Секция «Слабый и сильный искусственный интеллект в управленческих практиках»

Изучение самостоятельного обучения с помощью преобразователей внимания соседей

Научный руководитель – Беспалов Михаил Евгеньевич

Самодуров Максим Алексеевич

Студент (магистр)

Российский государственный университет им. А.Н. Косыгина, Москва, Россия

E-mail: samodurov.maxim@yandex.ru

Изучение самостоятельного обучения с помощью преобразователей внимания соседей Авторы: Самодуров М.А. Аннотация Методы, основанные на данных, достигли большого прогресса в широком спектре приложений машинного зрения и анализа данных благодаря новым возможностям сбора, аннотирования и обработки огромных объемов данных, причем обучение с учителем дало самые впечатляющие результаты. К сожалению, чрезвычайно трудоемкий процесс аннотирования данных ограничивает широкое применение глубокого обучения во многих приложениях. Для решения этой проблемы недавно было предложено несколько подходов, таких как обучение без учителя или обучение со слабым контролем. В настоящее время обучение с самоконтролем демонстрирует самые современные результаты и превосходит обучение с учителем для многих задач. Еще одной современной моделью нейронных сетей являются трансформаторные сети, которые могут обеспечить высокую производительность благодаря гибкости модели. Более того, качество аннотации напрямую влияет на качество работы сети. С этой точки зрения важно проанализировать, какие функции использует сеть в процессе обучения. Изучение механизма самовнимания позволяет выявить эти особенности и использовать их в процессе аннотирования. Настоящее исследование рассматривает проблему самоконтролируемого обучения трансформаторных сетей как многообещающий подход, позволяющий сделать шаг вперед в самоадаптации моделей нейронных сетей. В частности, мы изучаем кроссмодальную применимость самостоятельного обучения с использованием сети Transformer, предварительно обученной на цветных изображениях, для фильтрации данных в наборах данных тепловых изображений. Результаты оценки показывают, что сеть Transformer, основанная на механизме самообслуживания, идентифицирует одни и те же функции как в цветных, так и в наборах данных тепловых изображений. Ключевые слова самостоятельное обучение, нейронные сети, механизмы локального внимания 1. Введение Методы, основанные на данных, достигли большого прогресса в широком спектре приложений машинного зрения и анализа данных благодаря новым возможностям сбора, аннотирования и обработки огромных объемов данных, причем обучение с учителем дало самые впечатляющие результаты. К сожалению, чрезвычайно трудоемкий процесс аннотирования данных ограничивает широкое применение глубокого обучения во многих приложениях. Этот недостаток методов обучения с учителем сильно сдерживает реализацию глубокого обучения для многих приложений. Было предложено несколько подходов, решающие эту проблему, таких как обучение без учителя [1], обучение с полуконтролем [2], обучение со слабым учителем [3] и метаобучение [4]. В последнее время самообучение (SSL) привлекло значительное внимание в компьютерном зрении и позволило добиться значительных успехов в сокращении человеческого контроля. Действительно, выделяя репрезентативные характеристики из немаркированных данных, алгоритмы SSL уже превосходят контролируемое предварительное обучение во многих задачах [5]. Самоконтролируемое обучение нейронных сетей — это метод обучения, основанный на принципе использования знаний,

