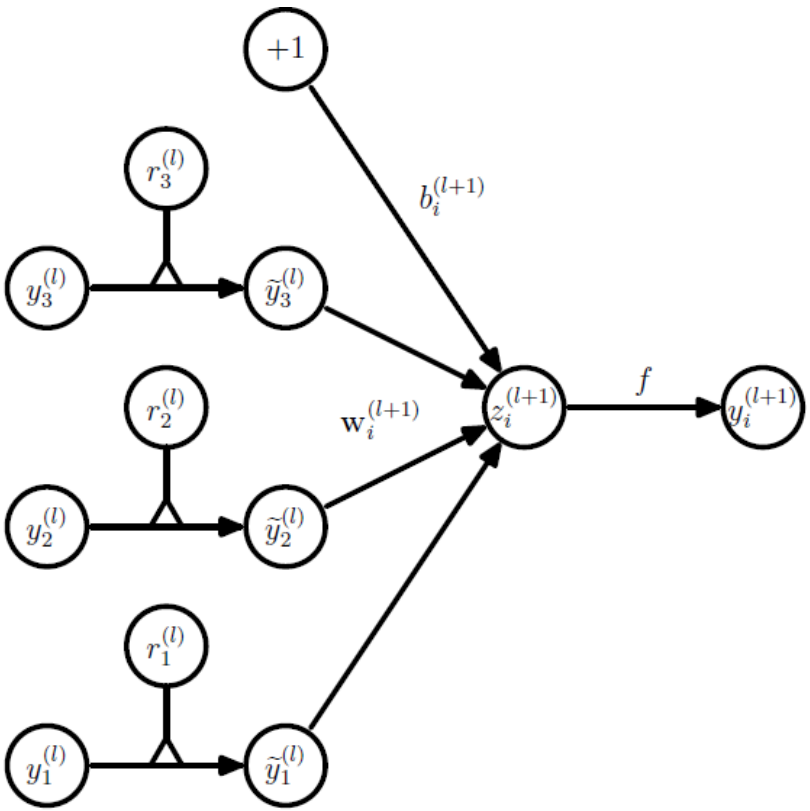
Прямой проход для стандартной нейронной сети:

$$z_i^{(l+1)} = w_i^{(l+1)} y^{(l)} + b_i^{(l+1)}$$

$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right)$$

Dropout

- $l \in \{1, \ldots, L\}$ — индексы скрытых слоев сети;
- $z^{(l)}$ — входной вектор слоя l;
- $y^{(l)}$ — выходной вектор слоя l ($y^{(0)} = x$ — исходные данные);
- $w^{(l)}$ и $b^{(l)}$ — вектор весов и сдвигов, соответствующих слою l;
- $f()$ — функция активации.



Прямой проход для нейронной сети с dropout:

$$r_j^{(l)} \sim \text{Bernoulli}\,(p)$$

$$\tilde{y}^{(l)} = r^{(l)} * y^{(l)}$$

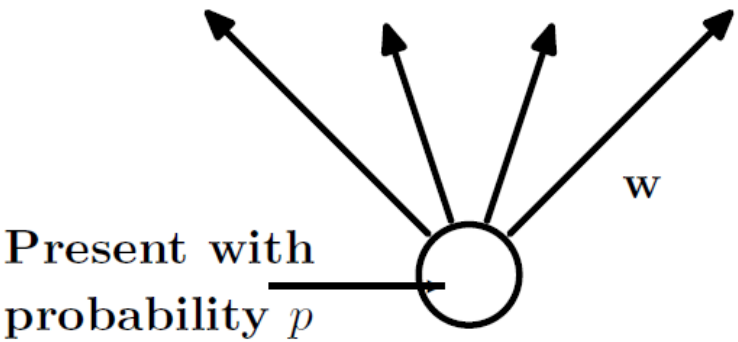
$$z_i^{(l+1)} = w_i^{(l+1)} \tilde{y}^{(l)} + b_i^{(l+1)}$$

