Нейронные сети

Neural network

Нейронная сеть

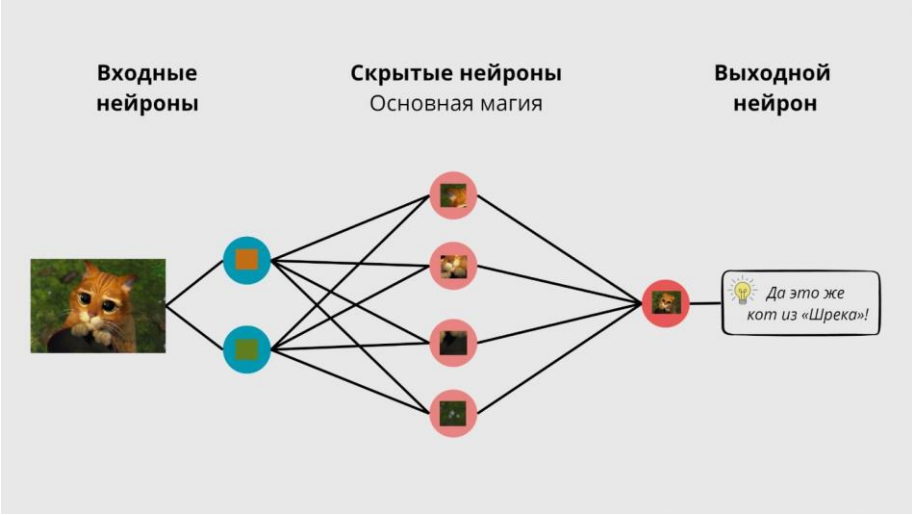
Нейронная сеть – компьютерная система, которая пытается имитировать работу человеческого мозга.

Они состоят из множества нейронов, которые соединены между собой для передачи информации. Эти системы обычно используют для обработки любого рода информации, например, для распознавания образов, классификации данных, прогнозирования.

Нейронная сеть

Нейрон — это вычислительная единица, которая получает информацию, производит над ней простые вычисления и передает ее дальше.

В том случае, когда нейросеть состоит из большого количества нейронов, вводят термин слоя. Есть **входной** слой, который получает информацию, n **скрытых** слоев, которые ее обрабатывают и **выходной** слой, который выводит результат.