которые сеть уже знает об изображении. В отличие от традиционных подходов, где данные подаются в сеть и их вывод сравнивается с желаемым результатом. Сети, обученные в режиме Self-Supervised, преобразуют каждое изображение в embedding — вектор в некотором пространстве, несущий числовую информацию об изображении. Преимуществом этого метода обучения является возможность использовать неразмеченные данные и позволить нейронной сети самостоятельно выбирать наиболее значимые области изображения. В этих методах выдающиеся результаты показывает зрительный преобразователь (visual transformers (ViT)), основанный на механизме self-attention. Однако существенным недостатком классических реализаций визуальных преобразователей является низкая скорость получения карт признаков изображений из-за квадратичной сложности исполнения. Одним из решений этой проблемы является использование механизмов локального внимания. В этом методе для каждого пикселя мы получаем взвешенную карту признаков только относительно ближайших к нему пикселей. Благодаря этому можно получить взвешенную карту признаков за линейное время. В настоящее время методика совместного внедрения демонстрирует впечатляющие результаты в самостоятельном обучении. В качестве отправной точки в исследовании используется метод самоконтроля DINO [10]. Платформа DINO упрощает самоконтролируемое обучение, напрямую прогнозируя результаты работы сети учителей, построенной с помощью кодировщика импульса, с использованием стандартных потерь перекрестной энтропии. Основными вкладами исследования являются: (1) оригинальная структура DinoNAT для самостоятельного обучения, основанного на дистилляции знаний без маркировки; (2) оценка эффективности деятельности DinoNAT рамки; (3) демонстрация кросс-модальной производительности DinoNAT framework для тепловизионного зрения. 2. Сопутствующая работа 2.1. Самостоятельное обучение Большое количество методов основано на различном представлении векторов – вложений одного и того же дополненного изображения. Например, в [9] максимизируется совместная информация из разных слоев сети. Этот метод позволяет сохранить информацию о входном изображении. Также минимизируется расстояние между дополненными представлениями изображения, поскольку увеличение не вносит сильных искажений в семантику изображения. В статье [12] представлен подход, основанный на подаче двух разных дополнений в две сети. При этом первая сеть обновляет свои веса с помощью метода обратного распространения ошибки, а вторая — с помощью механизма экспоненциального скользящего среднего. Рассматриваемый метод основан на [10]. Представляем этот процесс как задачу дистилляции знаний. Сеть учащихся и сеть учителей участвуют в процессе обучения. Сеть учителя получает только глобальные исправления (большие фрагменты изображения), тогда как сеть учеников получает как глобальные, так и локальные исправления (небольшие фрагменты изображения). В течение одной эпохи, Веса сетей учащихся обновляются методом обратного распространения ошибки, в то время как веса сетей учителей замораживаются. В конце эпохи веса сетей учителей обновляются с использованием механизма экспоненциального скользящего среднего. 2.2. Трансформаторные сети Сети-трансформеры [6], изначально предложенные для задач обработки естественного языка (natural language processing (NLP)), основаны на самовнимании, что позволяет исключить рекуррентные и свёрточные операции. Трансформаторные сети в настоящее время демонстрируют современную производительность не только для решения задач обработки естественного языка, но и для приложений компьютерного зрения. Применение подхода Transformer к задаче анализа изображений привело к созданию Vision Transformers (ViT) [8]. Они демонстрируют конкурентоспособную производительность по сравнению со сверточными сетями, но имеют такие недостатки, как высокие потребности в вычислительных и обучающих данных, а не извлечение уникальных функций. Vision Transformer (ViT) [12] был предложен в качестве классификатора изображений, использующего толь-

