Тятя! Тятя! Нейросети заменили продавца!

Ппилиф Ульянкин

Аннотация

В этой виньетке собрана коллекция ручных задачек про нейросетки. Вместе с Машей можно попробовать по маленьким шажкам с ручкой и бумажкой раскрыть у себя в теле несколько чакр и немного глубже понять модели глубокого обучения[1].

Вместо введения

> Я попала в сети, которые ты метил, я самая счастливая на всей планете.
>
> *Юлианна Караулова*

Однажды Маша услышала про какой-то Машин лёрнинг. Она сразу же смекнула, что именно она — та самая Маша, кому этот лёрнинг должен принадлежать. Ещё она смекнула, что если хочет владеть лёрнингом по праву, ни одна живая душа не должна сомневаться в том, что она шарит. Поэтому она постоянно изучает что-то новое.

Её друг Миша захотел стать адептом Машиного лёрнинга, и спросил её о том, как можно за вечер зашарить алгоритм обратного распространения ошибки. Тогда Маша открыла свою коллекцию учебников по глубокому обучению. В каких-то из них было написано, что ей никогда не придётся реализовывать алгоритм обратного распространения ошибки, а значит и смысла тратить время на его формулировку нет[2]. В каких-то она находила слишком сложную математику, с которой за один вечер точно не разберёшься.[3] В каких-то алгоритм был описан понятно, но оставалось много недосказанностей[4].

Маша решила, что для вечерних разборок нужно что-то более инфантильное. Тогда она решила поскрести по лёрнингу и собрать коллекцию ручных задачек, прорешивая которую, новые адепты Машиного лёрнинга могли бы открывать у себя диплернинговые чакры. Так и появилась эта виньетка.

[1] Ахахах глубже глубокого, ахахах

[2] Франсуа Шолле, Глубокое обучение на Python

[3] Goodfellow I., Bengio Y., Courville A. Deep learning. – MIT press, 2016.

[4] Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей - Санкт-Петербург, 2018.

Содержание

Листочек 1: всего лишь функция 3

Листочек 2: что выплёвывает нейросеть 8

Листочек 3: пятьдесят оттенков градиентного спуска 10

Листочек 4: алгоритм обратного распространения ошибки 12

Листочек 5: всего лишь кубики LEGO 17

Листочек 6: свёрточные сети 22

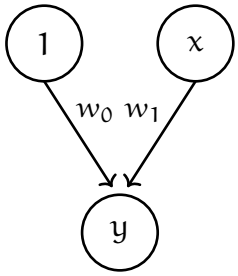
Листочек 7: рекуррентные сети 26

Листочек 1: всего лишь функция

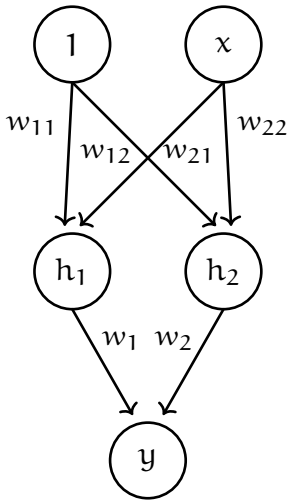
> Ты всего лишь машина, только имитация жизни. Робот сочинит симфонию? Робот превратит кусок холста в шедевр искусства?
>
> *Из фильма «Я, робот» (2004)*

Упражнение 1 (от регрессии к нейросетке)

Однажды вечером, по пути с работы[5] Маша зашла в свою любимую кофейню на Тверской. Там, на стене, она обнаружила очень интересную картину:



Хозяин кофейни, Добродум, объяснил Маше, что это Покрас-Лампас так нарисовал линейную регрессию, и её легко можно переписать в виде формулы: $y_i = w_0 + w_1 \cdot x_i$. Пока Добродум готовил кофе, Маша накидала у себя на бумажке новую картинку:

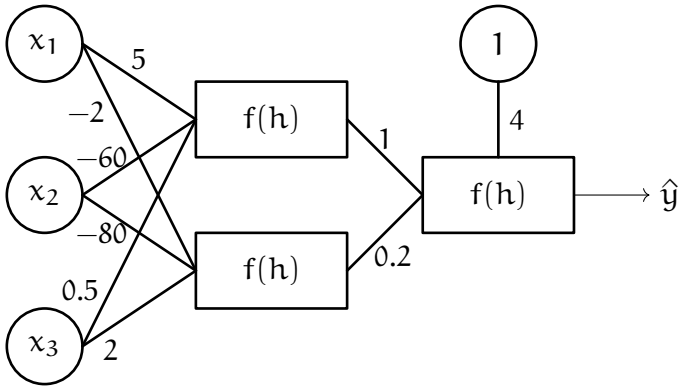


Как такая функция будет выглядеть в виде формулы? Правда ли, что y будет нелинейно зависеть от x? Если нет, как это исправить и сделать зависимость нелинейной?

[5] она работает рисёрчером.

Упражнение 2 (из картинки в формулу)

Добродум хочет понять, насколько сильно будет заполнена кофейня в следующие выходные. Для этого он обучил нейросетку. На вход она принимает три фактора: температуру за окном, x_1, факт наличия на Тверской митинга, x_2 и пол бариста на смене, x_3. В качестве функции активации Добродум использует ReLU.



а. В эти выходные за барной[6] стойкой стоит Агнесса. Митинга не предвидится, температура будет в районе 20 градусов. Спрогнозируйте, сколько человек придёт в кофейню к Добродуму?

б. На самом деле каждая нейросетка — это просто-напросто какая-то нелинейная сложная функция. Запишите нейросеть Добродума в виде функции.

Упражнение 3 (из формулы в картинку)

Маша написала на бумажке функцию:

$$y = \max(0, 4 \cdot \max(0, 3 \cdot x_1 + 4 \cdot x_2 + 1) + 2 \cdot \max(0, 3 \cdot x_1 + 2 \cdot x_2 + 7) + 6)$$

Теперь она хочет, чтобы кто-нибудь из её адептов нарисовал её в виде нейросетки. Нарисуйте.

Упражнение 4 (армия регрессий)

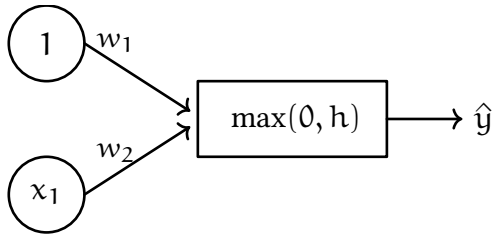
Парни очень любят Машу,[7] а Маша с недавних пор любит собирать персептроны и думать по вечерам об их весах и функциях активации. Сегодня она решила разобрать свои залежи из персептронов и как следует упорядочить их.

а. В ящике стола Маша нашла персептрон с картинки 1 Маша хочет подобрать веса так, чтобы он реализовывал логическое отрицание, то есть превращал $x_1 = 0$ в $y = 1$, а $x_1 = 1$ в $y = 0$.

