Содержание

Аннотация

В работе исследуется задача байесовского выбора архитектуры нейросетевой модели. Предлагается метод, оценивающий апостериорное совместное распределение структуры и параметров модели с помощью вариационной нижней оценки обоснованности. Вводятся априорные предположения о распределении параметров и структуры модели. Показано, что байесовский подход регуляризует норму гессиана оптимизируемой функции ошибки по структуре модели. Для оценки качества и устойчивости предложенного метода проводится вычислительный эксперимент на выборке Fashion-MNIST. Показано, что предлагаемый метод является более устойчивым в сравнении с базовыми градиентными методами выбора структуры.

1 Введение

Подходы, основанные на нейронных сетях, показывают высокое качество в решении многих задач таких, как детектирование объетов на изображении, построение языковых моделей и т.д [18, 14]. Однако выбор архитектуры, способной решить задачу с высоким качеством, требует больших экспертизы и временных затрат. Современный подход для преодоления данного ограничения, называемый поиском архитектуры нейросетевой модели (ПАНМ), заключается в автоматической генерации архитектуры модели для заданных задачи и набора данных. Архитектуры, полученные с помощью методов ПАНМ, продемонстрировали возможность решать задачи с большей точностью, чем созданные вручную [5].

Несмотря на то что такие модели обладают высокой обобщающей способностью, у них наблюдается недостаточная устойчивость [6]. При добавлении внешнего воздействия в параметры метода, точность модели с архитектурой, полученной с помощью ПАНМ, может резко снижаться. Одним из примеров проявления низкой устойчивости является уязвимость к состязательным атакам [6] – методам воздействия на модель или её входные параметры с целью снижения обобщающей способности. Такие воздействия потенциально могут быть очень опасны. В [19] приводится пример, описывающий применение атаки на знак градиента [8]. Такая атака подразумевает добавление во входное изображение специально подобранного шума, из-за которого объект на изображении начинает классифицироваться неверно. Именно поэтому становится важным делать модели более устойчивыми к подобным атакам и, в целом, ко всем внешним воздействиям.

Цели и задачи исследования. Основной целью исследования является повышение устойчивости градиентного метода поиска архитектуры нейросетевой модели с применением стохастических методов порождения параметров и структуры. Для реализации этой цели поставлены следующие задачи:

- изучить существующие методы решения задачи ПАНМ;
- изучить возможные методы внешнего воздействия на модели глубокого обучения;
- провести вариационный вывод оценки апостериорного распределения параметров и структуры;

- предложить теоретическую интерпретацию и обоснование предлагаемого метода;
- реализовать метод в виде программного кода на языке Python;
- провести анализ релевантных наборов данных;
- провести вычислительный эксперимент;
- провести сравнительный анализ точности и устойчивости относительно состязательных атак предлагаемого метода с другими подходами.

Научная новизна. Предложен метод построения более устойчивой архитектуры модели глубокого обучения, основанный на градиентном подходе поиска архитектуры. Показана связь между устойчивостью архитектуры и нормой гессиана по структуре. Показано, что применяемый подход регуляризует норму гессиана по структуре.

Методы исследования. Использованы методы глубокого обучения, стохастические методы порождения, метод вариационной нижней оценки обоснованности и градиентные методы оптимизации.

Практическая ценность. Предложенный метод предназначен для построения прикладных моделей, основанных на свёрточных нейронных сетях, и их применения в задачах классификации изображений.

Работа состоит из пяти разделов, заключения и списка литературы. Содержание изложено на второй странице. Список литературы включает 24 наименований.

Во **Введении** обосновываются цели и задачи исследования, его научная и практическая значимость.

В **Разделе 2** вводятся основные определения и понятия, формулируется вероятностная модель и ставится формальная постановка задачи.

В **Разделе 3** проводится анализ существующих методов решения задачи поиска архитектуры нейросетевой модели, а также повышения устойчивости таких методов относительно внешних воздействий.

В **Разделе 4** выводится функция ошибки, описывается предлагаемый метод, предлагается его теоретическая интерпретация, а также выбираются априорные предположения о распределении архитектуры.

В **Разделе 5** описываются используемые данные, процесс проведения вычислительного эксперимента, проводится сравнительный анализ точности и устойчивости с другими подходами.

В **Заключении** фиксируются основные результаты работы и указываются направления дальнейших исследований.

2 Постановка задачи

Повествование в данном разделе построено следующим образом. Сначала будут введены основные определения. Затем будет описана вероятностная модель. После будет сформулирована формальная постановка задачи.