ко Transformer Encoder, работающий со встроенным пространством фрагментов изображения, в основном для крупномасштабного обучения. Затем последовал ряд других методов, пытающихся повысить эффективность данных [13], в конечном итоге сделав такие модели, подобные Transformer, современными в классификации ImageNet-1K (без предварительного обучения на крупномасштабных наборах данных, таких как JFT -300M). Предыдущие работы, такие как DETR [4], исследовали гибриды CNN-Transformer для обнаружения объектов. ViT, с другой стороны, предложил модель, которая будет полагаться только на один непересекающийся сверточный слой (исправление и внедрение). ViT прошел предварительное обучение в основном на частном наборе данных JFT-300M, и было показано, что он превосходит современные CNN по многим критериям. Однако было также добавлено, что когда ViT предварительно обучается на наборах данных среднего масштаба, таких как ImageNet-1K и ImageNet-21K, он больше не достигает конкурентоспособных результатов. Модель преобразователя изображений с эффективным использованием данных (DeiT) продвинула ViT вперед с минимальными архитектурными изменениями, а также за счет использования передовых дополнений и методов обучения. Их усилия подчеркнули истинный потенциал классификатора изображений на основе Transformer в режимах данных среднего размера и вдохновили многих принять их методы обучения. 2.3. Внимание к себе (Self Attention) Масштабированное скалярное произведение внимания было определено Vaswani et al. [31] как операция над запросом и набором пар ключ-значение. Скалярное произведение запроса Q и ключа K вычисляется и масштабируется. Softmax применяется к выходным данным для нормализации весов внимания, а затем применяется к значению V. Это можно выразить следующим образом: Где d является вложенным измерением. Self-attention применяет внимание к скалярному произведению над линейными проекциями тех же входных данных, что и запрос, и пары ключ-значение. В Трансформерах обычно применяются многоголовые варианты внимания и самовнимания. Многоголовое внимание многократно применяет внимание скалярного произведения к различным вложениям, образуя, таким образом, головы внимания. Автономное самовнимание (Stand Alone Self Attention (SASA)) — один из первых паттернов самообслуживания со скользящим окном, призванный заменить свертки в существующих CNN. Он работает аналогично свертке с заполнением нулями и извлекает пары ключ-значение путем перемещения по карте объектов. Авторы сообщили о заметном улучшении точности, но заметили, что реализация имеет большую задержку, несмотря на более низкую теоретическую стоимость. SASA и его модификации не могли масштабироваться до более крупных окон и моделей из-за больших вычислительных затрат. Кроме того, не было решено уменьшение рецептивного поля в угловых случаях, вызванное набивкой. Window and Shifted Window (Swin) Attention было предложено Liu et al. как механизмы самоконтроля на основе нескользящих окон, которые разделяют карты объектов и применяют самоконтроле к каждому разделу отдельно. Эта операция по теоретической сложности аналогична SASA, но ее можно легко распараллелить посредством пакетного умножения матриц. Смещенный вариант следует за обычным и, как следует из названия, сдвигает разделение, чтобы обеспечить внеоконные взаимодействия, необходимые для роста восприимчивого поля. Предложенная ими модель Swin Transformer — одна из первых преобразователей иерархического зрения. Она создает пирамидальные карты объектов, уменьшая пространственную размерность и одновременно увеличивая глубину. Предлагаемое внимание соседей (Neighbourhood Attention (NA)) [7] локализует SA на ближайших соседей каждого пикселя, что не обязательно является фиксированным окном вокруг пикселя. Это изменение в разрешении позволяет всем пикселям сохранять одинаковую концентрацию внимания, которая в противном случае была бы уменьшена для угловых пикселей в альтернативах с нулевым дополнением (SASA). NA также приближается к SA по мере роста размера