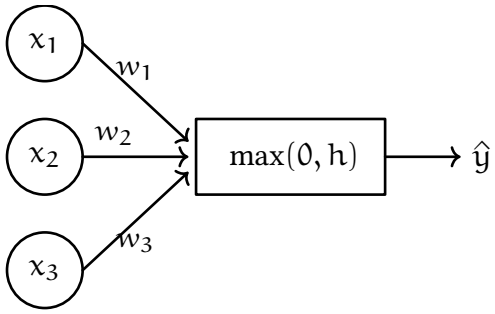
[6] барной... конечно, кофейня у него...

[7] когда у тебя есть лёрнинг, они так и лезут

Картинка 1



Картинка 2

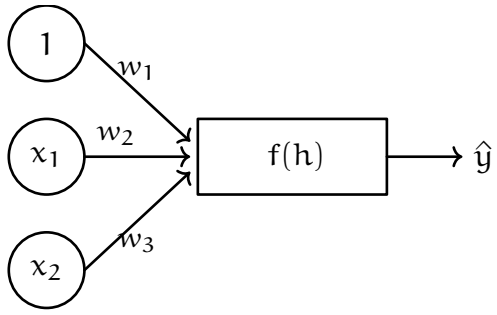


б. В тумбочке, среди носков, Маша нашла персептрон, с картинки 2, Маша хочет подобрать такие веса w_i, чтобы персептрон превращал x из таблички в соответствующие y:

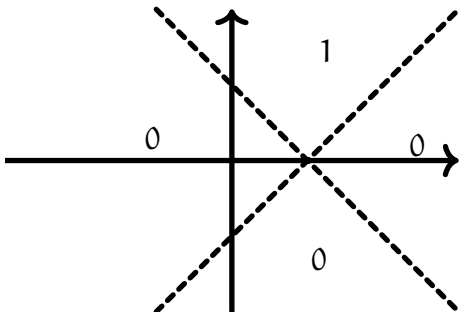
x_1	x_2	x_3	y
1	1	2	0.5
1	−1	1	0

в. Оказывается, что в ванной всё это время валялась куча персептронов с картинки 3 с неизвестной функцией активации.

Картинка 3



Картинка 4



Маша провела на плоскости две прямые: $x_1 + x_2 = 1$ и $x_1 - x_2 = 1$. Она хочет собрать из персептронов нейросетку, которая будет классифицировать объекты с плоскости так, как показано на картинке 4. В качестве функции возьмите единичную ступеньку (Функцию Хевисайда).

Упражнение 5 (логические функции)

Маша вчера поссорилась с Пашей. Он сказал, что у неё нет логики. Чтобы доказать Паше обратное, Маша нашла теорему, которая говорит о том, что с помощью нейросетки можно аппроксимировать почти любую функцию, и теперь собирается заняться аппроксимацией логических функций. Для начала она взяла самые простые, заданные следующими таблицами истинности:

x_1	x_2	$x_1 \cap x_2$
1	1	1
1	0	0
0	1	0
0	0	0

x_1	x_2	$x_1 \cup x_2$
1	1	1
1	0	1
0	1	1
0	0	0

x_1	x_2	x_1 XoR x_2
1	1	0
1	0	1
0	1	1
0	0	0

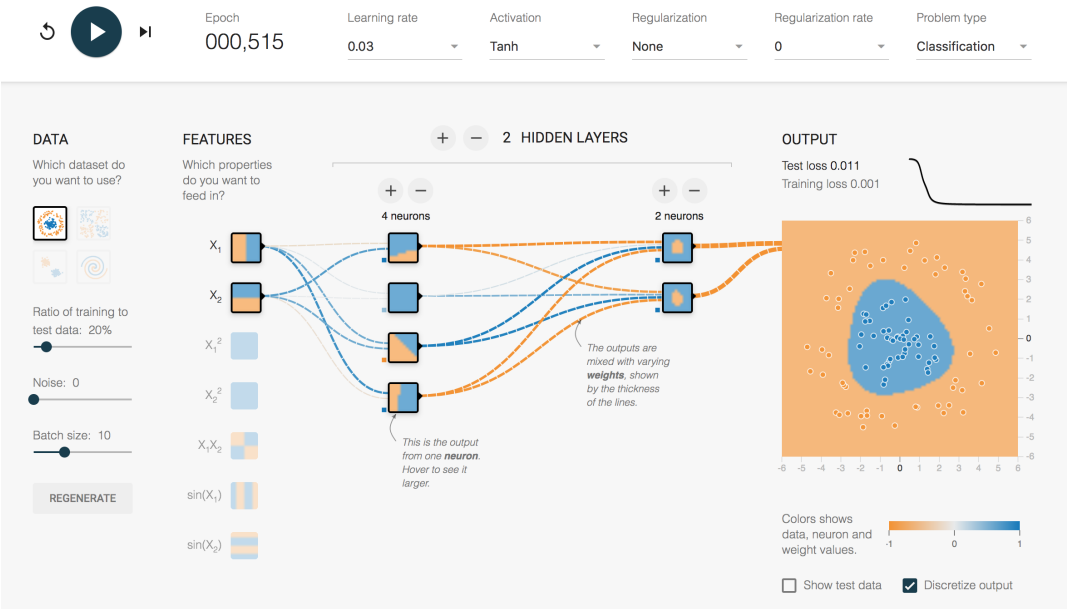
Первые два столбика идут на вход, третий получается на выходе. Первая операция — логическое "и вторая — "или". Операция из третьей таблицы называется "исключающим или (XoR). Если внимательно приглядеться, то можно заметить, что XoR — это то же самое что и $[x_1 \neq x_2]$[8].

Упражнение 6 (ещё немного про XoR)

Маша заметила, что на XoR ушло очень много персептронов. Она поняла, что первые два персептрона пытаются сварить для третьего нелинейные признаки, которых нейросетке не хватает. Она решила самостоятельно добавить персептрону вход $x_3 = x_1 \cdot x_2$ и реализовать XoR одним персептроном. Можно ли это сделать?

Упражнение 7 (избыток)

На сайте http://playground.tensorflow.org Маша стала играться с простенькими нейросетками и обучила для решения задачи классификации трёхслойного монстра.



Голубым цветом обозначен первый класс, рыжим второй. Внутри каждого нейрона визуализирована та разделяющая поверхность, которую он выстраивает. Так, первый слой ищет разделяющую линию. Второй слой пытается из этих линий выстроить более сложные фигуры

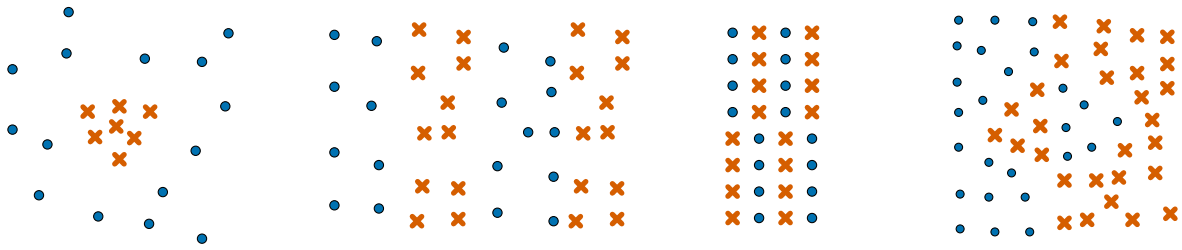
[8]Тут квадратные скобки обозначают индикатор. Он выдаёт 1, если внутри него стоит правда и 0, если ложь. Такая запись называется скобкой Айверсона. Попробуйте записать через неё единичную ступеньку Хевисайда.

