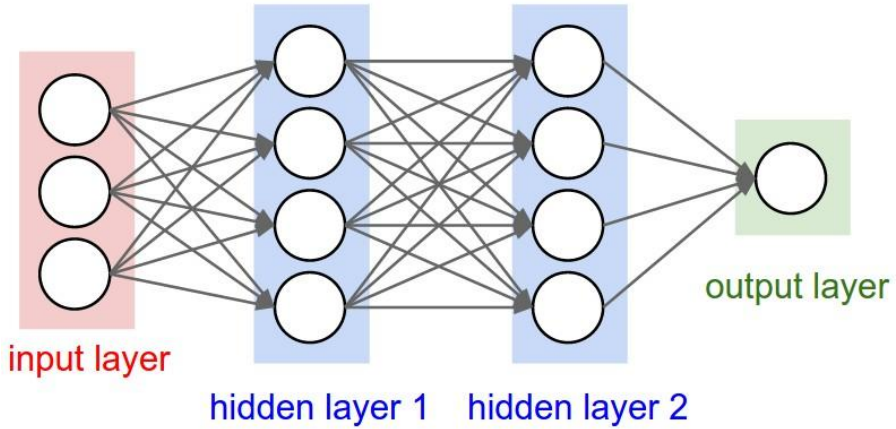
Основы практического использования нейронных сетей.

Лекция 2. Выбор архитектуры. Типы функции активации.

Дмитрий Буряк.
к.ф.-м.н.
dyb04@yandex.ru

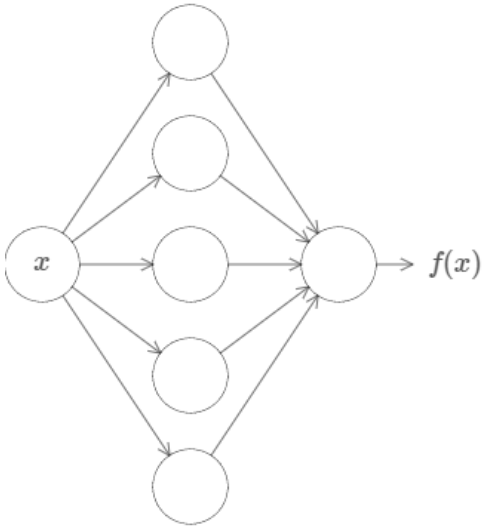
Типы слоев НС

- Входной слой.
  - Размер определяется исходными данными и методами предобработки.
- Выходной слой.
  - Размер определяется постановкой задачи.
- Внутренние слои.



Количество внутренних слоев

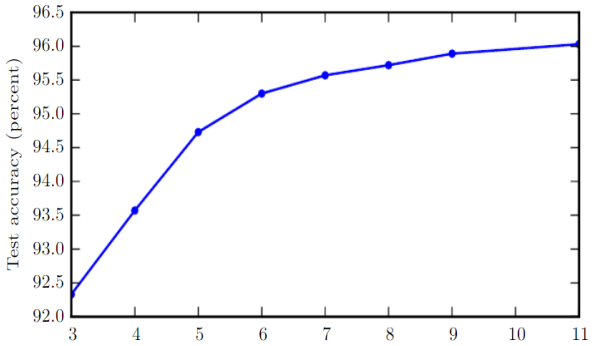
- Теорема об универсальном аппроксиматоре → 1 внутренний слой
- Не гарантирует возможность построения НС с 1 внутренним слоем
- Проблемы при построении такой НС:
  - алгоритм обучения не сходится к оптимальному набору параметров;
  - переобучение.



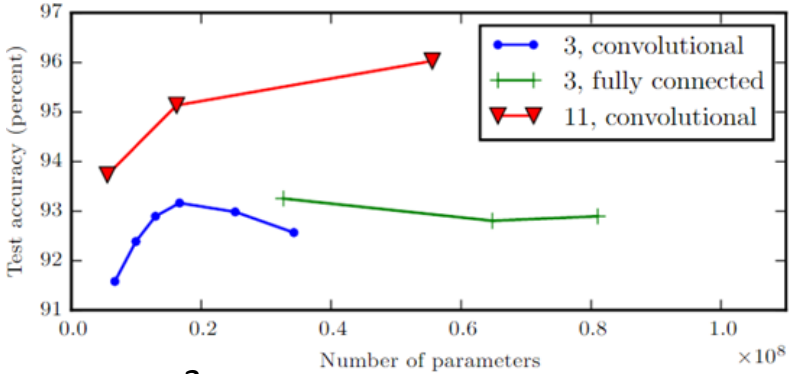
Увеличение числа внутренних слоев

- На практике 2 внутренних слоя эффективнее.
- При > 3 внутренних слоев → эффективность меняется слабо.
- Глубокие НС → ~10 – 100 внутренних слоев
  - особенности исходных данных и организации их обработки.