ее окрестности и эквивалентна SA в максимальной окрестности. Кроме того, NA имеет дополнительное преимущество, заключающееся в поддержании трансляционной эквивариантности, в отличие от блокированного и оконного самообслуживания. 2.4. Местные механизмы внимания Методы механизмов локального внимания основаны на концепции получения карты признаков внутри определенного окна, а не расчета ее для всего изображения. В статье [13] был проведен эксперимент по замене операций свертки механизмами локального внимания. В результате это нововведение увеличило скорость вывода модели и метрик на примере ImageNet. В [7] была предложена новая архитектура нейронной сети, основанная на механизме локального внимания, названном Neighborhood Attention Transformer (NAT). Авторам удалось получить конкурентоспособную архитектуру, показавшую более высокие показатели скорости и точности классификации по сравнению с моделью Swin. 3. Метод В этом разделе описан процесс обучения выбранной нейронной сети в режиме самообучения и ее контролируемое дополнительное обучение. Метод основан на [10]. Этот процесс представлен как задача дистилляции знаний 3.1. Обучение сети Сначала создаются две копии одной и той же модели для получения векторов внедрения. В качестве моделей ученика и учителя была взята архитектура Neighborhood Attention Transformer, схематическая схема которой представлена на рис. 1. Рисунок 1. Внешняя схема архитектуры NAT Далее на каждой итерации обучения сети из образа получаются глобальные и локальные патчи. Глобальные патчи — это части изображения, занимающие более 50% изображения (224x224 пикселей). Локальные патчи — части изображения меньшего размера (96x96 пикселей). Они подвергаются аугментации и подаются на вход сетей. Вход сети учителя — это глобальный патч, вход сети ученика — глобальный или локальный патч, но не такой же, как вход сети учителя. Выходными данными сетей являются вложения тех же размерностей. На основе полученных векторов получаем значение функции ошибок, представленное формулой: Где θ - весы, s – сеть-студент, x, x' – разные патчи подаются на вход в сети, H(a, b) – функция ошибки - перекрестная энтропия, Pa(b) –вектор вывода сети a. На основе полученного значения функции ошибок веса сети учеников обновляются, когда веса сети учителей замораживаются. В конце каждой эпохи веса сети учителей обновляются на основе весов сети учеников с использованием скользящего среднего: Wt = aWt + (1 − a)Ws, где Wt – весы модель-учителя, Ws – модельно-студентские веса, a – некоторое число от 0 до 1. Схема обучения DINO представлена на рис. 2. Ее можно представить как Алгоритм 1 Рисунок 2. Иллюстрация работы DINO. К входному изображению применяются различные дополнения учащихся и учителей, после чего входное изображение передается через нейронные сети для создания векторов внедрения. Вектор вывода учителя подвергается операциям заточки и центрирования, чтобы избежать проблем с коллапсом. Векторы учителя и ученика затем передаются через Softmax, и функция потерь рассчитывается путем обновления весов учащихся. Веса учителей обновляются каждую эпоху с использованием метода скользящего среднего. 3.2. Переобучение классификатора В качестве классификатора использовался метод ближайшего соседа, а также линейный классификатор. На каждой итерации предварительного обучения дополненное изображение подавалось на вход предварительно обученной сети с замороженными весами, на вход подавался полученный вектор внедрения. классификатора. Было рассчитано значение ошибки классификации и на его основе обновлены веса классификатора. 4. ВЫВОД Данная работа посвящена исследованию самообучения преобразователей с использованием механизмов локального внимания. Самостоятельное обучение — это модельный режим обучения, при котором разметка формируется на основе внутренней структуры самих объектов или на основе базовых знаний об объектах. Преимущество этого метода в том, что нет необходимости в дополнительной разметке. В качестве обучающей сети используется Neighborhood Attention Transformer — преобразо-

ватель, основанный на использовании механизма локального внимания. В рамках данной работы нейронная сеть Neighborhood Attention Transformer обучалась на образце ImageNet с использованием метода самоуправляемого обучения нейронных сетей DINO. Результаты, полученные в результате обучения, сравнивались с результатами, представленными в оригинальной статье метода DINO, и делались выводы о качестве сети (рис. 3). Образцы из набора данных ThermalGan Car Humans Buildings Рисунок 3. Примеры изображений и полученные карты внимания.