и так далее. Чем ярче связь между нейронами, тем больше весовой коэффициент, относящейся к ней. Синие связи — положительные, рыжие — отрицательные. Чем тусклее связь, тем он ближе к нулю.

Маша заметила, что с её архитектурой что-то не так. Какие у неё проблемы?

Упражнение 8 (минималочка)

Шестилетняя сестрёнка ворвалась в квартиру Маши и разрисовала ей все обои:



Маша по жизни оптимистка. Поэтому она увидела не дополнительные траты на ремонт, а четыре задачи классификации. И теперь в её голове вопрос: сколько минимально нейронов нужно, чтобы эти задачи решить?

Упражнение 9 (универсальный регрессор)

Маша доказала Паше, что у неё всё в полном порядке с логикой. Теперь она собирается доказать ему, что с помощью двухслойной нейронной сетки можно приблизить любую непрерывную функцию от одного аргумента $f(x)$ со сколь угодно большой точностью[9].

Hint: Вспомните, что любую непрерывную функцию можно приблизить с помощью кусочно-линейной функции (ступеньки). Осознайте как с помощью пары нейронов можно описать такую ступеньку. Соедините все ступеньки в сумму с помощью выходного нейрона.

Упражнение 10 (число параметров)

Та, кому принадлежит машин лёрнинг собирается обучить полносвязную нейронную сеть для решения задачи регрессии, На вход в ней идёт 12 переменных, в сетке есть 3 скрытых слоя. В первом слое 300 нейронов, во втором 200, в третьем 100. Сколько параметров предстоит оценить Маше?

[9] http://neuralnetworksanddeeplearning.com/chap4.html

Листочек 2: что выплёвывает нейросеть

Плюют в душу обычно те, кому не удалось в неё влезть.

Пацанский паблик категории Б

Упражнение 11 (про сигмоиду)

Любую s-образную функцию называют сигмоидой. Наиболее сильно прославилась под таким названием функция $f(t) = \frac{e^t}{1+e^t}$. Слава о ней добралась до Маши и теперь она хочет немного поисследовать её свойства[10].

а. Что происходит при $t \to +\infty$? А при $t \to -\infty$?

б. Как связаны между собой $f(t)$ и $f(-t)$?

в. Как связаны между собой $f'(t)$ и $f'(-t)$?

г. Как связаны между собой $f(t)$ и $f'(t)$?

д. Найдите $f(0)$, $f'(0)$ и $\ln f(0)$.

е. Найдите обратную функцию $f^{-1}(t)$

ж. Как связаны между собой $\frac{d \ln f(t)}{dt}$ и $f(-t)$?

з. Постройте графики функций $f(t)$ и $f'(t)$.

и. Говорят, что сигмоида — это гладкий аналог единичной ступеньки. Попробуйте построить на компьютере графики $f(t)$, $f(10 \cdot t)$, $f(100 \cdot t)$, $f(1000 \cdot t)$. Как они себя ведут?

Упражнение 12 (про logloss)

У Маши три наблюдения, первое наблюдение — кит, остальные — муравьи. Киты кодируются $y_i = 1$, муравьи — $y_i = 0$. В качестве регрессоров Маша берёт номера наблюдений $x_i = i$. После этого Маша оценивает логистическую регрессию с константой. В качестве функции потерь используются логистические потери.

а. Выпишите для данной задачи функцию потерь, которую минимизирует Маша.

б. При каких оценках коэффициентов логистической регрессии эта функция достигает своего минимума?

Упражнение 13 (про softmax)

Маша чуть внимательнее присмотрелась к своему третьему наблюдению и поняла, что это не кит, а бобёр. Теперь ей нужно решать задачу классификации на три класса. Она решил использовать для этого нейросеть с softmax-слоем на выходе.

Маша уже обучила нейронную сетку и хочет построить прогнозы для двух наблюдений. Слой, который находится перед softmax выдал для этих двух наблюдений следующий резуль-

[10] Часть задач украдена отсюда: https://github.com/bdemeshev/mlearn_pro

тат: $(1, -2, 0)$ и $(0.5, -1, 0)$.

а. Чему равны вероятности получить кита, муравья и бобра для этих двух наблюдений?

б. Пусть первым был кит, а вторым бобёр. Чему будет равна logloss-ошибка?

в. Пусть у Маши есть два класса. Она хочет выучить нейросеть. Она может учить нейронку с одним выходом и сигмоидой в качестве функции активации либо нейронку с двумя выходами и softmax в качестве функции активации. Как выходы этих двух нейронок взаимосвязаны между собой?

г. Объясните, почему softmax считают сглаженным вариантом максимума.

Упражнение 14 (про разные выходы)

Та, в чьих руках находится лёрнинг, решила немного поэкспериментировать с выходами из своей сетки.

а. Для начала Маша решила, что хочет решать задачу классификации на два класса и получать на выходе вероятность принадлежности к первому. Что ей надо сделать с последним слоем сетки?

б. Теперь Маша хочет решать задачу классификации на K классов. Что ей делать с последним слоем?

в. Новые вводные! Маша хочет спрогнозировать рейтинг фильма на "Кинопоиске". Он измеряется по шкале от 0 до 10 и принимает любое непрерывное значение. Как Маша может приспособить для этого свою нейронку?

г. У Маши есть куча новостей. Каждая новость может быть спортивной, политической или экономической. Иногда новость может относится сразу к нескольким категориям. Как Маше собрать нейросетку для решения этой задачи? Как будет выглядеть при этом функция ошибки?

д. У Маши есть картинки с уточками и чайками. Маша хочет научить нейросеть искать на картинке птицу, обводить её в прямоугольник (bounding box), а затем классифицировать то, что попало в прямоугольник. Как должен выглядеть выход из такой нейросети? Как должна выглядеть функция потерь?

е. Маша задумалась, как можно спрогнозировать число людей в кафе так, чтобы на выходе сетка всегда прогнозировала целое число. Надо ли как-то при этом менять функцию потерь?

Hint: вспомните про пуассоновскую регрессию.

Листочек 3: пятьдесят оттенков градиентного спуска

Повторять до сходимости — это как жарить до готовности

Неизвестный студент Вышки

Упражнение 15 (50 оттенков спуска)

Маша Нестерова, хозяйка машин лёрнинга[11], собрала два наблюдения: $x_1 = 1, x_2 = 2$, $y_1 = 2, y_2 = 3$ и собирается обучить линейную регрессию $y = w \cdot x$. Маша очень хрупкая девушка, и ей не помешает помощь.

а. Получите теоретическую оценку методом наименьших квадратов.

б. Сделайте три шага градиентного спуска. В качестве стартовой точки используйте $w_0 = 0$. В качестве скорости обучения возьмите $\eta = 0.1$.

в. Сделайте четыре шага стохастического градиентного спуска. Пусть в SGD сначала попадает первое наблюдение, затем второе.

г. Если вы добрались до этого пункта, вы поняли градиентный спуск. Маша довольна. Начинаем заниматься тупой технической бессмыслицей. Сделайте два шага Momentum SGD. Возьмите $\alpha = 0.9, \eta = 0.1$

д. Сделайте два шага Momentum SGD с коррекцией Нестерова.