Задача распознавания многоразрядных чисел



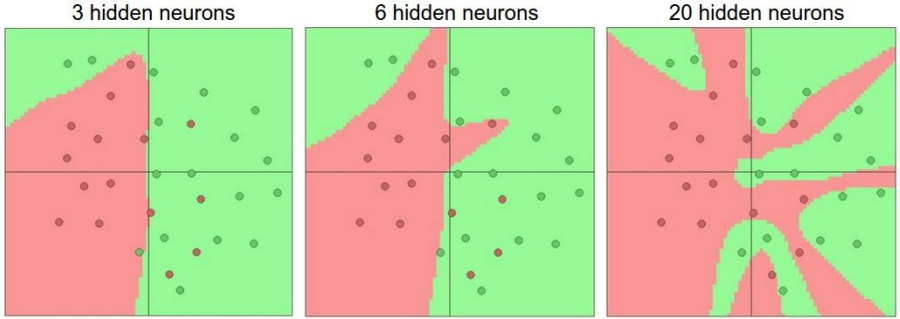
Зависимость точности от количества слоев



Зависимость точности от числа параметров

Увеличение размеров слоя

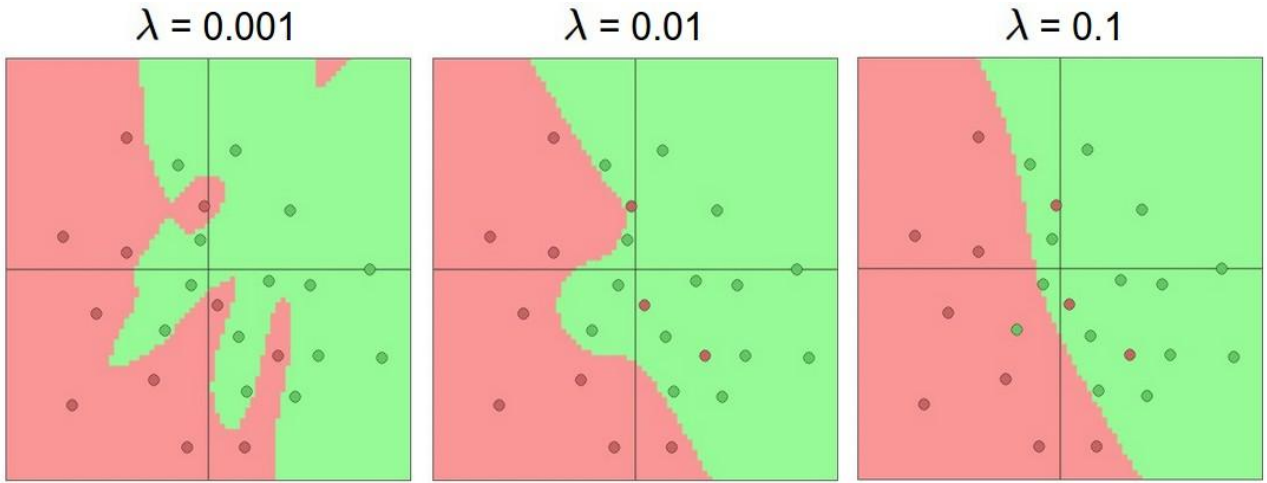
- Увеличение размеров и числа слоев → рост емкости сети.



- Емкость VS переобучение
- Решение проблемы переобучения
  - оптимизация размеров сети;
  - регуляризация;
  - drop-out;
  - генерация дополнительных данных для обучения (augmentation).

Регуляризация.

- Влияние коэффициента регуляризации (L2) на переобучение.
- Сеть с 20 нейронами во внутреннем слое.



Влияние на вид функции ошибки.