2.1 Основные понятия и определения

Определение 1. *Моделью* называется дифференцируемая по параметрам функция

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = y : \mathbb{W} \times \mathbb{X} \to \mathbb{Y},$$

где $\mathbf{w} \in \mathbb{W}$ - параметры модели, $\mathbf{x} \in \mathbb{X}$ - признаковое описание входного объекта, $y \in \mathbb{Y}$ - метка входного объекта.

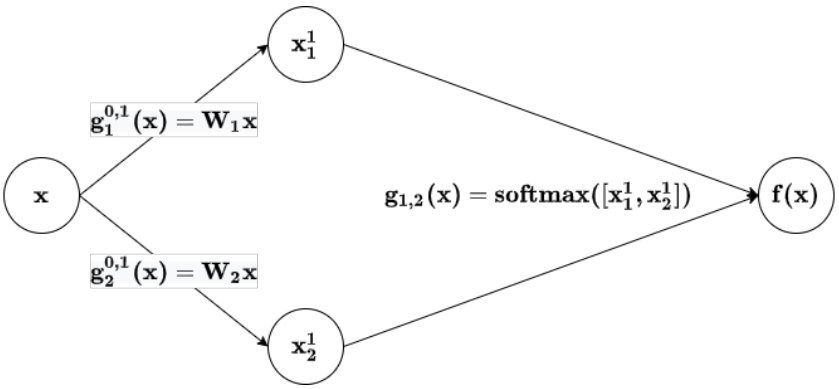


Рис. 1: Пример представления модели $\mathbf{f} = \mathbf{softmax}(\gamma_0^{0,1} \cdot \mathbf{W_1}(\mathbf{x}) + \gamma_1^{0,1} \cdot \mathbf{W_2}(\mathbf{x}))$ в виде графа.

Модель представляется в виде направленного ациклического графа $G = (V, E)$ (рис. 1). На каждом ребре $(j, k) \in E$ задаётся вектор дифференцируемых функций $\mathbf{g}^{j,k}$, называемых операциями. Каждая вершина $v_k \in V$ является результатом применения операций над выходами с предшествующих вершин, т.е. справедливо

$$v_k = \sum_{j<k} \mathbf{g}^{j,k}(v_j).$$

Определение 2. Пусть на ребре (j, k) задан вектор операций $\mathbf{g}^{j,k}$, $|\mathbf{g}^{j,k}| = N^{j,k}$. *Структурными параметрами* назовём вектор $\boldsymbol{\gamma}^{j,k} = [0, 1]^{N^{j,k}}$.

Определение 3. *Структурой* модели называется конкатенация её структурных параметров $\boldsymbol{\Gamma} = \{\boldsymbol{\gamma}^{j,k} | (j, k) \in E\}$.

Определение 4. *Архитектурой* модели называется совокупность её параметров и структуры.

В настоящей работе $\mathbf{f}(\mathbf{w}, \mathbf{x})$ рассматривается как вероятностная модель $p(\mathbf{w}, \mathbf{\Gamma})$ архитектуры. Параметрам и структуре сопоставляется распределение, соответствующее априорным представлениям о статистических свойствах заданных объектов.

Определение 5. *Гиперапараметрами* $\mathbf{h} \in \mathbb{H}$ модели назовём параметры распределения $p(\mathbf{w}, \mathbf{\Gamma} \,|\, \mathbf{h})$.

В качестве примера можно предположить, что совместное распределение параметров и структуры модели задано через нормальное распределение $p(\mathbf{w}, \mathbf{\Gamma}) \propto \mathcal{N}(\mathbf{m}, \mathbf{A^{-1}})$. Тогда гиперепараметрами модели будет являться набор $\mathbf{h} = \{\mathbf{m}, \mathbf{A^{-1}}\}$.

2.2 Формальная постановка задачи

Задан набор данных $\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ где каждому входному объекту $\mathbf{x}_i \in \mathbb{X}$ соответствует целевая переменная $y_i \in \mathbb{Y}$. Элементы $(\mathbf{x}_i, y_i)$ являются случайными величинами, взятыми из совместного распределения $\mathbf{p}(\mathbf{x}, \mathbf{y})$. Задаются априорные распределения на параметры $\mathbf{w} \sim p(\mathbf{w} \,|\, \mathbf{h})$ и структуру $\mathbf{\Gamma} \sim p(\mathbf{\Gamma} \,|\, \mathbf{h})$ модели.