е. Сделайте два шага RMSprop. Возьмите $\alpha = 0.9, \eta = 0.1$

ж. Сделайте два шага Adam. Возьмём $\beta_1 = \beta_2 = 0.9, \eta = 0.1$

Упражнение 16 (логистическая регрессия)

Маша решила, что нет смысла останавливаться на обычной регрессии, когда она знает, что есть ещё и логистическая:

$$z = w \cdot x \qquad p = P(y = 1) = \frac{1}{1 + e^{-z}}$$
$$\text{logloss} = -[y \cdot \ln p + (1 - y) \cdot \ln(1 - p)]$$

Запишите формулу, по которой можно пересчитывать веса в ходе градиентного спуска для логистической регрессии.

Оказалось, что $x = -5$, а $y = 1$. Сделайте один шаг градиентного спуска, если $w_0 = 1$, а скорость обучения $\gamma = 0.01$.

Упражнение 17 (вопросики)

Убедитесь, что вы можете дать ответы на следующие вопросы:

[11] Лёрнинг ей папа подарил

- Как вы думаете, почему считается, что SGD лучше работает для оптимизации функций, имеющих больше одного экстремума?
- Предположим, что у функции потерь есть несколько локальных минимумов. Как можно адаптировать градиентный спуск так, чтобы он находил глобальный минимум чаще?
- Что будет происходить со стохастическим градиентным спуском, если длина его шага не будет уменьшаться от итерации к итерации?

Упражнение 18 (скорости обучения)

В стохастическом градиентном спуске веса изменяются по формуле

$$w_t = w_{t-1} - \eta_t \cdot \nabla L(w_{t-1}, x_i, y_i),$$

где наблюдение i выбрано случайно, скорость обучения зависит от номера итерации.

Условия Роббинса-Монро гарантируют сходимость алгоритма к оптимуму для выпуклых дифференцируемых функций. Они говорят, что ряд из скоростей $\sum_{t=0}^{\infty} \eta_t$ должен расходиться, а ряд $\sum_{t=0}^{\infty} \eta_t^2$ сходиться. То есть скорость спуска должна падать не слишком медленно, но и не слишком быстро. Какие из последовательностей, перечисленных ниже, можно использовать для описания изменения скорости алгоритма?

а. $\eta_t = \frac{1}{t}$
б. $\eta_t = \frac{0.1}{t^{0.3}}$
в. $\eta_t = \frac{1}{\sqrt{t}}$
г. $\eta_t = \frac{1}{t^2}$
д. $\eta_t = e^{-t}$
е. $\eta_t = \lambda \cdot \left(\frac{s_0}{s_0+t}\right)^p$, где λ, p и s_0 — параметры

Листочек 4: алгоритм обратного распространения ошибки

К толковому выбору приводит опыт,а к нему приводит выбор бестолковый.

JSON Стэтхэм

Упражнение 19 (граф вычислений)

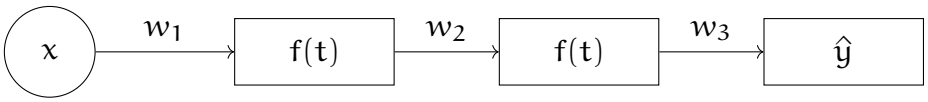
Как найти производную a по b в графе вычислений? Находим не посещённый путь из a в b, перемножаем все производные на рёбрах получившегося пути. Добавляем это произведение в сумму. Так делаем для всех путей. Маша хочет попробовать этот алгоритм на функции

$$f(x, y) = x^2 + xy + (x + y)^2.$$

Помогите ей нарисовать граф вычислений и найти $\frac{\partial f}{\partial x}$ и $\frac{\partial f}{\partial y}$. В каждой вершине графа записывайте результат вычисления одной элементарной операции: сложений или умножения[12].

Упражнение 20 (придумываем backpropagation)

У Маши есть нейросеть с картинки ниже, где w_k — веса для k слоя, $f(t)$ — какая-то функция активации. Маша хочет научиться делать для такой нейронной сетки градиентный спуск.



а. Запишите Машину нейросеть, как сложную функцию.

б. Предположим, что Маша решает задачу регрессии. Она прогоняет через нейросетку одно наблюдение. Она вычисляет знчение функции потерь $L(w_1, w_2, w_3) = \frac{1}{2} \cdot (y - \hat{y})^2$. Найдите производные функции L по всем весам w_k.

в. В производных постоянно повторяются одни и те же части. Постоянно искать их не очень оптимально. Выделите эти части в прямоугольнички цветными ручками.

г. Выпишите все производные в том виде, в котором их было бы удобно использовать для алгоритма обратного распространения ошибки, а затем, сформулируйте сам алгоритм. Нарисуйте под него удобную схемку.

Упражнение 21 (сигмоида)

В **не**глубоких сетях в качестве функции активации можно использовать сигмоиду

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z},$$

[12]По мотивам книги Николенко "Глубокое обучение"(стр. 79)

Маша хочет использовать сигмоиду внутри нейросети. Предполагается, что после прямого шага, наши вычисления будут использованы в другой части нейросети. В конечном итоге, по выходу из нейросети мы вычислим какую-то функцию потерь L.

У сигмоиды нет параметров. Чтобы обучить нейросеть, Маше понадобится производная $\frac{\partial L}{\partial z}$. Выпишите её в матричном виде через производные $\frac{\partial L}{\partial \sigma}$ и $\frac{\partial \sigma}{\partial z}$.

Упражнение 22 (линейный слой)

Маша знает, что главный слой в нейронных сетях — линейный. В матричном виде его можно записать как $Z = XW$.

Маша хочет использовать этот слой внутри нейросети. Предполагается, что после прямого шага наши вычисления будут использованы в другой части нейросети. В конечном итоге, по выходу из нейросети мы вычислим какую-то функцию потерь L.

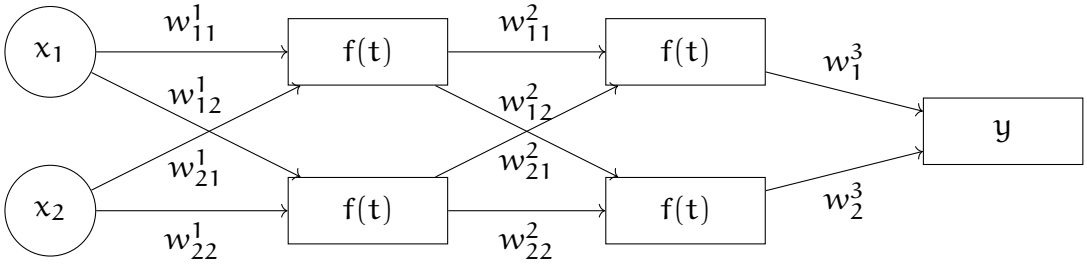
Чтобы обучить нейросеть, Маше понадобятся производные $\frac{\partial L}{\partial X}$ и $\frac{\partial L}{\partial W}$. Аккуратно найдите их и запишите в матричном виде[13]. Предполагается, что

$$X = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix} \qquad W = \begin{pmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{pmatrix}$$

$$Z = XW = \begin{pmatrix} z_{11} & z_{12} & z_{13} \\ z_{21} & z_{22} & z_{23} \end{pmatrix} = \begin{pmatrix} x_{11}w_{11} + x_{12}w_{21} & x_{11}w_{12} + x_{12}w_{22} & x_{11}w_{13} + x_{12}w_{23} \\ x_{21}w_{11} + x_{22}w_{21} & x_{21}w_{12} + x_{22}w_{22} & x_{21}w_{13} + x_{22}w_{23} \end{pmatrix}$$

Упражнение 23 (Backpropagation в матричном виде)

У Маши есть нейросеть с картинки ниже, где w_{ij}^k — веса для k слоя, $f(t)$ — какая-то функция активации. Маша хочет научиться делать для такой нейронной сетки градиентный спуск.