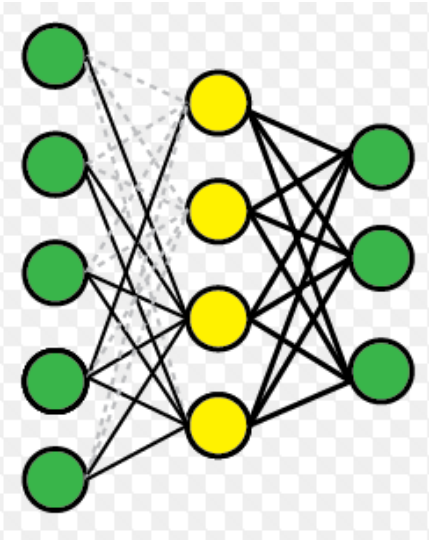
- Рост размера сети

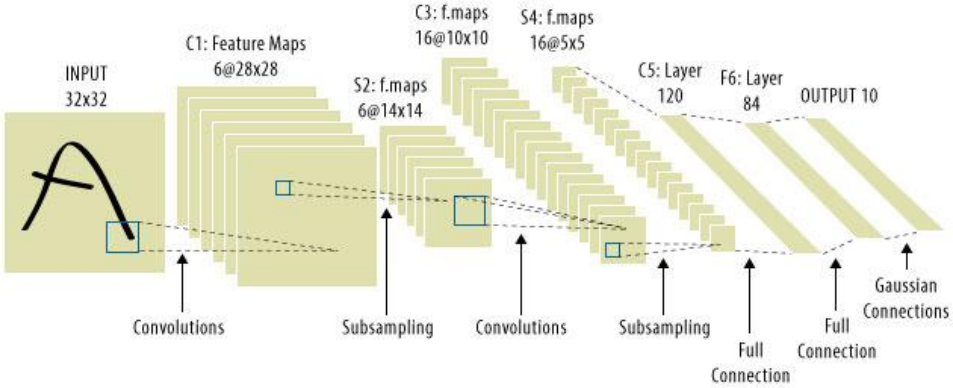
  → увеличение числа локальных экстремумов функции ошибки

  → уменьшения разброса значений экстремумов (эффективности получаемых решений).

Частичная связанность.

- Прореживание связей → уменьшение числа настраиваемых параметров
- Сети свертки → совместное использование весов.





Функция активации. Выходной слой

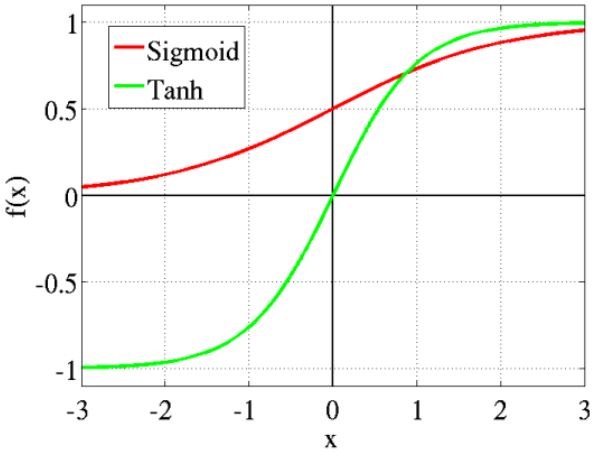
- Вид функции активации влияет на выбор функции ошибки.
- Линейная функция активации: аффинное преобразование:
  - $\hat{y} = W^\top h + b.$
  - эффективны алгоритмы обучения, основанные на оценке градиента

Функция активации. Выходной слой (2)

- Сигмоидальная функция активации:
  - $\sigma(x) = \frac{1}{1 + \exp(-x)}$ $\hat{y} = \sigma\left(\boldsymbol{w}^\top \boldsymbol{h} + b\right)$
  - Задачи классификации с двумя классами
- Softmax
  - $\text{softmax}(\boldsymbol{z})_i = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$ $\boldsymbol{z} = \boldsymbol{W}^\top \boldsymbol{h} + \boldsymbol{b}$
  - Задачи классификации с N классами;
  - Моделирование значения вероятности принадлежности к классу.