Целью работы является нахождение совместного апостериорного распределения параметров и структуры модели. Предполагается, что параметры модели не зависят от её структуры. Вероятностная модель задается следующим образом:

$$p(\mathbf{w}, \mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) = p(\mathbf{w} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) \cdot p(\mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}).$$

В качестве оптимальных параметров $\mathbf{w}^*, \mathbf{\Gamma}^*$ предлагается использовать те, которые доставляют максимум апостериорной вероятности.

Для оценки апостериорного распределения введём вариационное совместное распределение параметров и структуры $q(\mathbf{w}, \mathbf{\Gamma} \,|\, \boldsymbol{\theta})$ с параметризацией $\boldsymbol{\theta}$. Предполагается независимость параметров от структуры, то есть $q(\mathbf{w}, \mathbf{\Gamma} \,|\, \boldsymbol{\theta}) = q_{\mathbf{w}}(\mathbf{w} \,|\, \boldsymbol{\theta}_{\boldsymbol{w}}) \cdot q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\theta}_{\mathbf{\Gamma}})$.

Предлагается минимизировать расстояние Кульбака-Лейблера между оценочным и истинным апостериорным распределением:

$$\mathcal{L}(\mathbf{w}, \mathbf{\Gamma}) = \mathcal{D}_{KL}\big(q(\mathbf{w}, \mathbf{\Gamma} \,|\, \boldsymbol{\theta}) \;||\; p(\mathbf{w}, \mathbf{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h})\big).$$

Также предполагается, что набор данных разделён на обучающую и валидационную выборку, то есть представим в виде $\mathfrak{D} = \mathfrak{D}_{train} \sqcup \mathfrak{D}_{val}$. Через $\mathcal{L}_{train}$, $\mathcal{L}_{val}$

обозначается функция ошибки на обучающей и валидационной выборках, соответственно. Вывод точного вида функции ошибки будет представлен в разделе 4.1.

Таким образом, ставится двухуровневая оптимизационная задача:

$$\begin{aligned} &\min_{\mathbf{\Gamma}} \ \mathcal{L}_{val}(\hat{\mathbf{w}}, \mathbf{\Gamma}) \\ &s.t. \ \ \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} \mathcal{L}_{train}(\mathbf{w}, \mathbf{\Gamma}). \end{aligned} \tag{1}$$

На первом уровне вычисляются параметры модели, доставляющие минимум функции ошибки на обучающей выборке. На втором уровне, используя оптимальные параметры, находится оптимальная структура модели, на которой функция ошибки на валидационной выборке достигает минимума.

3 Обзор существующих методов

В данном разделе будут подробно рассмотрены методы решения исследуемой задачи. Сначала будут описаны методы, основанные на обучении с подкреплением. Затем речь пойдет о подходах, использующих градиентные методы обучения архитектуры. После будет рассмотрен байесовский подход, предполагающий введение априорных предположений о распределении архитектуры. Наконец, будет описан обзор методов, направленных на повышение устойчивости ПАНМ.

3.1 Методы, основанные на обучении с подкреплением

Рассмотрим подробнее методы, в основе которых лежит подход обучения с подкреплением. В работе [15] роль агента выполняет контроллер LSTM [11], формирующий архитектуру путём последовательной генерации операций на каждом слое (рис. 2).

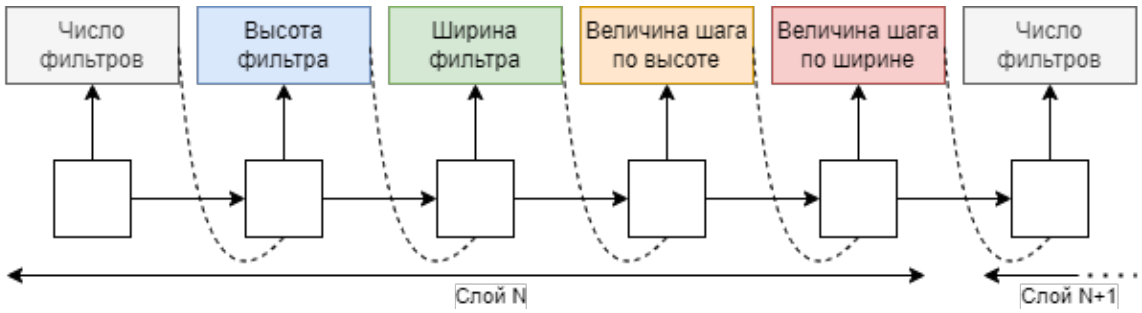


Рис. 2: Процесс поиска архитектуры методом [15]. Реккурентный контроллер последовательно генерирует параметры свёрточного слоя.