а. Запишите Машину нейросеть, как сложную функцию. Сначала в виде нескольких уравнений, а затем в матричном виде.

б. Выпишите все производные в том виде, в котором их было бы удобно использовать для алгоритма обратного распространения ошибки, а затем, сформулируйте сам алгоритм. Нарисуйте под него удобную схемку.

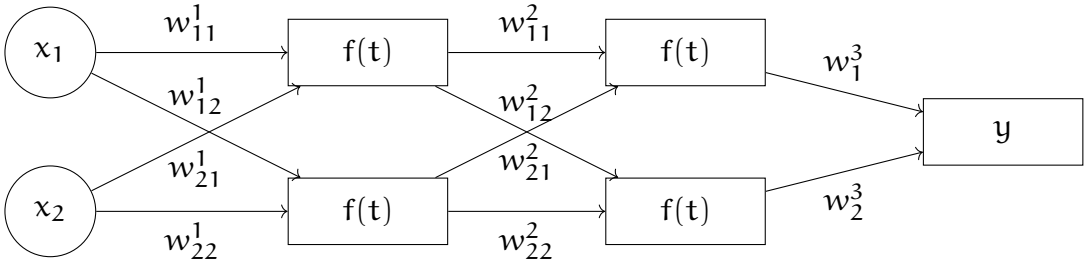
[13] https://web.eecs.umich.edu/~justincj/teaching/eecs442/notes/linear-backprop.html

Упражнение 24 (Backpropagation своими руками)

У Маши есть нейросеть с картинки ниже. Она использует функцию потерь

$$L(W_1, W_2, W_3) = \frac{1}{2} \cdot (\hat{y} - y)^2.$$

В качестве функции активации Маша выбрала сигмоиду $\sigma(t) = \frac{e^t}{1+e^t}$.



Выпишите для Машиной нейросетки алгоритм обратного распространения ошибки в общем виде. Пусть Маша инициализировала веса нейронной сети нулями. У неё есть два наблюдения

№	x_1	x_2	y
1	1	1	1
2	5	2	0

Сделайте руками два шага алгоритма обратного распространения ошибки. Пусть скорость обучения $\eta = 1$. Стохастический градиентный спуск решил, что сначала для шага будет использоваться второе наблюдение, а затем первое. Объясните, почему инициализировать веса нулями — плохая идея. Почему делать инициализацию весов любой другой константой — плохая идея?

Упражнение 25 (Незаметный backpropagation)

Маша собрала нейросеть:

$$y = \max\left(0; X \cdot \begin{pmatrix} 1 & -1 \\ 0.5 & 0 \end{pmatrix}\right) \cdot \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$

Теперь Маша внимательно смотрит на неё.

а) Первый слой нашей нейросетки — линейный. По какой формуле делается forward pass? Сделайте его для матрицы

$$X = \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix}.$$

б) Найдите для первого слоя производную выхода по входу. При обратном движении по нейросетке, в первый слой пришёл накопленный градиент

$$d = \begin{pmatrix} -0.5 & 0 \\ 0 & 0 \end{pmatrix}.$$

Каким будет новое накопленное значение градиента, которое выплюнет из себя линейный слой? По какой формуле делается backward pass?

в) Второй слой нейросетки — функция активации, ReLU. По какой формуле делается forward pass? Сделайте его для матрицы

$$H_1 = \begin{pmatrix} 2 & -0.5 \\ 0 & 1 \end{pmatrix}.$$

г) Найдите для второго слоя производную выхода по входу. При обратном движении по нейросетке во второй слой пришёл накопленный градиент

$$d = \begin{pmatrix} -0.5 & -1 \\ 0 & 0 \end{pmatrix}.$$

Каким будет новое накопленное значение градиента, которое выплюнет из себя ReLU? По какой формуле делается backward pass?

д) Третий слой нейросетки — линейный. По какой формуле делается forward pass? Сделайте его для матрицы

$$O_1 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}.$$

е) Найдите для третьего слоя производную выхода по входу. При обратном движении по нейросетке, в третий слой пришёл накопленный градиент $d = (-1, 0)^T$. Каким будет новое накопленное значение градиента, которое выплюнет из себя линейный слой?

ж) Мы решаем задачу Регрессии. В качестве функции ошибки мы используем

$$MSE = \frac{1}{2n} \sum (\hat{y}_i - y_i)^2.$$

Пусть для рассматриваемых наблюдений реальные значения $y_1 = 2, y_2 = 1$. Найдите значение MSE.

з) Чему равна производная MSE по прогнозу? Каким будет накопленное значение градиента, которое MSE выплюнет из себя в предыдущий слой нейросетки?

и) Пусть скорость обучения $\gamma = 1$. Сделайте для весов нейросети шаг градиентного спуска.

к) Посидела Маша, посидела, и поняла, что неправильно она всё делает. В реальности перед ней не задача регрессии, а задача классификации. Маша применила к выходу из нейросетки сигмоиду. Как будет для неё выглядеть forward pass?

л) В качестве функции потерь Маша использует logloss . Как для этой функции потерь вы-

глядит forward pass? Сделайте его.

м) Найдите для logloss производную прогнозов по входу в сигмоиду. Как будет выглядеть backward pass, если $y_1 = 0, y_2 = 1$? Как поменяется оставшаяся часть алгоритма обратного распространения ошибки?

Упражнение 26 (Нестеров и backprop)

К Маше приехал её папа и загрузил её интересным вопросом. В алгоритме обратного распространения ошибки мы можем делать шаг как минимум двумя способами:

а. Зафиксировали все w_{t-1}, нашли все градиенты, сделали сразу по всем весам шаг градиентного спуска.

б. Нашли градиенты для последнего слоя и сделали шаг для его весов, получили w_t^k. Для поиска градиентов предпоследнего слоя используем веса w_t^k, а не w_{t-1}^k. Все остальные слои обновляем по аналогии.

Как думаете, какой из способов будет приводить к более быстрой сходимости и почему[14]?

[14] Я придумал эту задачу и не смог найти статью, где делали бы что-то похожее. Если вы видели такую, пришлите мне её плиз.

Листочек 5: всего лишь кубики LEGO

Цитата про лего

автор цитаты

Функции активации

Желание - Ржавый - Семнадцать - Рассвет - Печь - Девять - Добросердечный - Возвращение на Родину - Один - Грузовой вагон.

Код активации Зимнего Солдата

Упражнение 27 (про сигмоиду)

Любую "sобразную функцию называют сигмоидой. Наиболее сильно прославилась под таким названием функция $f(t) = \frac{e^t}{1+e^t}$. Слава о ней добралась до Маши и теперь она хочет немного поисследовать её свойства.