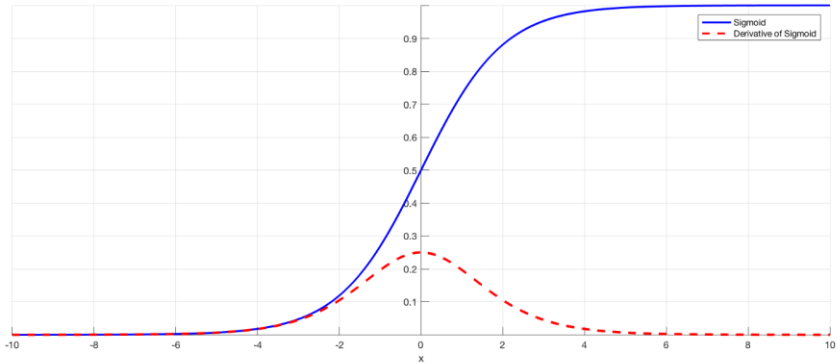
Функция активации. Сигмоиды

- Логистическая сигмоида, гиперболический тангенс
  - большая область насыщения → падение эффективности градиентных методов обучения;
  - проблема «исчезающего градиента» (vanishing gradient)
  - предпочтение гиперболическому тангенсу
  - редкое применение на фоне ReLU для сетей прямого распространения
  - активное использование в рекуррентных сетях, вероятностных моделях и др.



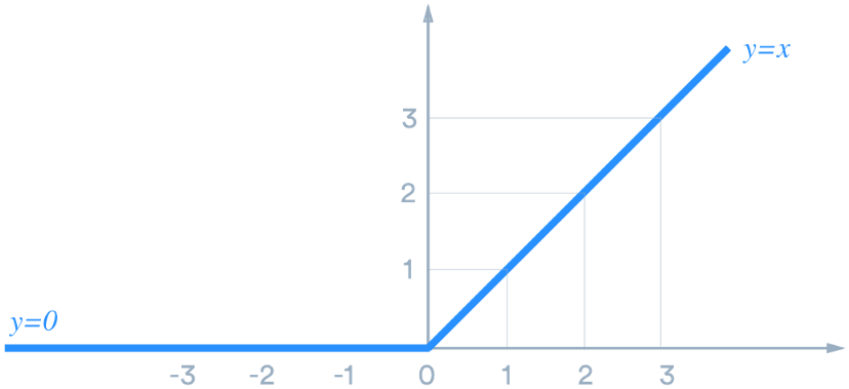
Vanishing Gradient Problem

- ❑ Уменьшение величины градиента с ростом числа слоев, пройденных в ходе обратного распространения ошибки.
- ❑ Сигмоиды имеют большие области насыщения с малыми значениями производных .
- ❑ Градиент функции ошибки в промежуточном слое зависит от произведения всех производных функций активации в прошедших слоях.
- ❑ Коррекция весов в первых слоях сети существенно меньше чем в последних.