$$y_i^{(l+1)} = f\left(z_i^{(l+1)}\right)$$

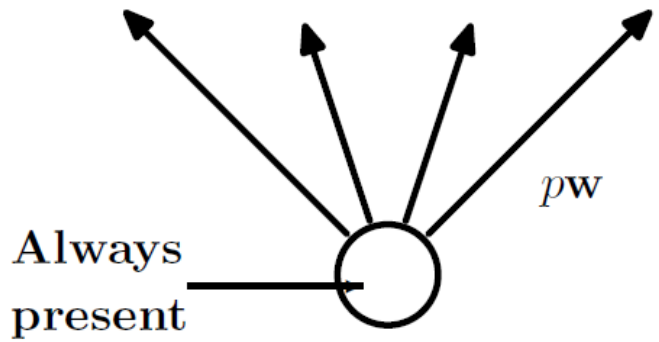
Dropout

Этап тестирование:

$$y_i^{(l+1)} = f\left(pw_i^{(l+1)} y^{(l)} + b_i^{(l+1)}\right)$$

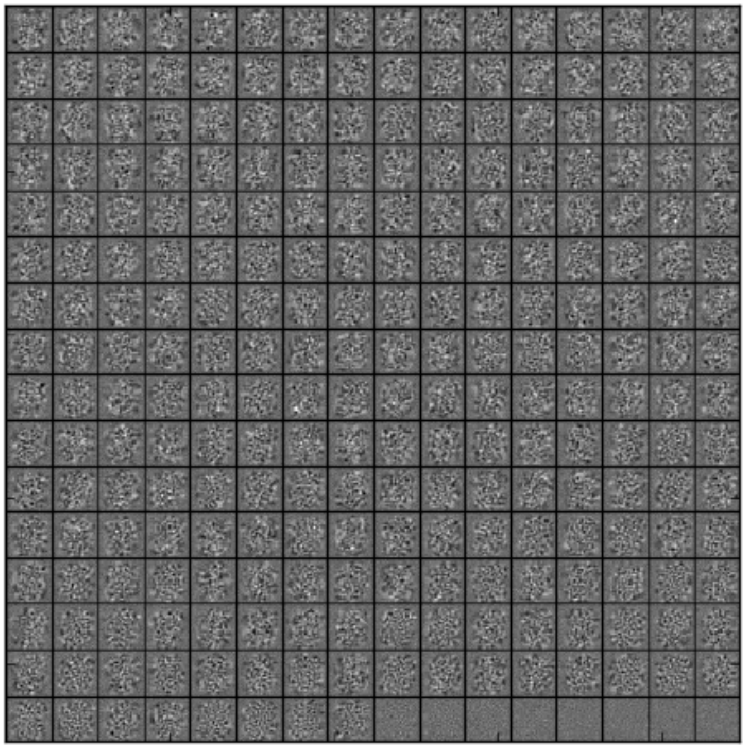


(a) At training time

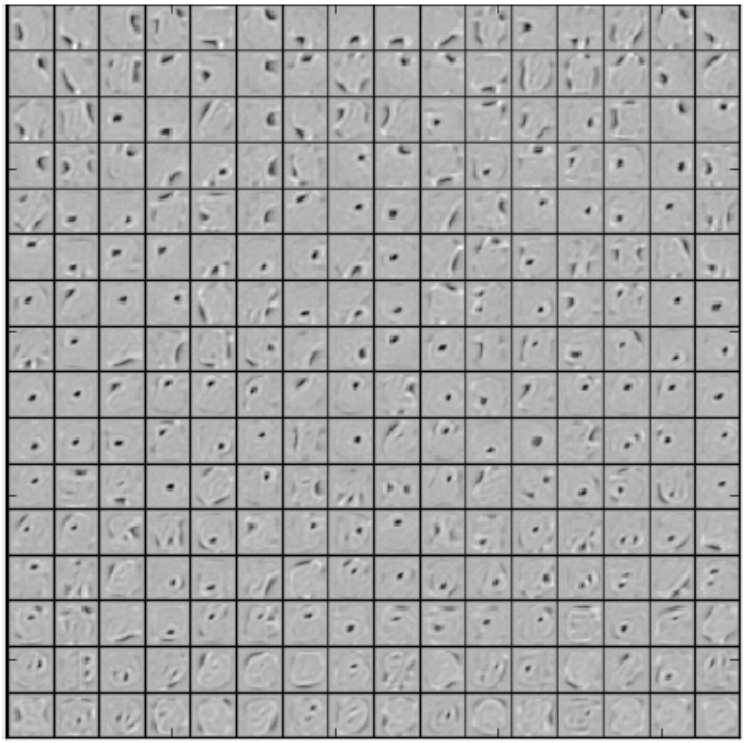


(b) At test time

Dropout



(a) Without dropout



(b) Dropout with $p = 0.5$.

Признаки, выделенные на датасете MNIST при помощи автоэнкодера

Dropout

Преимущества:

- Предотвращает переобучение;
- Возможность комбинирования ответов больших нейронных сетей во время тестирования.

Недостатки:

- Необходимость изменять нейронную сеть для проведения тестирования.

Inverted Dropout