а. Выпишите формулы для forward pass и backward pass через слой с сигмоидой.

б. Какое максимальное значение принимает производная сигмоиды? Объясните как это способствует затуханию градиента и параличу нейронной сети?

Упражнение 28 (про тангенс)

Функция $f(t) = \tanh(t) = \frac{2}{1+e^{-2t}} - 1$ называется гиперболическим тангенсом.

а. Что происходит при $t \to +\infty$? А при $t \to -\infty$?

б. Как связаны между собой $f(t)$ и $f'(t)$?

в. Выпишите формулы для forward pass и backward pass через слой с тангенсом.

г. Правда ли, что тангенс способствует затуханию градиента и параличу нейронной сети? Какое максимальное значение принимает производная тангенса?

д.

пункт про то, почему часто функцию юзают в RNN

Упражнение 29 (про ReLU)

Функция $f(t) = \mathrm{ReLU}(t) = \max(t, 0)$ называется ReLU.

а.

Задача про ReLU и сигмоиду (Николенко)

Задача про паралич сигмоиды и ReLU

Parametric Rectifier (PReLU) Выписать уравнения для бэкпропа по параметру α

Упражнение 30 (softplus)

Про то почему её не используют

Что-то про Mish и Swish

Упражнение 31 (температура генерации)

Иногда в функцию softmax добавляют дополнительный параметр T, который называют температурой. Тогда она приобретает вид

$$f(z) = \frac{e^{\frac{z_i}{T}}}{\sum_{k=1}^{K} e^{\frac{z_k}{T}}}$$

Обычно это делается, когда с помощью нейросетки нужно сгенерировать какой-нибудь новый объект. Пусть у нас есть три класса. Наша нейросеть выдала на последнем слое числа $1, 2, 5$.

а. Какое итоговое распределение вероятностей мы получим, если $T = 10$?

б. А если $T = 1$?

в. А если $T = 0.1$?

г. Какое распределение получится при $T \to 0$?

д. А при $T \to \infty$?

е. Предположим, что объектов на порядок больше. Например, это реплики, которые Алиса может сказать вам в ответ на какую-то фразу. Понятное дело, что вашей фразе будет релевантно какое-то подмножество ответов. Какое значение температуры сэмплирования T смогут сделать реплики Алисы непредсказуемыми? А какие сделают их однотипными?

Регуляризация

Цитата про регуляризацию

Автор цитаты

Упражнение 32 (Маша и покемоны)

Маша измерила вес трёх покемонов, $y_1 = 6$, $y_2 = 6$, $y_3 = 10$. Она хочет спрогнозировать

вес следующего покемона. Модель для веса покемонов у Маши очень простая, $y_i = \beta + \varepsilon_i$, поэтому прогнозирует Маша по формуле $\hat{y}_i = \hat{\beta}$.

Для оценки параметра β Маша использует следующую целевую функцию:

$$\sum (y_i - \hat{\beta})^2 + \lambda \cdot \hat{\beta}^2$$

а) Найдите оптимальное $\hat{\beta}$ при $\lambda = 0$.

б) Найдите оптимальное $\hat{\beta}$ при произвольном λ. Правда ли, что чем больше λ, тем меньше β?

в) Подберите оптимальное λ с помощью кросс-валидации leave one out («выкинь одного»). При такой валидации на первом шаге мы оцениваем модель на всей выборке без первого наблюдения, а на первом тестируем её. На втором шаге мы оцениваем модель на всей выборке без второго наблюдения, а на втором тестируем её. И так далее n раз. Каждое наблюдение является отдельным фолдом.

г) Найдите оптимальное $\hat{\beta}$ при λ_{CV}.

Упражнение 33 (а вот и моя остановочка)

Сделать задачу по связи ранней остановки и регуляризатора. Как в книжке про диплернинг

Упражнение 34 (дропаут)

Маша собирается обучить нейронную сеть для решения задачи регрессии, На вход в неё идёт 12 переменных, в сетке есть 3 скрытых слоя. В пером слое 300 нейронов, во втором 200, в третьем 100.

a) Сколько параметров предстоит оценить Маше? Сколько наблюдений вы бы на её месте использовали?

b) Пусть в каждом слое была отключена половина нейронов. Сколько коэффициентов необходимо оценить?

c) Предположим, что Маша решила после первого слоя добавить в свою сетку Dropout с вероятностью p. Какова вероятность того, что отключится весь слой?

d) Маша добавила Dropout с вероятностью p. после каждого слоя. Какова вероятность того, что один из слоёв отключится и сетка не сможет учиться?

e) Пусть случайная величина N — это число включённых нейронов. Найдите её математическое ожидание и дисперсию. Если Маша хочет проредить сетку на четверть, какое значение p она должна поставить?

f) Пусть случайная величина P — это число параметров в нейросети, которое необходимо оценить. Найдите её математическое ожидание и дисперсию. Почему найденное вами

математическое ожидание выглядит очень логично? Что оно вам напоминает? Обратите внимание на то, что смерть одного из параметров легко может привести к смерти другого.

Добавить вопросиков про дропконнект

Бэкпроп через дропаут

Нормализация по батчам

Чашка хорошего чая восстановит мою нормальность.

Артур из «Автостопом по галактике»

Бэкпроп через батчнорм, смысл батчнорма, нарисовать граф через него, про то что это уродливая процедура

родить задачу из статьи dropout vs batchnorm

Инициализация

цитата об этом

автор

Упражнение 35 (инициализация весов)

а. Маша использует для активации симметричную функцию. Для инициализации весов она хочет использовать распределение

$$w_i \sim U\left[-\frac{1}{\sqrt{n_{in}}}; \frac{1}{\sqrt{n_{in}}}\right].$$

Покажите, что это будет приводить к затуханию дисперсии при переходе от одного слоя к другому.

б. Какими нужно взять параметры равномерного распределения, чтобы дисперсия не затухала?

в. Маша хочет инициализировать веса из нормального распределения. Какими нужно взять параметры, чтобы дисперсия не затухала?

г. Несимметричный случай

Упражнение 36 (ReLU и инициализация весов)

Внутри нейрона в качестве функции активации используется ReLU. На вход идёт 10 признаков. В качестве инициализации для весов используется нормальное распределение, $N(0, 1)$. С какой вероятностью нейрон будет выдавать на выход нулевое наблюдение, если

Предположения на входы? Какое распределение и с какими параметрами надо использовать, чтобы этого не произошло? Сюда же про инициализацию Хе.

Упражнение 37 (сложная задача про инициализацию)

Так чтобы плотность частного надо было искать и все офигели (Воронцов)

Стрельба по ногам

цитата об этом

автор

Упражнение 38 (Проблемы с архитектурой)

Миша принёс Маше несколько разных архитектур. Они выглядят довольно странно. Помогите Маше разобраться, что именно Миша сделал неправильно.

Упражнение 39 (Ещё один backpropagation)

У Маши есть трёхслойная нейросеть:

$$y = f(f(X \cdot W_3) \cdot W_2) \cdot W_1$$

а. Маща использует в качестве функции активации $f(t) = \mathrm{ReLU}(t) = \max(0; t)$, а в качестве функции потерь $L(W_1, W_2) = \frac{1}{2} \cdot (y - \hat{y})^2$.