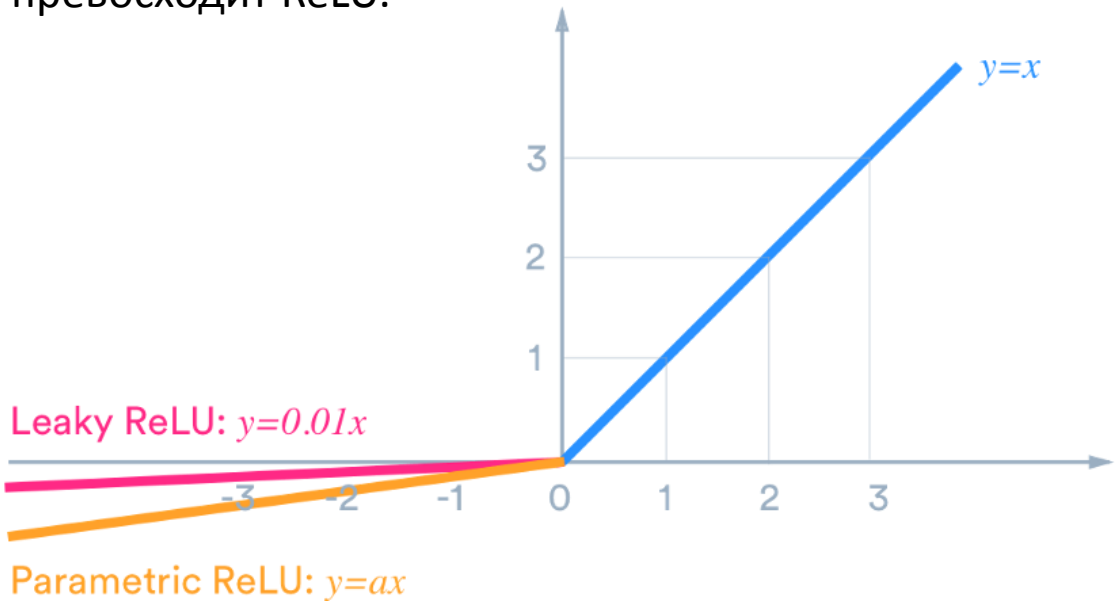
Функция активации. ReLU

- ReLU:
  - $g(z) = \max\{0, z\} \quad \boldsymbol{z} = \boldsymbol{W}^\top \boldsymbol{h} + \boldsymbol{b}$
  - Низкая вычислительная сложность;
  - Хорошая сходимость градиентных методов оптимизации;
  - Постоянная производная при x>0;
  - Дополнительная устойчивость к переобучению (производная равна 0 при x<0);
  - Проблема «умирающий ReLU» (решение – невысокая скорость обучения)



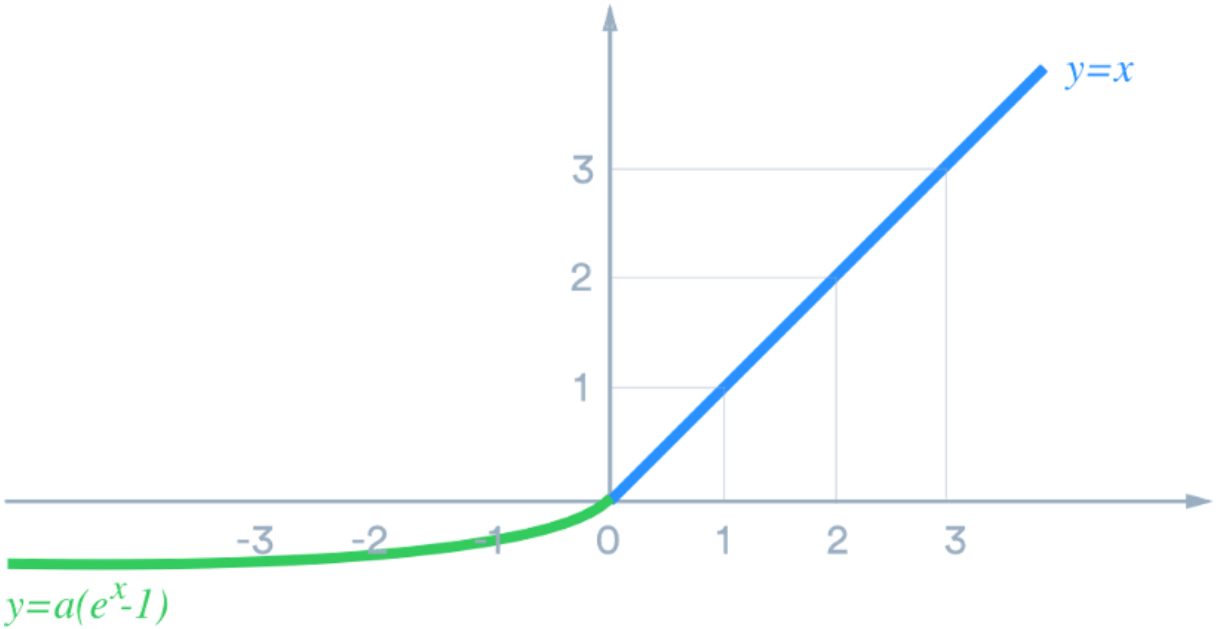
Функция активации. Варианты ReLU

- Варианты ReLU: leaky ReLU, parameteric ReLU
  - $h_i = g(\boldsymbol{z}, \boldsymbol{\alpha})_i = \max(0, z_i) + \alpha_i \min(0, z_i)$
  - Решение проблемы «умирающий ReLU»;
  - Не всегда превосходит ReLU.