Источники и литература

1) Счастье Угочи Дике и др. «Обучение без учителя на основе искусственной нейронной сети: обзор». Международная конференция IEEE по киборгам и бионическим системам (CBS), 2018 г. IEEE. 2018, стр. 322–327.
2) Йеспер Э. Ван Энгелен и Хольгер Х. Хоос. «Опрос по полуконтролируемому обучению». Machine Learning 109.2 (2020), стр. 373–440.
3) Чжи-Хуа Чжоу. «Краткое введение в обучение со слабым учителем». Национальный научный обзор 5.1 (2018), стр. 44–53.
4) Хоакин Ваншорен. «Метаобучение». Автоматизированное машинное обучение. Спрингер, Чам, 2019 г., стр. 35–61.
5) Прия Гоял, Матильда Карон, Бенджамин Лефодо, Мин Сюй, Пэнчао Ван, Вивек Пай, Маннат Сингх, Виталий Липчинский, Ишан Мисра, Арманд Жулен и Петр Бояновский, «Самостоятельная предварительная тренировка визуальных функций в дикой природе», arXiv:2103.01988 2021, https://doi.org/10.48550/arXiv.2103.01988.
6) Ашиш Васвани, Ноам Шазир, Ники Пармар, Якоб Ушкорейт, Лайон Джонс, Эйдан Н Гомес, Лукаш Кайзер и Илья Полосухин. «Внимание – это все, что вам нужно». Достижения в области нейронных систем обработки информации, 30 октября 2017 г.
7) Али Хассани, Стивен Уолтон, Цзячен Ли, Шен Ли, Хамфри Ши; «Трансформатор внимания соседей». Материалы конференции IEEE/CVF по компьютерному зрению и распознаванию образов (CVPR), 2023 г., стр. 6185-6194.
8) Алексей Досовицкий, Лукас Байер, Александр Колесников, Дирк Вайсенборн, Сяохуа Чжай, Томас Унтертинер, Мостафа Дегани, Матиас Миндерер, Георг Хейгольд, Сильвен Гелли и др. «Изображение стоит 16x16 слов: Трансформаторы для распознавания изображений в масштабе». препринт arXiv:2010.11929, 2020
9) Бахман, Филип и Хьельм, Р. Девон и Бухвальтер, Уильям, «Изучение представлений путем максимизации взаимной информации между представлениями».
10) Карон, Матильда и Туврон, Хьюго и Мисра, Ишан и Жегу, Эрве и Майрал, Жюльен и Бояновский, Петр и Жулен, Арманд, «Новые свойства самоуправляемых преобразователей зрения», arXiv:2104.14294, 2021, https: //doi.org/10.48550/arXiv.2104.14294
11) Максим Окуаб и др., DINOv2: «Изучение надежных визуальных функций без надзора», arXiv:2304.07193 ,2023, https://doi.org/10.48550/arXiv.2304.07193
12) Гриль, Жан-Бастьен и Струб, Флориан и Альче, Флоран и Таллек, Корантен и Ришмон, Пьер и Бучацкая, Елена и Дёрш, Карл и Авила Пирес, Бернардо и Го, Жаохан и Гешлаги Азар, Мохаммед и Пиот, Билал и Кавукчуоглу, Корай и Мунос, Реми и Валко, Михал, «Достижения в области нейронных систем обработки информации».

13) Праджит Рамачандран, Ники Пармар, Ашиш Васвани, Ирван Белло, Ансельм Левская и Джонатон Шленс, «Автономное внимание к себе в моделях видения».

Иллюстрации

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \quad (1)$$

Рис. : Масштабированное скалярное произведение внимания было определено Vaswani et al. [31] как операция над запросом и набором пар ключ-значение. Скалярное произведение запроса Q и ключа K вычисляется и масштабируется. Softmax применяется к выходным данным для нормализации весов внимания, а затем применяется к значению V. Это можно выразить следующим образом:

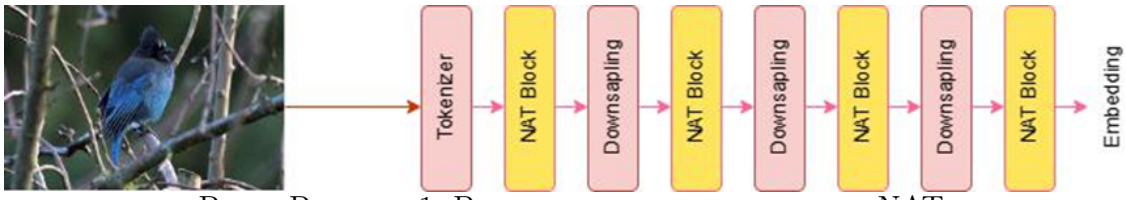

Рис. : Рисунок 1. Внешняя схема архитектуры NAT

$$\min_{\theta_S} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{x' \in V, x' \neq x} H(P_t(x), P_s(x')), \quad (2)$$

Рис. : На основе полученных векторов получаем значение функции ошибок, представленное формулой:

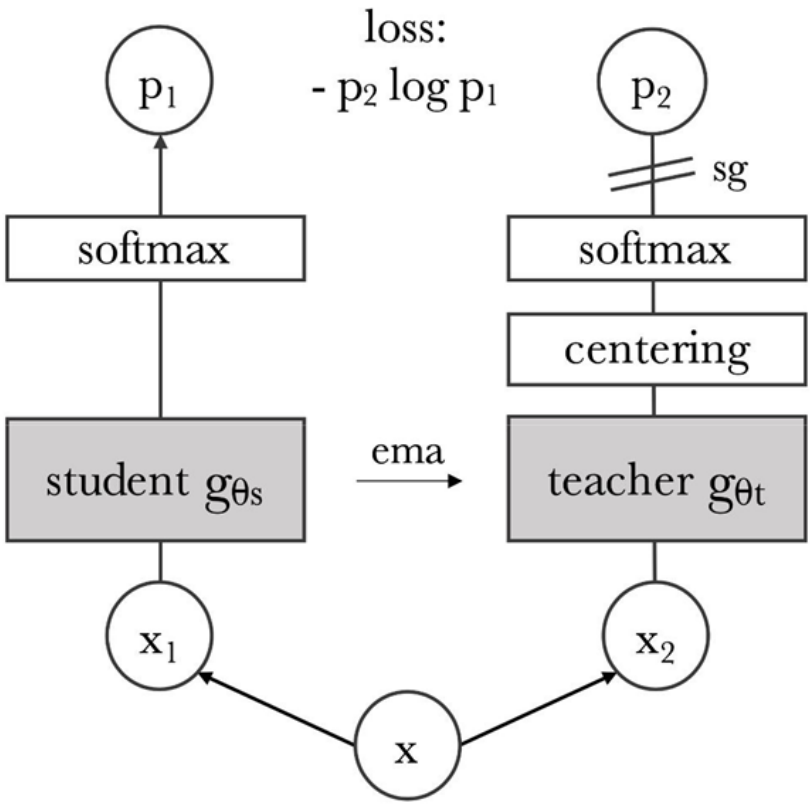

Рис. : Рисунок 2. Иллюстрация работы DINO. К входному изображению применяются различные дополнения учащихся и учителей, после чего входное изображение передается через нейронные сети для создания векторов внедрения. Вектор вывода учителя подвергается операциям заточки и центрирования, чтобы избежать проблем с коллапсом. Векторы учителя и ученика затем передаются через Softmax, и функция потерь рассчитывается путем обновления весов учащихся. Веса учителей обновляются каждую эпоху с использованием метода скользящего среднего.

Algorithm 1: DINO training

```
/* This is comment                                                    */
```

Input:

Teacher network model G_t with θ_t parameters
Student network model G_s with θ_s parameters
tp_s, tp_t: student and teacher temperatures
C – center (K)
q, m – network and center momentum rates

Output:

θ_s parameters, that match probability $P_t(x)$

```
1  Training procedure ;
2  Procedure Training():
       /* load a minibatch x with n samples                           */
3      for x ∈ Loader do
4          x_1 = augment(x)
5          x_2 = augment(x)
6          s_1 = G_s(x_1), s_2 = G_s(x_2) ;          /* G_s output */
7          t_1 = G_t(x_1), t_2 = G_t(x_2) ;          /* G_t output */
8          loss = H(t_1, s_2)/2 + H(t_2, s_1)/2
9          loss.backward()
10         update(G_s); ;                                  /* SGD */
11         G_t.params = q · G_t.params + (1 − q) · G_s.params
12         C = m · C + (1 − m) · cat([t1, t2]).mean(dim = 0)
13     return θ_s;
14 Loss Function ;
15 Function Loss(t,s):
16     t = t.detach() ;                          /* stop gradient */
17     s = softmax(s / tps, dim=1)
18     t = softmax((t - C) / tpt, dim=1) ;     /* center + sharpen */
19     return (t ∗ log(s)).sum(dim = 1).mean();
```

Рис. : Данная работа посвящена исследованию самообучения преобразователей с использованием механизмов локального внимания. Самостоятельное обучение — это модельный режим обучения, при котором разметка формируется на основе внутренней структуры самих объектов или на основе базовых знаний об объектах. Преимущество этого метода в том, что нет необходимости в дополнительной разметке.

Рис. : Рисунок 3. Примеры изображений и полученные карты внимания.