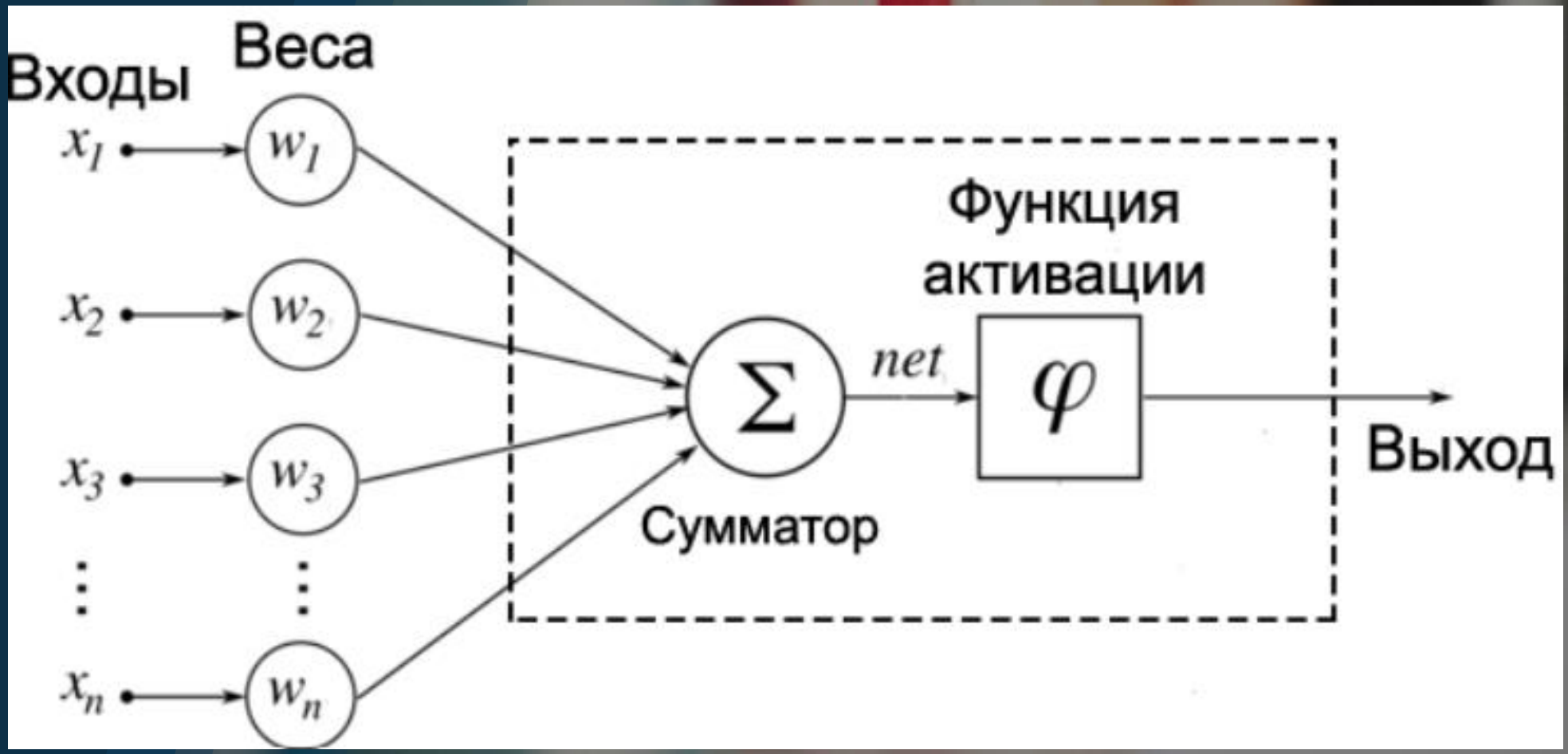


Нейронная сеть

У каждого из нейронов есть 2 основных параметра: входные данные (input data) и выходные данные (output data). В случае входного нейрона: input=output.

В остальных, в поле input попадает суммарная информация всех нейронов с предыдущего слоя, после чего, она нормализуется, с помощью функции активации (f(x)) и попадает в поле output.

Как работает нейронная сеть



Как работает нейронная сеть

1. Прохождение сигнала через слои нейронов.

На этом этапе данные поступают на вход системы и проходят через все слои. Каждый слой обрабатывает информацию по-разному, выделяя определенные признаки или структуры.

2. Обучение и корректировка веса.

На этом этапе нейронка «обучается» на основе примеров и корректирует свой вес таким образом, чтобы минимизировать ошибки в предсказании результатов. Для этого используется метод градиентного спуска.

3. Выдача результата.

Финальный этап представляет собой получение результата в виде предсказания, классификации, прогноза или рекомендации.

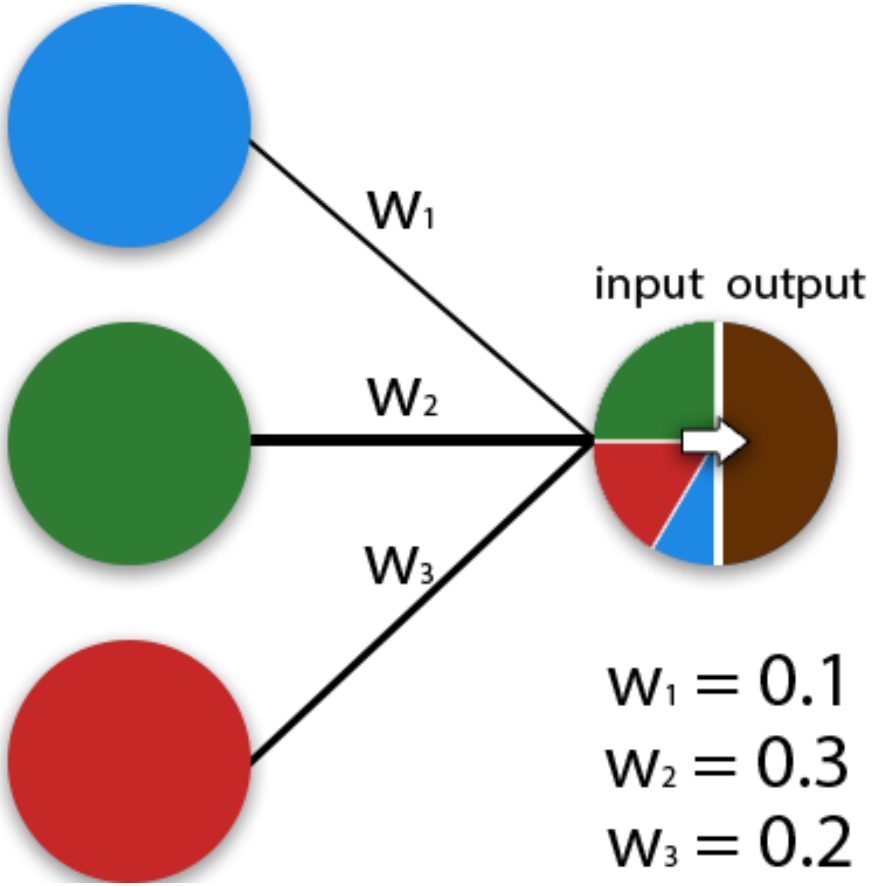
Синапсис

Синапс это связь между двумя нейронами.

У синапсов есть 1 параметр — **вес**. Благодаря ему, входная информация изменяется, когда передается от одного нейрона к другому.

Допустим, есть 3 нейрона, которые передают информацию следующему. Тогда у нас есть 3 веса, соответствующие каждому из этих нейронов. У того нейрона, у которого вес будет больше, та информация и будет доминирующей в следующем нейроне (пример — смешение цветов).

На самом деле, совокупность весов нейронной сети или матрица весов — это своеобразный мозг всей системы. Именно благодаря этим весам, входная информация обрабатывается и превращается в результат.