Наградой выступает точность обученной на валидационной выборке модели. Таким образом, с каждой итерацией контроллер генерирует все более оптимальные архитектуры (рис. 3).

Данный подход показал очень высокое качество в решении задач детектирования объектов на изображении и построения языковых моделей. Но серьезным недостатком такого метода является высокая требовательность к вычислительным ресурсам и, как следствие, крайне длительный процесс поиска.

В работе [4] была предложена модификация, позволяющая заметно сократить время поиска оптимальной структуры нейросетевой модели. Семейство всех архитектур представляется в виде параметризованного ациклического направленного графа, на ребрах которого определяются возможные операции над выходами с предыдущих

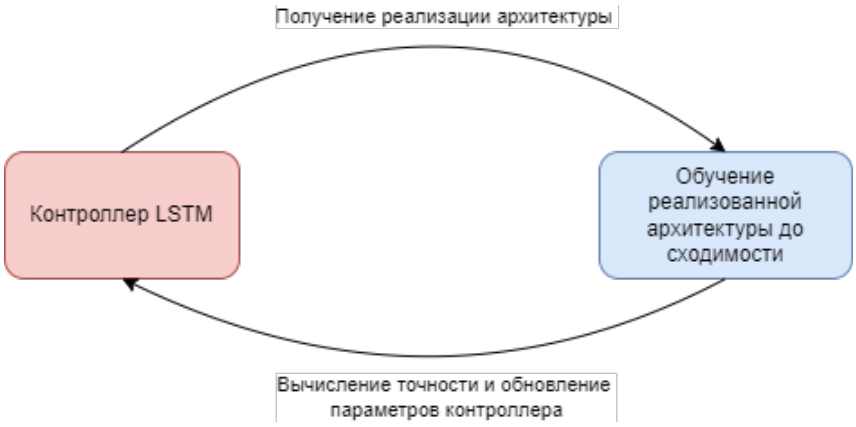


Рис. 3: Схематичная визуализация метода, описанного в [15].

слоёв.

Конкретная архитектура получается путём выбора конкретной операции на ребре графа. В процессе поиска параметры обученных операций не сбрасывались до изначальных, а сохранялись и использовались в последствии другими архитектурами. Такой процесс использования общего пространства параметров позволил резко сократить время поиска, при этом сохранив высокую обобщающую способность найденных структур.

3.2 Градиентные методы

В работе [16] был предложен подход DARTS, позволяющий решать задачу ПАНМ градиентными методами. Семейство всех архитектур так же представляется в виде ациклического направленного графа, как в вышеописанных работах. Производится процедура релаксации дискретного пространства поиска структуры в непрерывное. Вводятся обучаемые параметры структуры, являющиемся весами соответствующих им операций и смешанная операция (рис. 4), являющаяся взешенной гладкой функцией от операций на текущем ребре графа.

Такой подход позволяет брать производную по структурным параметрам, что позволяет обучать поиск градиентными методами. После обучения происходит проекция структуры со смешанными операциями на структуру с одной операцией на каждом ребре путём выбора операции, которой соответствует структурный параметр с наибольшим значением.

В методе, описанном в [20], ставится оптимизационная задача обучения параметров распределения структуры с помощью градиентных методов. Структурные пара-