Идея: умножать функцию активации на коэффициент не во время тестового этапа, а во время обучения.

Этап обучение:

$$y_i^{(l+1)} = f\left(\frac{1}{p} w_i^{(l+1)} (r^{(l)} * y^{(l)}) + b_i^{(l+1)}\right)$$

Этап тестирования:

$$y_i^{(l+1)} = f\left(w_i^{(l+1)} y^{(l)} + b_i^{(l+1)}\right)$$

Батч-нормализация

- Проблема: долгое время обучения нейронных сетей;
- Стандартные решения:
  - Увеличение learning rate;
  - Уменьшение число параметров сети;
  - Изменение learning rate в процессе обучения особым способом;
- Предлагаемое решение: батч-нормализация;
- Идея: добавить в архитектуру сети нормализацию по каждому из обучающих минибатчей.

Батч-нормализация

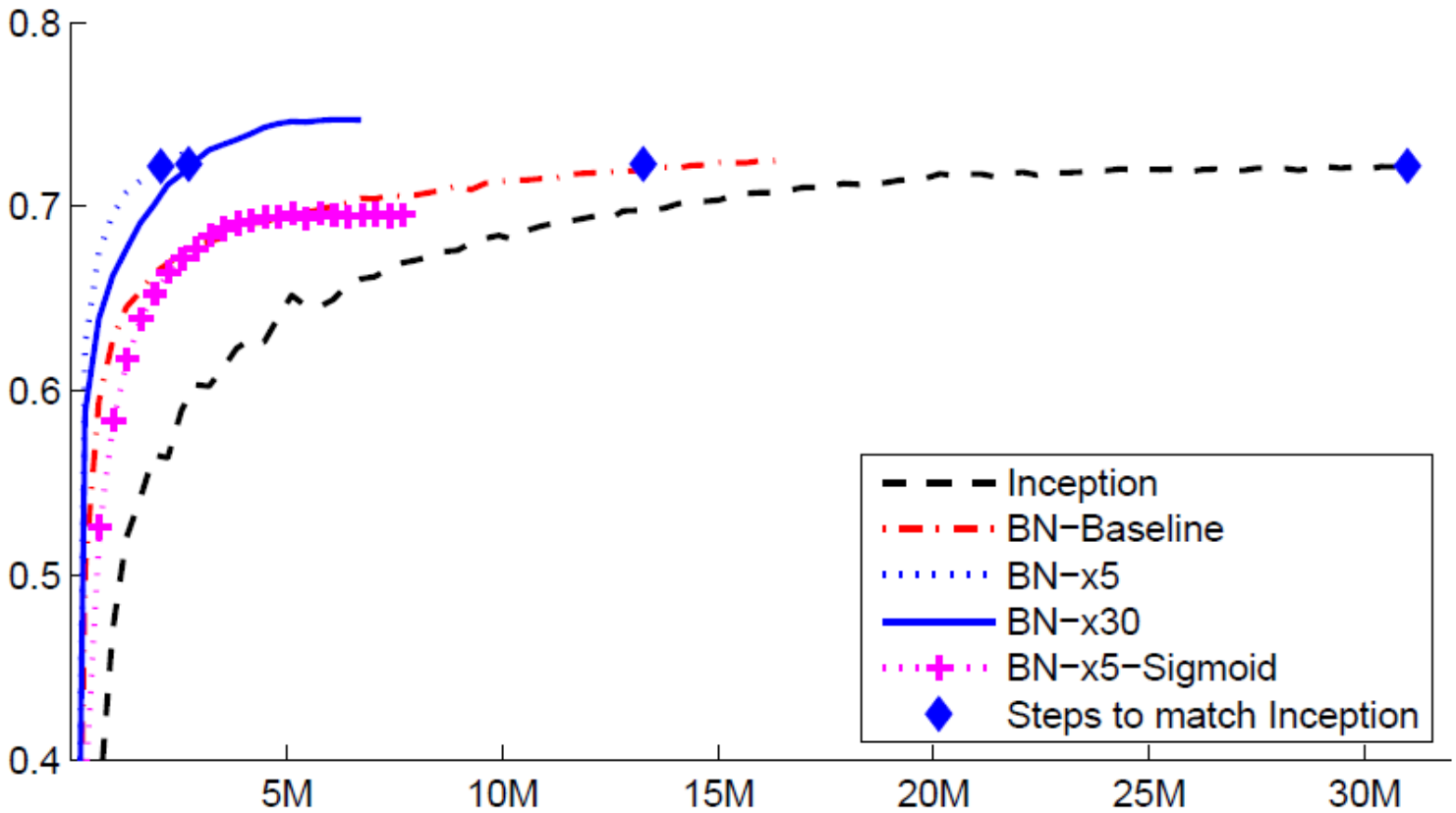
Дано: значения $x \in B$, где $B = \{x_{1...m}\}$