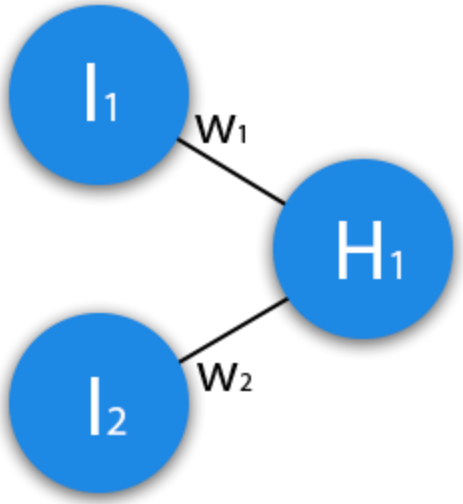


Нейронная сеть

В данном примере изображена часть нейронной сети, где буквами I обозначены входные нейроны, буквой H — скрытый нейрон, а буквой w — веса.

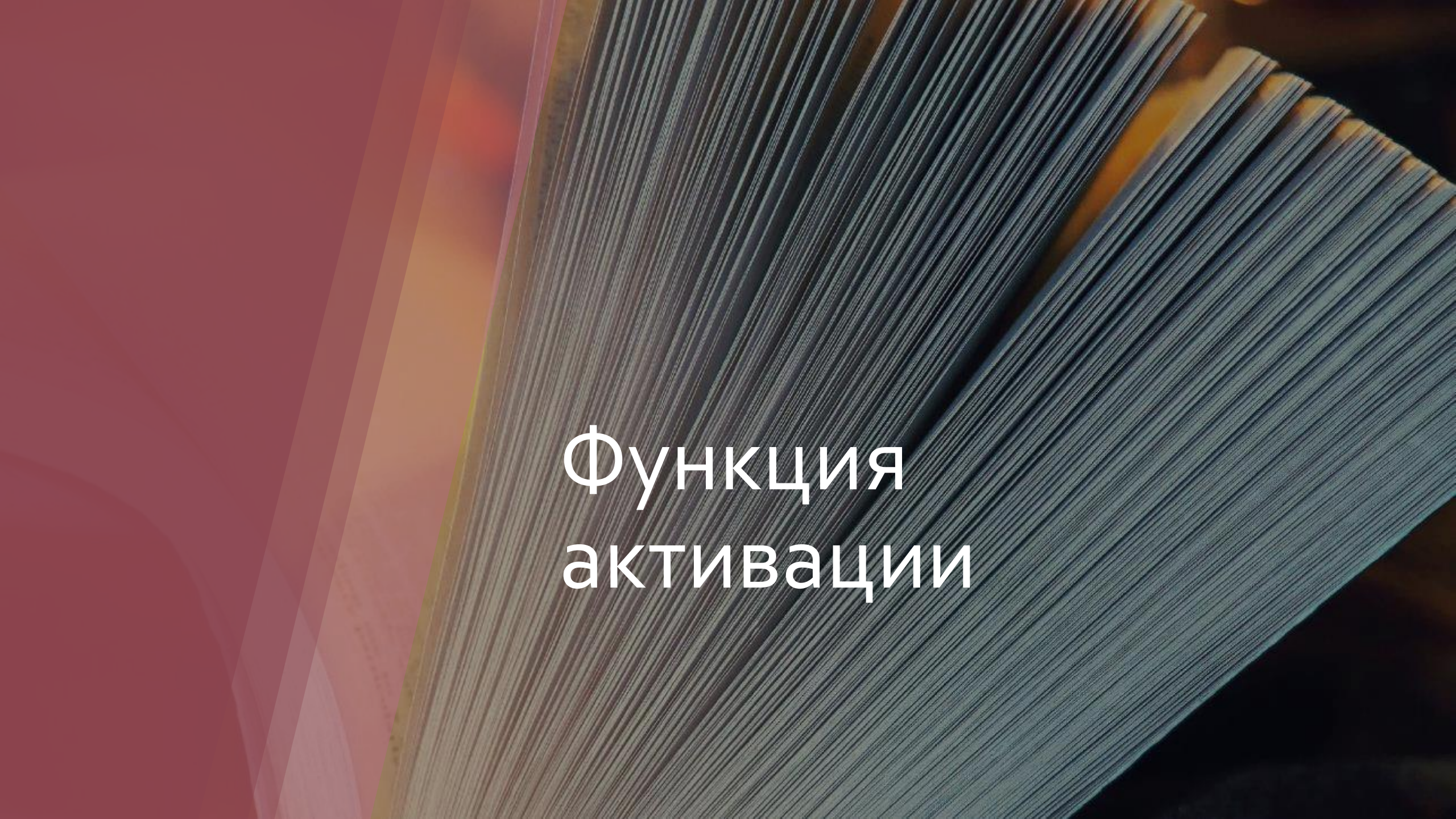
Из формулы видно, что входная информация — это сумма всех входных данных, умноженных на соответствующие им веса. Тогда дадим на вход 1 и 0. Пусть w1=0.4 и w2 = 0.7 Входные данные нейрона H1 будут следующими: 1*0.4+0*0.7=0.4. Теперь когда у нас есть входные данные, можем получить выходные данные, подставив входное значение в функцию активации. Выходные данные передаем дальше. И так, повторяем для всех слоев, пока не дойдем до выходного нейрона.