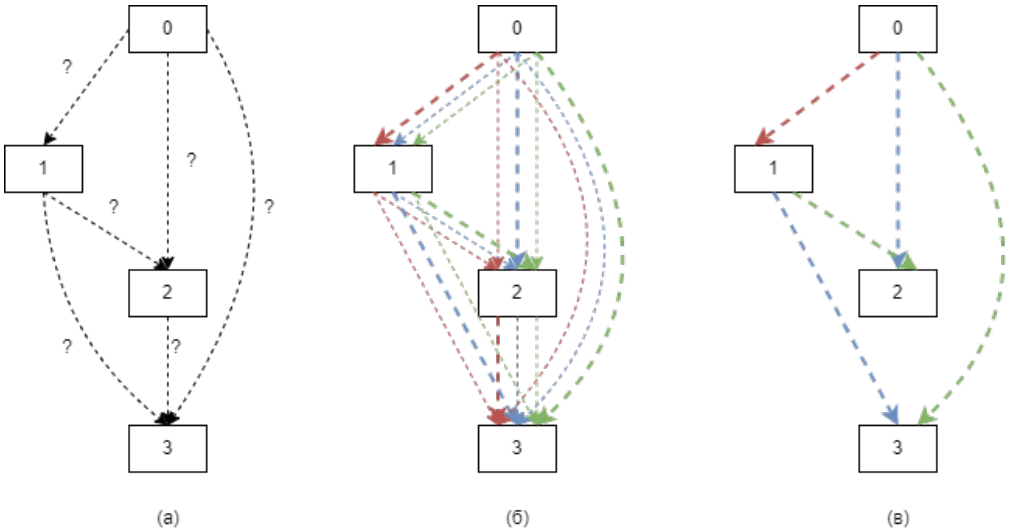


Рис. 4: Визуализация метода DARTS [16]. На этапе (а) происходит инициализация модели и пространства поиска. На этапе (б) показан процесс обучения структурных параметров. Этап (в) демонстрирует окончательную архитектуру.

метры операций задаются с помощью распределения Gumbel-Softmax [13], которое также позволяет релаксировать дискретное пространство поиска на непрерывное. С помощью такой репараметризации становится возможным рассчитывать градиенты по структурным параметрам и проводить обучение методами стохастического градиентного спуска.

3.3 Байесовский подход к ПАНМ

В [1] отмечен недостаток градиентного ПАНМ, который заключается в предпочтении нулевых операций и операций пропуска в силу нахождения локального оптимума. Из-за которого качество генерируемых архитектур становится ниже. Для борьбы с таким эффектом был предложен байесовский подход к градиентному ПАНМ. Параметры и структура модели являются независимыми случайными величинами, на которые задаются априорные нормальные распределения. Для оценки неизвестного апостериорного распределения параметров и структуры минимизируется матожидание правдоподобия. Также приводятся дополнительные регуляризационные члены, стабилизирующие процесс поиска. Примерами являются l_1 и l_2 регуляризации структурных параметров модели.

Метод, описанный в [3], также формулируется как задача вероятностного распределения. Структурные параметры смешанной операции представляются как случайные величины, на которые задается априорное распределение Дирихле. Выбор

такого распределения позволяет. Для контроля дисперсии распределения Дирихле используется дополнительная регуляризация параметра концентрации.

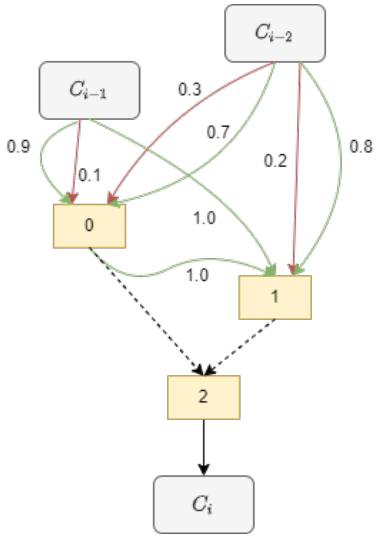


Рис. 5: Общая схема задания смешанной операции в байесовском подходе к ПАНМ. Структурные параметры приобретают смысл вероятности соответствующей операции. Около каждого ребра с операцией указаны значения обученных структурных параметров. Зелёным отмечены те рёбра, которые входят в окончательную архитектуру.

В работе [7] апостериорное распределения параметров и структуры оценивается параметризованным вариационным распределением. Для минимизации расстояния Кульбака-Лейблера между ними предлагается оптимизировать вариационную нижнюю оценку обоснованности. Оценочные вариационные распределения задаются как нормальные распределения с диагональной матрицей ковариации. Дополнительно, для уменьшения неопределённости в распределении структурных параметрах, к функции потерь добавляется регуляризация в виде совместной по структурным параметрам энтропии.

В [23] описан другой вероятностный подход к задаче ПАНМ, а именно байесовская оптимизация. Предлагается моделировать функцию сбора данных через отдельную нейронную сеть. На каждой итерации очередная сгенерированная архитектура кодируется спцеиальным образом, затем обучается до сходимости. Полученные пары *архитектура-точность* с предыдущих итераций используются для тренировки полносвязной нейронной сети для прогнозирования точности сгенерированной архитектуры. После обучения выбирается та структура, которая доставляет максимальную спрогнозированную точность.

Также в работах [9], [17] поиск оптимальной архитектуры моделируется с помо-

щью скрытых марковских цепей и гауссовских случайных процессов.