б.

в.

г.

Для всех пунктов запишите уравнения для прямого и обратного проходов по сетке. Выпишите для всех весов уравнения, по которым будет делаться шаг градиентного спуска.

Листочек 6: свёрточные сети

> Урра! Отлично сработано, ребятки. Давайте завтра не придем? Возьмем отгул на денек? Вы пробовали шаурму? В двух кварталах отсюда делают какую-то шаурму. Не знаю, что это, но мне хочется.
>
> *Тони Старк (Мстители, 2012)*

Упражнение 40 (Свёртка своими руками)

У Маши есть картинка и свёртка, которую она хочет применить к этой картинке.

3	3	2	1	0
0	0	1	3	1
3	1	2	2	3
3	1	2	2	3
3	1	2	2	3

Картинка Маши

0	1	2
2	2	0
0	1	2

Свёртка Маши

а. Сделайте свёртку картинки без сдвигов и дополнений. К тому, что получилось применить max pooling и average pooling размера 2×2.

б. Примените к исходной картинке свёртку с параметром сдвига (stride) равным 2.

в. Примените к исходной картинке свёртку с параметром сдвига (stride) равным 3.

г. Примените к исходной картинке свёртку с дополнением нулями (zero padding) и параметром сдвига (stride) равным 0.

Упражнение 41 (Ядра)

У Маши есть куча ядер для свёрток. Догадайтесь какое из них что делает:

0	0	0
0	1	0
0	0	0

-1	-1	-1
-1	8	-1
-1	-1	-1

-1	-1	-1
-1	8	-1
-1	-1	-1

-1	-1	-1
-1	8	-1
-1	-1	-1

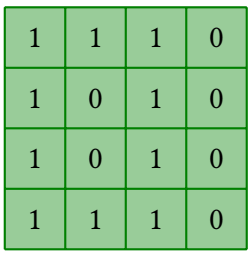
-1	-1	-1
-1	8	-1
-1	-1	-1

Упражнение 42 (Крестики и нолики)

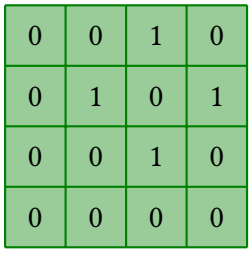
Маша хочет научить компьютер играть в крестики и нолики. На первом шаге ей надо научить алгоритм распознавать есть ли крестик на картинке. Под ноликом понимается любая фигура с дырой в середине. Под крестиком понимается любая фигура из двух пересекающихся линий.

Алгоритм должен быть устроен следующим образом. На первом шаге одна или несколько свёрток проходят по картинке. На втором шаге по результатам свёрток принимается решение. Например, берётся максимальное получившееся число и сравнивается с каким-то порогом. Классификатор крестиков и ноликов должен работать безупречно. Помогите Маше придумать такой классификатор.

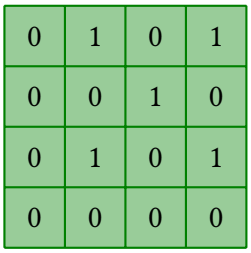
а. В мире Маши на картинках могут быть нарисованы либо крестики либо нолики. Все картинки, подающиеся на вход алгоритма могут быть только размера 4×4. Примеры крестиков и ноликов нарисованы ниже.



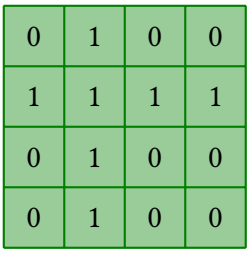
нолик



нолик



крестик



крестик

б. Предположим, что теперь кроме крестиков и ноликов в нашем мире существуют ещё и другие любые картинки. Как можно модернизировать ваш алгоритм, чтобы он по-прежнему стабильно работал с безупречным качеством?

Упражнение 43 (Свёрточный и полносвязный)

Маше рассказали, что свёрточный слой — это полносвязный слой с некоторыми ограничениями. Она хочет разобраться, что это за ограничения. На вход в слой идёт чёрно-белое изображение размера 4×4. Каждый пиксель изображения — отдельная переменная.

а. Нарисуйте с помощью кругляшей и стрелочек полносвязным слой, который обрабатывает картинку. Подпишите все веса. Нарисуйте свёрточный слой в таком же формате. На картинке часть связей исчезнет, а часть весов станет одинаковой.

б. Запишите свёрточный слой с помощью перемножения матриц в виде $H = X \cdot W$. Как выглядит матрица W? Как через свёрточный слой можно сделать шаг обратного распространения ошибки?

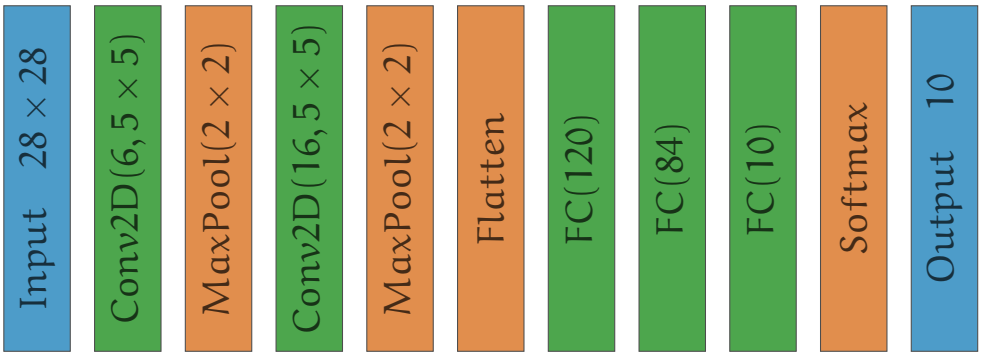
Упражнение 44 (Число параметров)

На вход в нейронную сетку идёт изображение рукописной цифры размера 28×28.

а. Маша вытягивает изображение в длинный вектор и использует полносвязную сетку для классификации изображений. В сетке идёт один полносвязный слой из 1000 нейронов. После идёт слой, который осуществляет классификацию изображения на 10 классов.

Сколько параметров нужно оценить?

б. Маша вместо полносвязной сетки использует свёрточную. На первом шаге делается 6 свёрток размера 5×5. На втором шаге делается max-pooling размера 2×2. На третьем 16 свёрток размера 5×5. На четвёртом max-pooling размера 2×2. На пятом картинка вытягивается в длинный вектор. Далее идут три полносвязных слоя размеров $120, 84, 10$. В конце делается softmax. После каждой свёртки и полносвязного слоя, кроме последнего, в качестве функции активации используется ReLU.



Сколько параметров необходимо будет оценить в такой модели? Какого размера будут выходы из каждого слоя?

в. Маша использует архитектуру из пункта б, но все свёртки делает с дополнением нулями (zero padding). Как изменится число оцениваемых параметров? Какого размера будут выходы из каждого слоя?

г. Маша использует архитектуру из пункта б, но все свёртки делает с дополнением нулями (zero padding) и параметром сдвига (stride) равным 2. ак изменится число оцениваемых параметров? Какого размера будут выходы из каждого слоя?