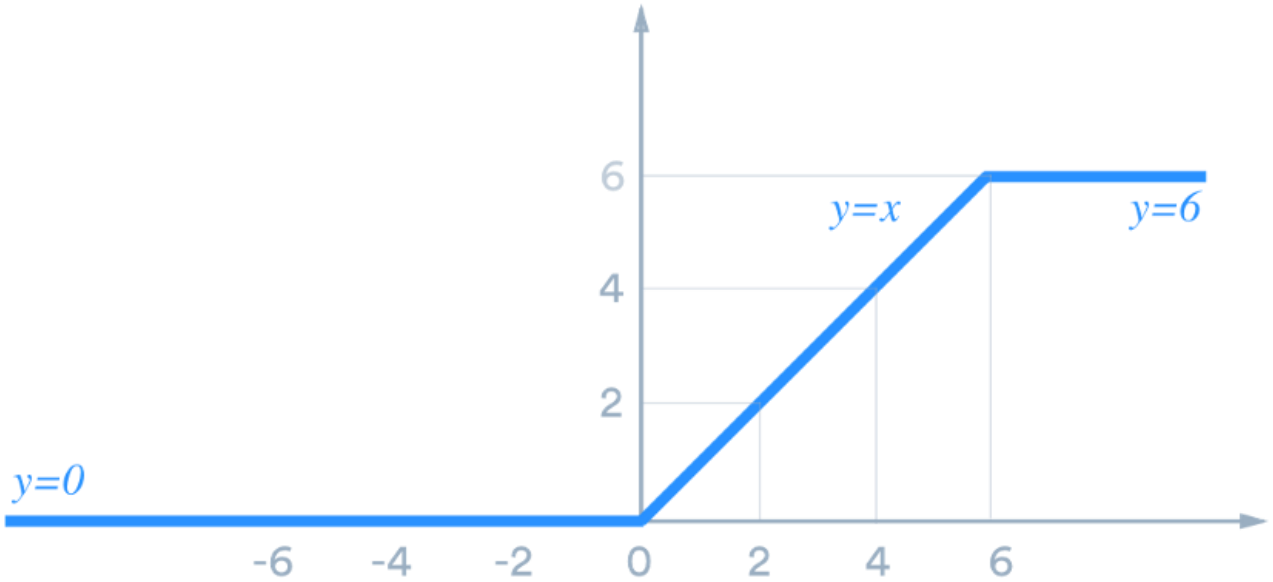
Функция активации. ELU

- Exponential Linear (ELU, SELU)
  - Решение проблемы «умирающий ReLU»;
  - Ограниченная снизу.



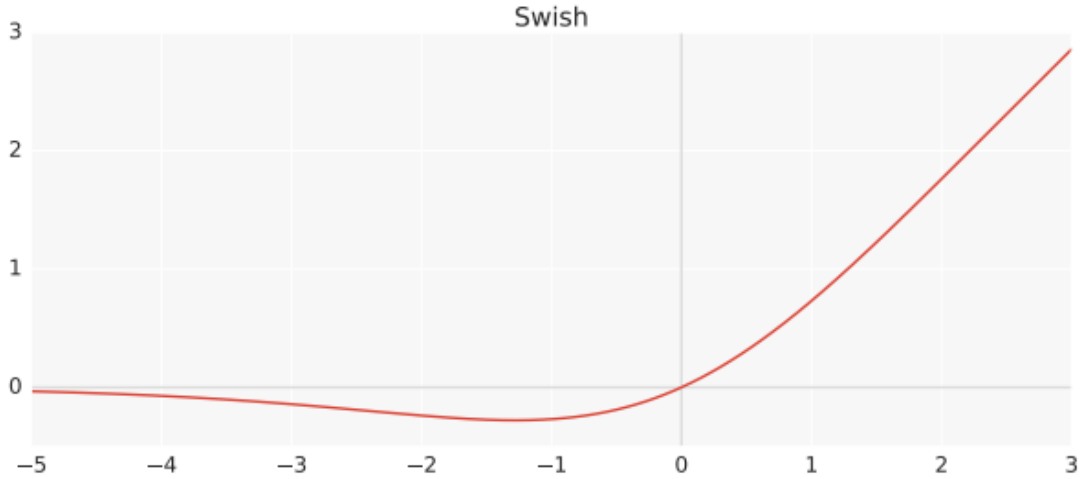
Функция активации. ReLU6

- ReLU6
  - Ограниченная сверху;
  - Применяется в ряде известных архитектур (MobileNet)