Требуется: $\{y_i = BN_{\gamma,\beta}(x_i)\}$

Алгоритм:

1. $\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$
2. $\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B)^2$
3. $\hat{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$
4. $y_i = \gamma\hat{x}_i + \beta \equiv BN_{\gamma,\beta}(x_i)$

Батч-нормализация



Точность классификации «начального» алгоритма и его вариаций с батч-нормализацией vs количество итераций на этапе обучения.

Батч-нормализация

Преимущества:

- Уменьшение времени обучения нейронной сети;
- Возможность строить сети без применения dropout и с малым коэффициентом L_2 регуляризации.

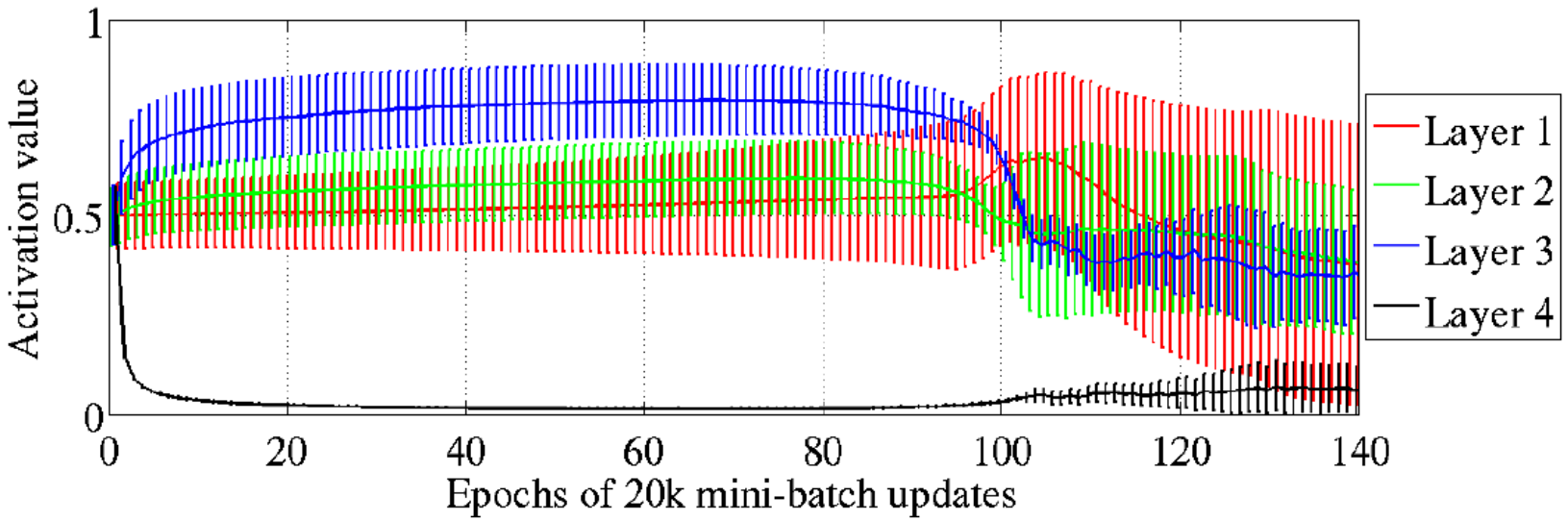
Недостатки:

- Появляется еще $2L$ параметров, которые необходимо обучить.

Методы инициализации

- Проблема:
  - Перенасыщение нейронов;
  - Скорость сходимости.
- Решение: методы инициализации;
- Идея: осознанный выбор начальных значений весов для слоев, составляющих модель.

Методы инициализации



Средние значения и дисперсии выходных значений функций активаций различных слоев

Методы инициализации

Предположим, что веса и входные значения не коррелируют и имеют нулевое матожидание.

$$Var(\sum_{i=1}^{n_{in}} w_i x_i) = \sum_{i=1}^{n_{in}} Var(w_i x_i) =$$

$$= \sum_{i=1}^{n_{in}} Var(W)Var(X) = n_{in} Var(W)Var(X)$$

Прямой проход:

$$Var(W) = \frac{1}{n_{in}}$$

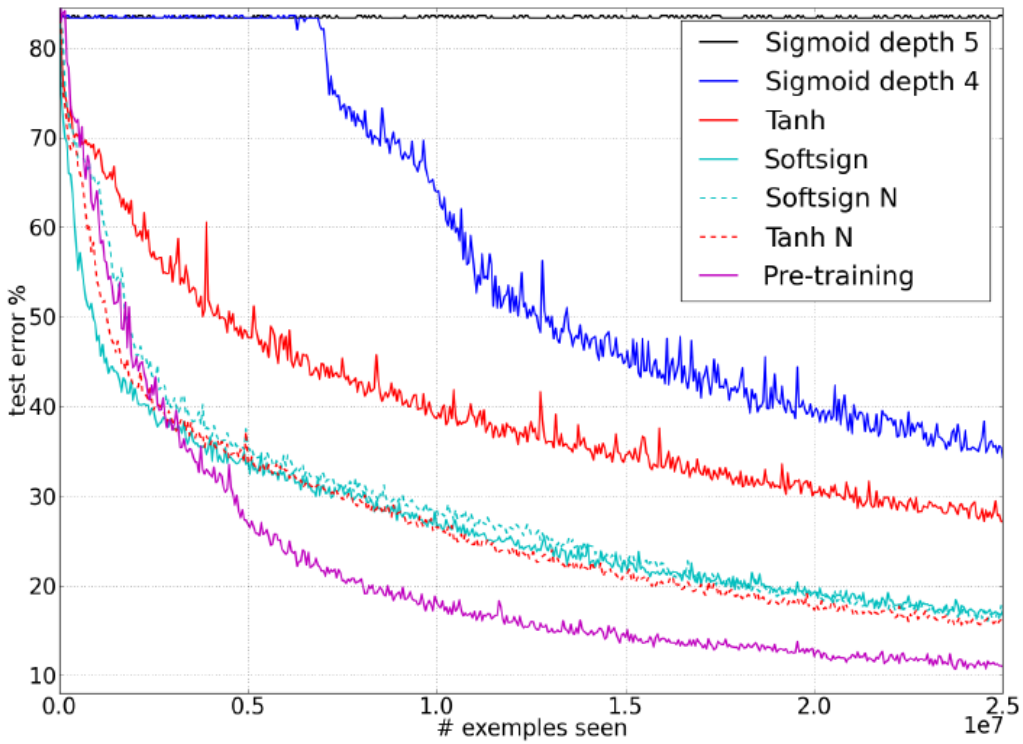
Обратный проход:

$$Var(W) = \frac{1}{n_{out}}$$

Методы инициализации

Метод инициализации Завьера (Xavier):