Запустив такую сеть в первый раз можно увидеть, что ответ далек от правильно, потому что сеть не натренирована. Чтобы улучшить результаты будем ее тренировать.



$$1) H_{1_{input}} = (I_1 * w_1) + (I_2 * w_2)$$

$$2) H_{1_{output}} = f_{activation}(H_{1_{input}})$$



Функция активации

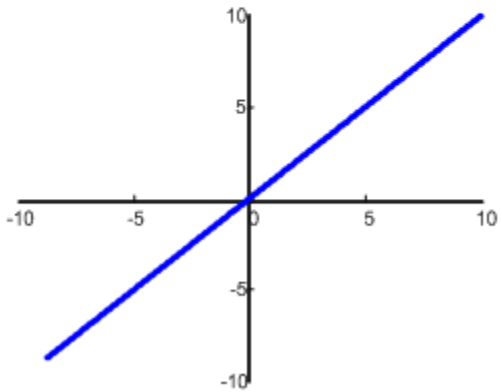
Линейная функция

Функция активации — это способ нормализации входных данных. То есть, если на входе будет большое число, пропустив его через функцию активации, можно получить выход в нужном диапазоне. Функций активации достаточно много. Главные их отличия — это диапазон значений.

Линейная функция

Эта функция почти никогда не используется, за исключением случаев, когда нужно протестировать нейронную сеть или передать значение без преобразований

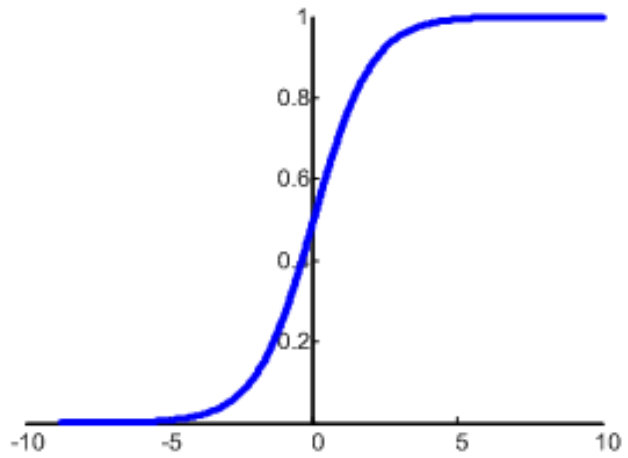
$$f(x) = x$$



Сигмоид

Это самая распространенная функция активации, ее диапазон значений [0,1]. Именно на ней показано большинство примеров в сети, также ее иногда называют логистической функцией.

$$f(x) = \frac{1}{1+e^{-x}}$$

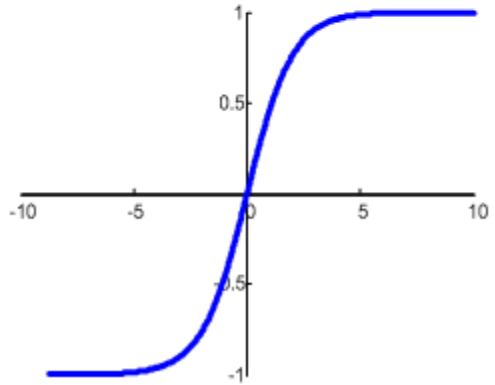


Гиперболический тангенс

Имеет смысл использовать гиперболический тангенс, только тогда, когда значения могут быть и отрицательными, и положительными, так как диапазон функции [-1,1].

Использовать эту функцию только с положительными значениями нецелесообразно так как это значительно ухудшит результаты нейросети.

$$f(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$$



Основные определения и термины

Тренировочный сет — это последовательность данных, которыми оперирует нейронная сеть.

Итерация - это своеобразный счетчик, который увеличивается каждый раз, когда нейронная сеть проходит один тренировочный сет. Другими словами, это общее количество тренировочных сетов пройденных нейронной сетью.

Эпоха

При инициализации нейронной сети эта величина устанавливается в 0 и имеет потолок, задаваемый вручную. Чем больше эпоха, тем лучше натренирована сеть и соответственно, ее результат. Эпоха увеличивается каждый раз, когда мы проходим весь набор тренировочных сетов.

Основные определения и термины

Ошибка — это процентная величина, отражающая расхождение между ожидаемым и полученным ответами. Ошибка формируется каждую эпоху и должна идти на спад. Если этого не происходит, значит, что-то делаете не так.

Ошибку можно вычислить разными путями, но мы рассмотрим лишь три основных способа: Mean Squared Error (далее MSE), Root MSE и Arctan. Каждый метод считает ошибки по разному. У Arctan, ошибка, почти всегда, будет больше, так как он работает по принципу: чем больше разница, тем больше ошибка. У Root MSE будет наименьшая ошибка, поэтому, чаще всего, используют MSE, которая сохраняет баланс в вычислении ошибки.