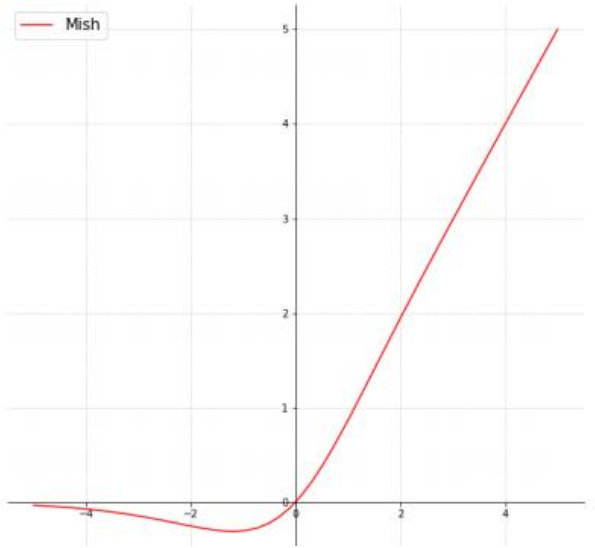
Функция активации. Swish

- Swish
  - $f(x) = x * \sigma(\beta x)$ β – фиксирован или обучаем
  - Аналогична ReLU при x>0;
  - Решение проблемы «умирающей ReLU»;
  - Немонотонная функция при x<0;
  - Во ряде случаев эффективнее ReLU.



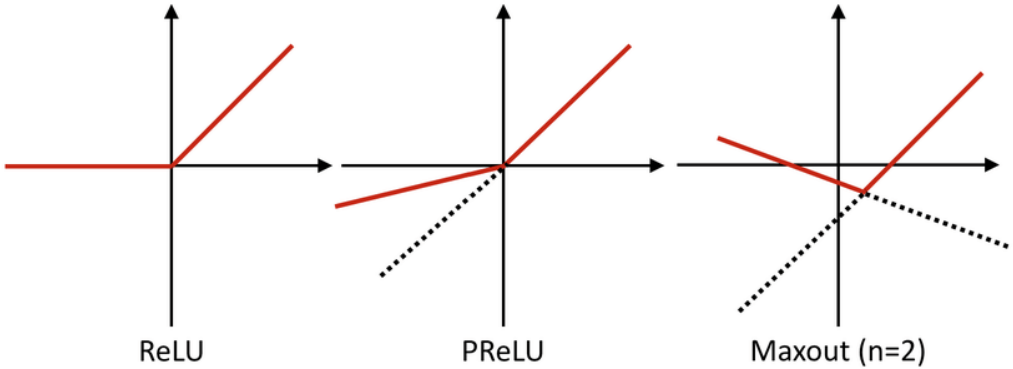
Функция активации. Mish

- Mish
  - $f(x) = x * \tanh(softplus(x)) \quad softplus(x) = \ln(1 + e^x)$
  - Аналогична ReLU при x>0;
  - Решение проблемы «умирающей ReLU»;
  - Немонотонная функция при x<0;
  - Во ряде случаев эффективнее ReLU и Swish.



Функция активации. Maxout

- Maxout
  - $\max(w_1^T x + b_1, w_2^T x + b_2)$
  - Обощение ReLU и Leaky ReLU
  - Удваивает число параметров



Функция активации. Другие виды

- Другие виды функций активации:
  - многообразие функций имеют сходую эффективность
  - применение «неканонических» функций только при существенных отличиях от известных аналогов.
- Не использовать функцию активации
  - $\boldsymbol{h} = g(\boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}) \rightarrow \boldsymbol{h} = g(\boldsymbol{V}^\top \boldsymbol{U}^\top \boldsymbol{x} + \boldsymbol{b})$
  - уменьшение числа весов: *np* → *(n+p)q*
- Радиальные базисные функции (RBF)
  - $h_i = \exp\left(-\frac{1}{\sigma_i^2}||\boldsymbol{W}_{:,i} - \boldsymbol{x}||^2\right)$
  - Сложности при оптимизации.
- Softplus
  - $g(a) = \zeta(a) = \log(1 + e^a)$
  - Теоретически лучше ReLU, практически менее эффективен.