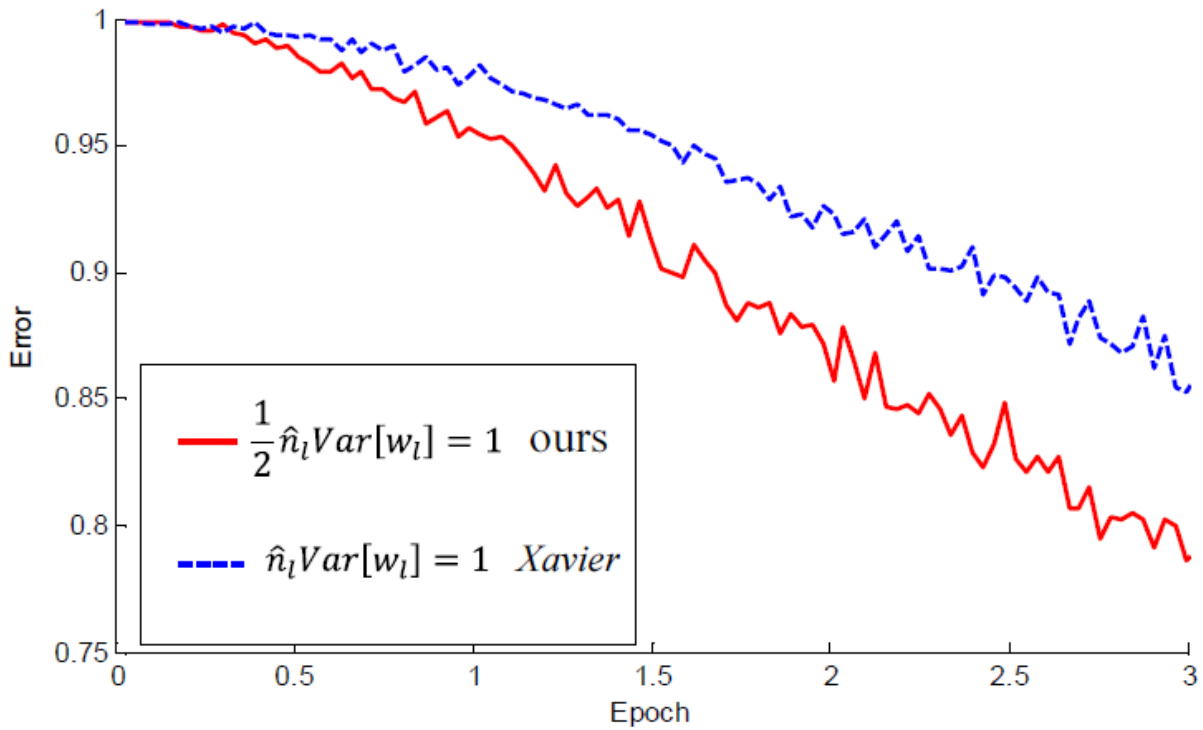
$$Var(W) = \frac{2}{n_{in}+n_{out}}$$



Методы инициализации

Метод инициализации Ге (He):

$$Var(W) = \frac{2}{n_{in}}$$



Список литературы

- Dropout:
  - Dropout: A Simple Way to Prevent Neural Networks from Overfitting — N.Srivastava, G.Hinton (2014)
  - Dropout — метод решения проблемы переобучения в нейронных сетях
- Батч-нормализация:
  - Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift — S.Ioffe, C.Szegedy (2015)
  - Batch Normalization для ускорения обучения нейронных сетей
- Методы инициализации:
  - Understanding the difficulty of training deep feedforward neural networks — X.Glorot, Y.Bengio (2010)
  - Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification — K.He, X.Zhang (2015)
  - An Explanation of Xavier Initialization