За каждый сет, считаем ошибку, отняв от идеального ответа, полученный. Далее, либо возводим в квадрат, либо вычисляем квадратный тангенс из этой разности, после чего полученное число делим на количество сетов.

Основные определения и термины

MSE

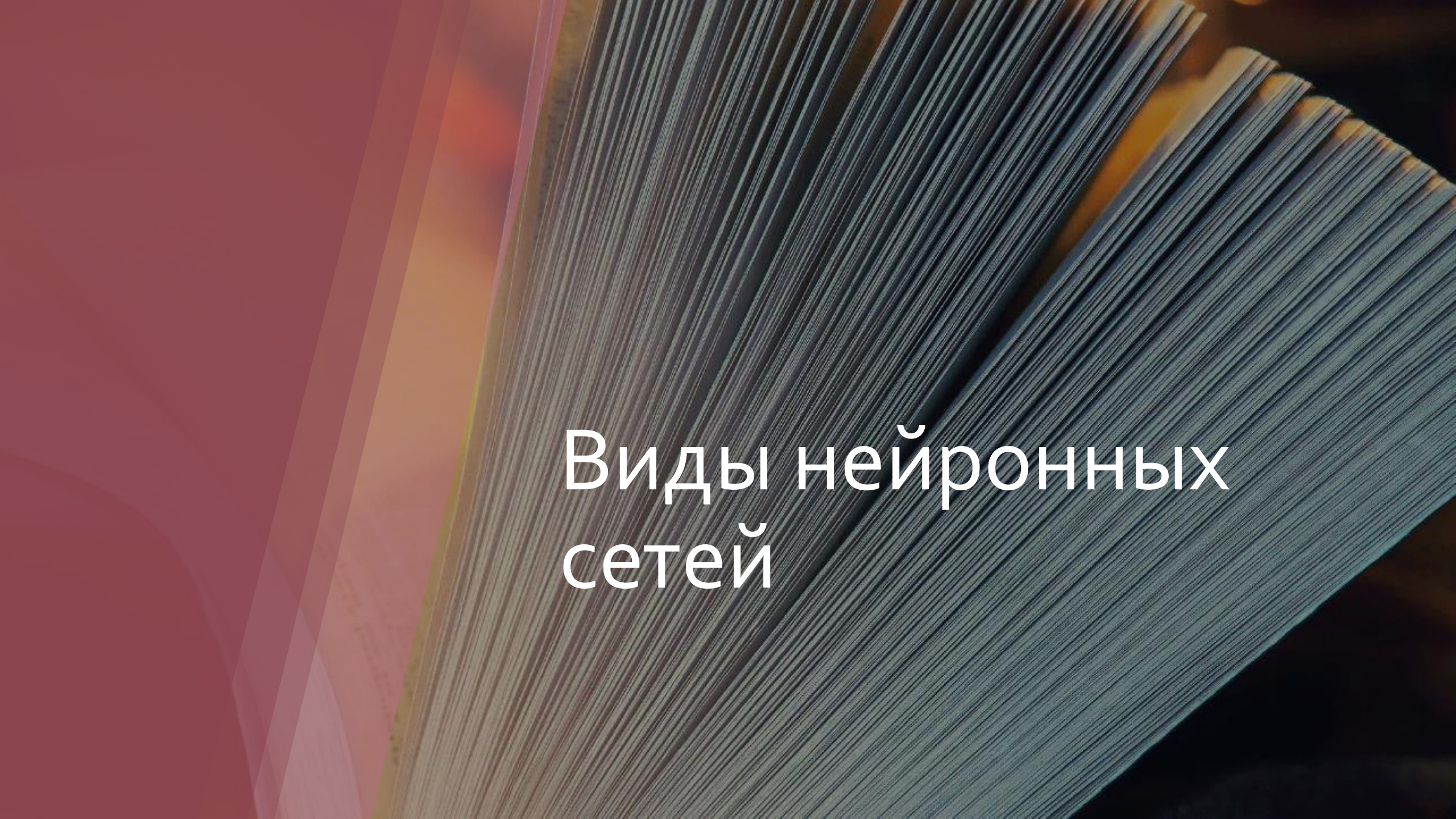
$$\frac{(i_1-a_1)^2+(i_2-a_2)^2+\ldots+(i_n-a_n)^2}{n}$$