Упражнение 45 (Поле обзора)

Маша хочет найти котика размера 512×512 пикселей. Для этого она использует свёртки размера 5×5 без дополнения нулями (zero padding) После каждого свёрточного слоя Маша делает max-pooling.

а. Через сколько слоёв поле восприятия Машиной нейросетки впервые охватит котейку?

б. Маша хочет поменять max-pooling на свёртки со сдвигом (stride) так, чтобы котейка находился за такое же число слоёв. Какой размер сдвига ей надо выбрать?

в. Пусть s — величина сдвига, k — размер свёртки, m — размер пулинга, n — номер слоя. Выпишите формулу, по которой можно найти размер поля видимости (receptive field).

Упражнение 46 (Снова число параметров)

Маша собирает разные архитектуры. Помогите ей оценить число параметров для каждой из них.

а. У Маши есть свёрточный слой. На вход в свёрточный слой идёт изображение с C_{in} каналами размера $W \times H$. Маша использует C_{out} фильтров размера $W_k \cdot H_k$. Сколько параметров ей предстоит оценить?

б.

в.

Сюда вариант с подсчётом числа параметров в сепарабельной свёртке

Упражнение 47 (Скользящее среднее)

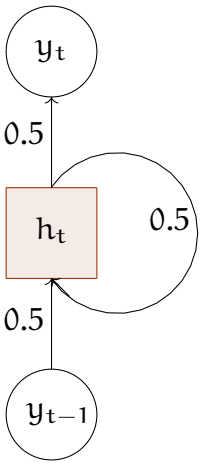
Скользящее среднее — это свёртка, которая работает для вектора. Опишите как именно она работает. Какой физический смысл стоит за размером такой свёртки и дополнением нулями?

Листочек 7: рекуррентные сети

> А сегодня в завтрашний день, не все могут смотреть. Вернее смотреть могут не только лишь все, не каждый может это делать.
>
> *Рекуррентная сеть глубины один*

Упражнение 48 (Туда и обратно)

Маша хочет сделать шаг обратного распространения ошибки через рекуррентную ячейку для последовательности $y_0 = 0, y_1 = 1, y_2 = -1, y_3 = 2$. Скрытое состояние инициализировано как $h_0 = 0$. Все веса инициализированы как 0.5. Во всех уравнениях, описывающих ячейку нет констант. В качестве функций активаций Маша использует ReLU. В качестве функции потерь Маша использует MSE.



а. Сделайте прямой шаг через ячейку. Для каждого элемента последовательности постройте прогноз. Посчитайте значение ошибки.

б. Сделайте обратный шаг распространения ошибки. Посчитайте для каждого из весов градиенты и обновите значения весов.

Упражнение 49 (Число параметров)

У Маши есть очень длинный временной ряд. Она хочет обучить несколько нейросетей предсказывать его дальнейшее значение. В своих моделях Маша нигде не использует константы.

а. Маша выделяет окно длины 100. Оно движется по последовательности. Для каждого окна Маша предсказывает следующее значение в ряду. В сетку подаются наблюдения с 1—го по 100—е. Прогнозируется 101—ое наблюдение. Затем на вход подаются наблюдения со 2—го по 100—е. Прогнозируется 102—ое наблюдение. И так далее до конца последовательности.

   На первом слое используется 20 нейронов. На втором слое используется один нейрон. Сколько параметров нужно оценить?

б. Маша использует одну простую RNN-ячейку. Сколько параметров ей необходимо оценить?

в. Маша хочет предсказывать значение y_t по трём последовательностям y_{t-1}, y_{t-2} и y_{t-3}. На первом слое сети Маша использует два рекуррентных нейрона. На втором слое она использует один рекуррентный нейрон. Матрица какого размера идёт на вход в первый слой? Матрица какого размера передаётся во второй слой? Какое число параметров необходимо оценить Маше?

г. Мы находимся в условиях прошлого пункта, но используетм LSTM-ячейки с забыванием. Сколько параметров надо оценить?

д. Мы находимся в условиях прошлого пугкта, но используем GRU-ячейки. Сколько параметров надо оценить?

Упражнение 50 (Из картинки в формулу)

У Маши есть два рекуррентных нейрона. Помогите ей изобразить их в виде вычислительных графов.

Однонаправленный:

$$\begin{aligned} h_t &= f_h(b_h + W \cdot h_{t-1} + V \cdot x_t) \\ y_t &= f_y(b_y + U \cdot h_t) \end{aligned}$$

Двунаправленный:

$$\begin{aligned} h_t &= f_h(b_h + W \cdot h_{t-1} + V \cdot x_t) \\ s_t &= f_s(b_s + W' \cdot s_{t+1} + V' \cdot x_t) \\ o_t &= b_y + U \cdot h_t + U' \cdot s_t \\ y_t &= f_y(o_t) \end{aligned}$$

Упражнение 51 (Замочные скважины)

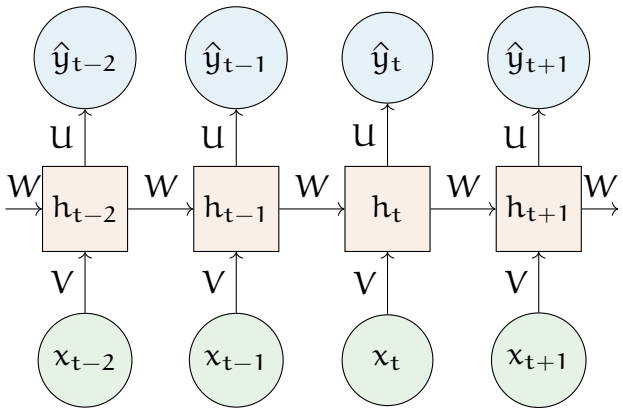
В 2000 году Шмидхубер и Герс предложили модификацию LSTM с замочными скважинами. Она описывается следующей системой из уравнений

$$\begin{aligned} c'_t &= \phi_c(W_c x_t + V_c h_{t-1} + b_c) \\ i_t &= \phi_i(W_i x_t + V_i h_{t-1} + U_i c_{t-1} + b_i) \\ f_t &= \phi_f(W_f x_t + V_f h_{t-1} + U_f c_{t-1} + b_f) \\ o_t &= \phi_o(W_o x_t + V_o h_{t-1} + U_o c_{t-1} + b_o) \\ c_t &= f_t \odot c_{t-1} + i_t \odot c'_t \\ h_t &= o_t \odot \phi_h(c_t) \end{aligned}$$

Изобразите эту ячейку в виде вычислительного графа. Объясните, чем именно она отличается от базовой модификации LSTM. Какой в этом смысл?

Упражнение 52 (Лишние части)

Выпишите уравнения, описывающие LSTM-ячейку с забыванием и GRU-ячейку. Какие последовательности и веса нужно занулить, чтобы эти ячейки превратились в простую RNN-ячейку?


$\hat{y}_{t-2}$
$\hat{y}_{t-1}$
$\hat{y}_t$
$\hat{y}_{t+1}$
U
W
h_{t-2}
h_{t-1}
h_t
h_{t+1}
V
x_{t-2}
x_{t-1}
x_t
x_{t+1}

какие-нибудь упражнения про w2v

упражнение про разные модные виды ячеек типа резнетов и тп

задачи про атеншны и тп