Root MSE

$$\sqrt{\frac{(i_1-a_1)^2+(i_2-a_2)^2+\ldots+(i_n-a_n)^2}{n}}$$

Arctan

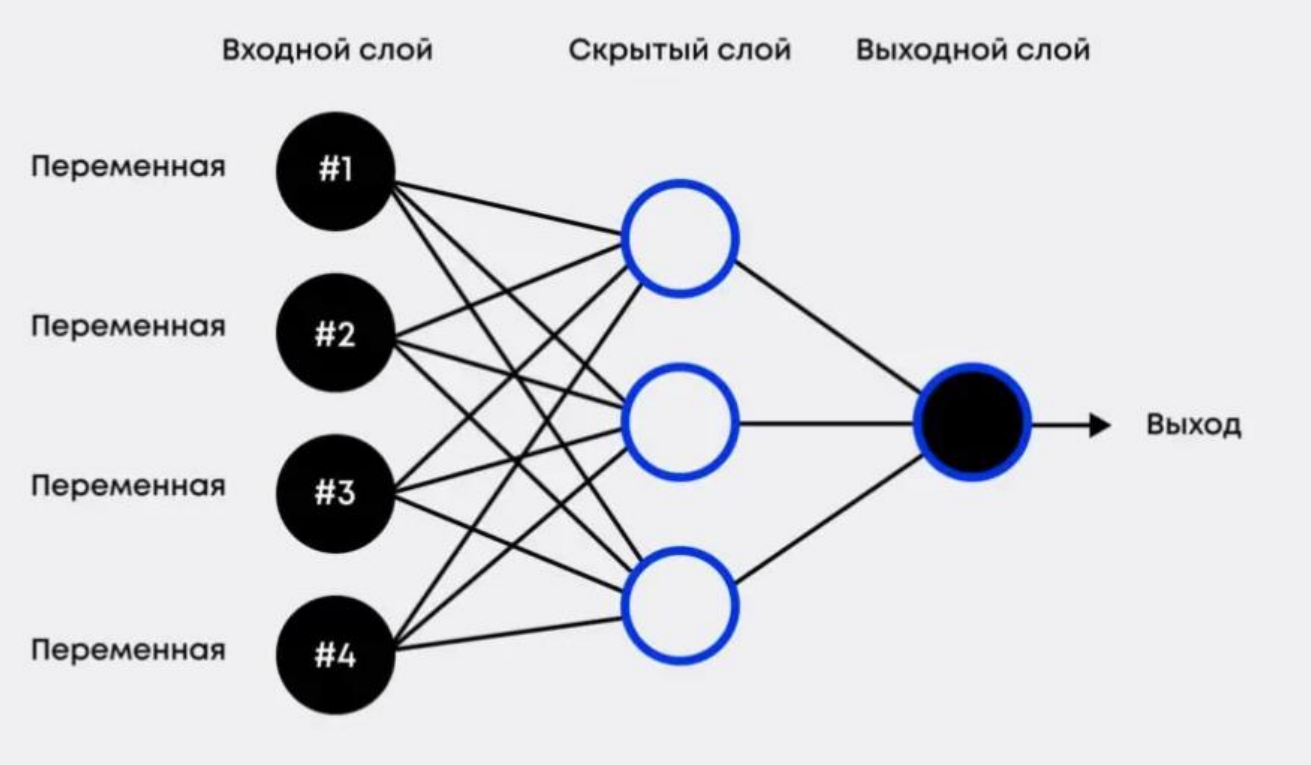
$$\frac{\arctan^2(i_1-a_1)+\ldots+\arctan^2(i_n-a_n)}{n}$$



Виды нейронных сетей

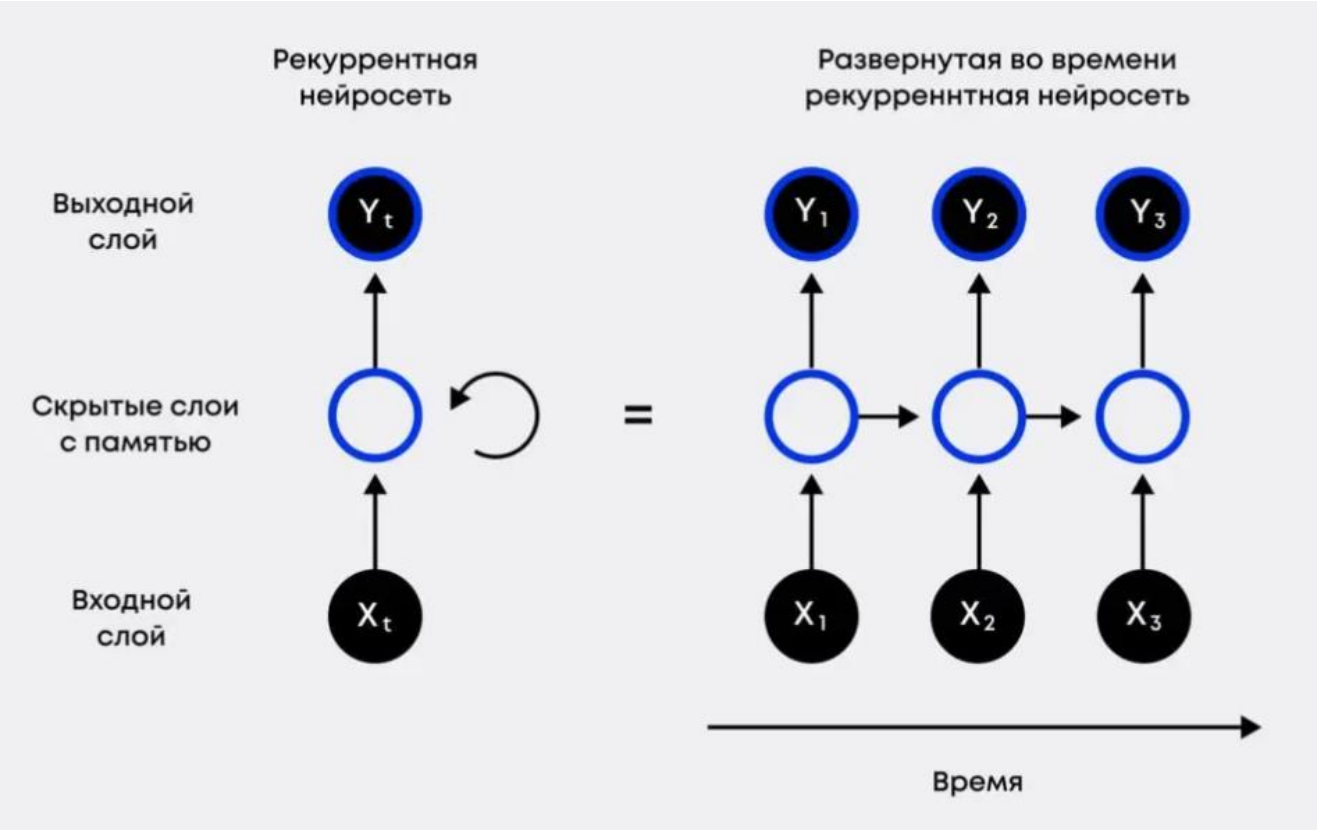
Прямые (Feedforward)

Самый простой и распространенный вид, в котором сигнал движется только в одном направлении от входного слоя через скрытые слои к выходному слою.



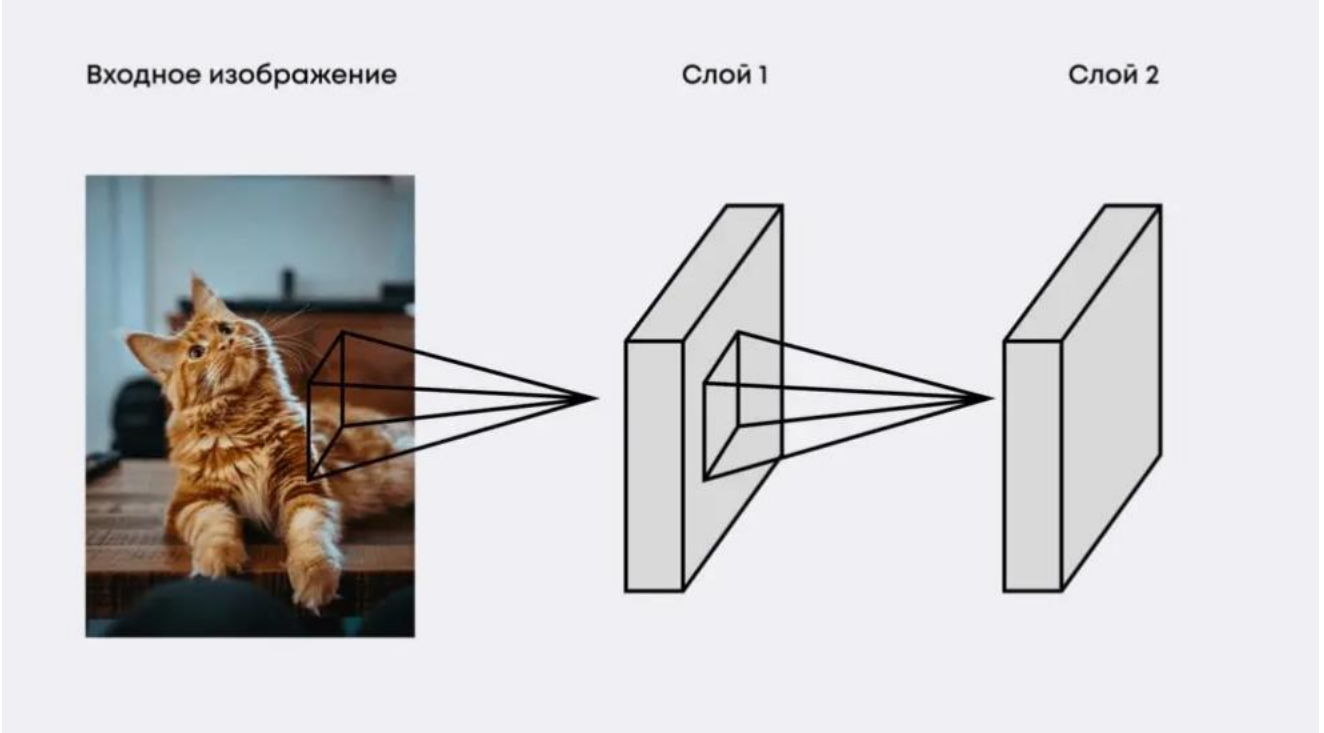
Рекуррентные (Recurrent)

Система, в которой возможно обратное движение сигнала от выхода к входу. Эти сети широко используются для распознавание речи, анализа временных рядов и генерации текста.



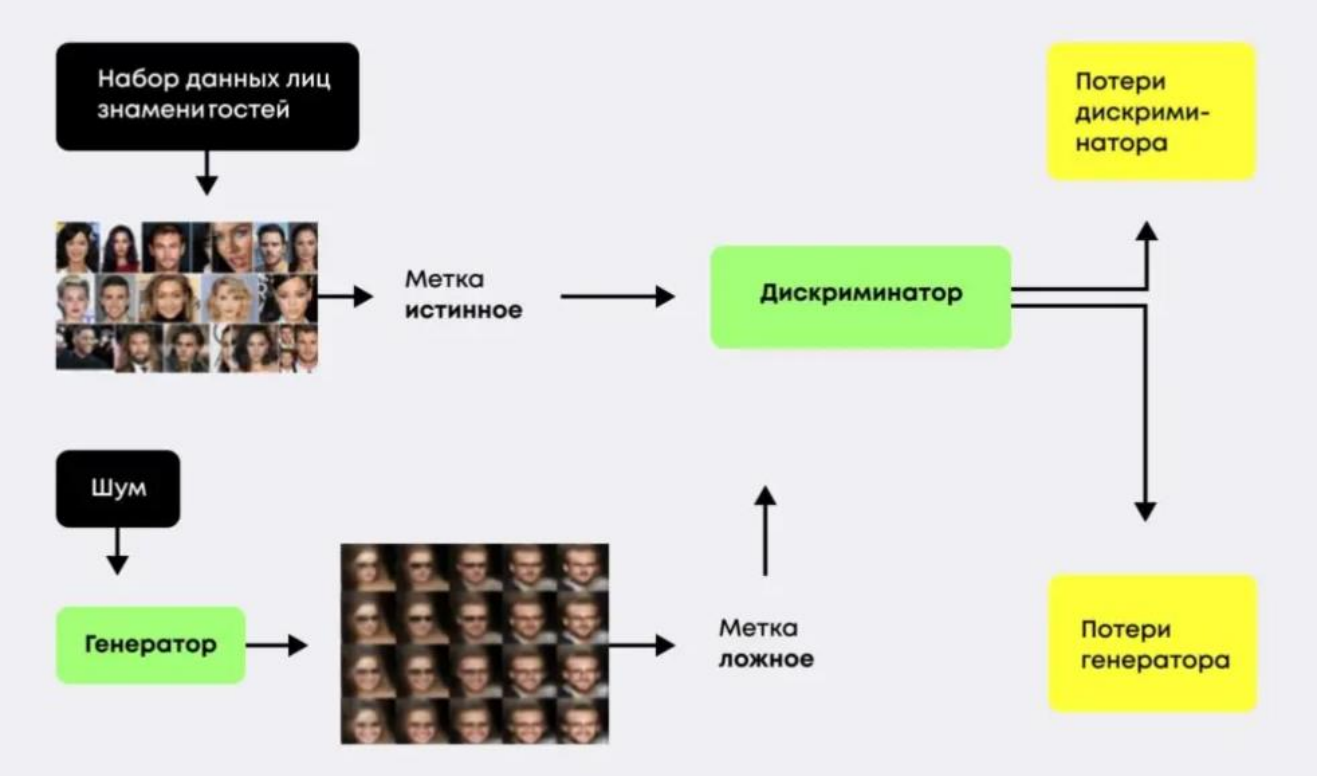
Сверхточные (Convolutional)

Используются для обработки изображений или других форматов сеток: графические модели. Они находят своё применение в распознавании объектов, сегментации и классификации изображений, а также в обработке и анализе видео.



Состязательные (Generative Adversarial)

Система, в которой две нейронные сети соревнуются друг с другом в процессе обучения: генеративная сеть создает новые примеры данных, а дискриминативная сеть пытается определить, является ли такой пример реальным или ложным.



Преимущества и недостатки

Преимущества:

— Нейронные сети могут обрабатывать большие объемы данных и тем самым предоставлять подробную информацию о том, или ином явлении;

— Создание и обучение не требует знания специфических физических или математических закономерностей, что делает их доступными для широкого круга специалистов;

— Нейронка способна обрабатывать информацию в режиме реального времени – это полезно в случаях, когда требуется оперативное решение;

— Они могут обучаться на существующих данных и делать точные предсказания в автоматическом режиме, что значительно увеличивает эффективность работы в различных предметных областях.

Преимущества и недостатки

Недостатки:

— Нейросети могут работать только с данными, которые были использованы при их обучении. Если данные меняются, модель может потребовать доработку, приемлемую точность или отказаться от использования данных вовсе;

— Создание и обучение занимают много времени и требуют значительных ресурсов, особенно при работе с крупными объемами данных;

— Результаты недостаточно точные и могут оказаться чрезмерно сложными для анализа;

— Нейронные сети не могут объяснить свои решения, что важно для задач с пояснительным анализом (например, медицинские диагнозы).

— Коррекция ошибок может быть трудной процедурой.