3.4 Устойчивость методов ПАНМ

Актуальность проблемы повышения устойчивости относительно состязательных атак подтверждается в работе [24]. Суть описанного подхода заключается в добавлении в пространство поиска метода ENAS [4] дополнительных операций, использование которых ведёт к снижению обобщающей способности архитектуры. В процессе обучения увеличивается частота генерации архитектур с низкой точностью, что больше информации методу о пространстве поиска.

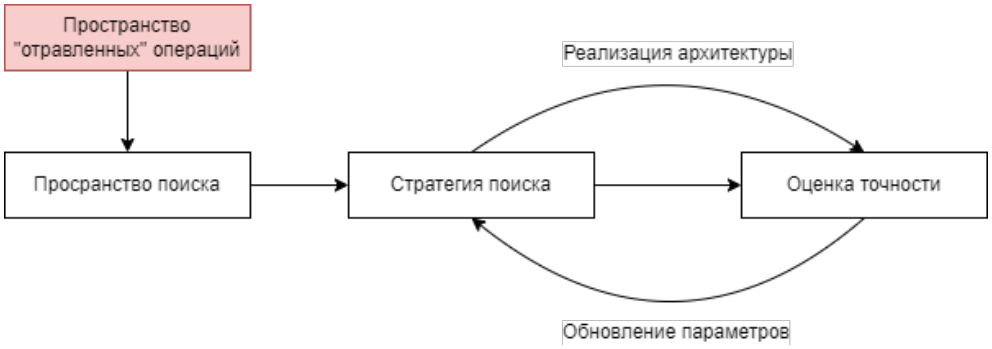


Рис. 6: Визуализация метода [4].

Схематичное изображение метода представлено на рис. 4. Отличительной чертой такого подхода является то, что злоумышленнику не требуется никаких априорных знаний о решаемой задаче или обучающей выборке.

В [12] для повышения устойчивости относительно состязательных атак предлагаются две дифференцируемые меры устойчивости. Первая основана на сертификате устойчивости – пороге мощности состязательной атаки, при котором зашумлённый объект все еще будет классифицироваться верно. Исходя из верхних и нижних оценок сертификата устойчивости для каждого блока модели оценивается нижняя оценка сертификата для всей модели, которая затем оптимизируется в процессе обучения. Другой метрикой является добавление регуляризации якобиана по входным объектам. Такой подход является менее затратным по вычислительным ресурсам, чем использование сертификата.

В работе [22] проводится анализ устойчивости градиентного метода поиска DARTS [16]. Оказывается, имеет место резкая потеря качества архитектуры при проецировании смешанных операций на симплекс, поэтому градиентные методы не

обладают достаточной устойчивостью. Эмпирическим путём была установлена связь между потерей качества и ростом спектра гессиана по структуре. Для контроля его значения была предложена регуляризация максимального собственного значения. При достижении собственного значения заданного порога процесс обучения прерывается.

Работа, описанная в [2], также адресует проблеме потери качества метода DARTS. Для повышения точности и устойчивости предлагается регуляризация, основанная на внесении шума в структуру модели. Подобное зашумление позволяет регулировать норму гессиана по структуре модели.

В настоящей работе предлагается более обоснованный и обобщенный подход с помощью байесовского вывода. Метод, описанный в [2], является частным случаем метода, описываемого в текущем исследовании.

4 Описание метода

В данном разделе повествование будет вестись следующим образом. Сначала будет проведён вывод функции ошибки. Затем будет приведена теоретическая интерпретация предлагаемого метода. После этого будут сформулированы априорные предположения о распределении параметров и структуры архитектуры модели, а также будет предоставлено алгоритмическое описание метода.

4.1 Функция потерь

По теореме Байеса имеем

$$p(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) = \frac{p(\mathbf{y} \,|\, \mathbf{w}, \boldsymbol{\Gamma}, \mathbf{X}) \cdot p(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{h})}{p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{h})}.$$

В силу большого числа скрытых параметров обоснованность модели аналитически невычислима. Предлагается использовать вариационную нижнюю оценку обоснованности [10]. Для этого перепишем логарифм обоснованности следующим образом:

$$\log p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{h}) = \iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{h}) d\boldsymbol{\Gamma} d\mathbf{w}. \tag{4.1}$$

Преобразуем полученный интеграл:

$$\iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{h}) d\boldsymbol{\Gamma} d\mathbf{w} = \iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{h})}{p(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h})} d\boldsymbol{\Gamma} d\mathbf{w} =$$

$$= \iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma}) q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta})}{p(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) q(\mathbf{w}, \boldsymbol{\Gamma} | \boldsymbol{\theta})} d\boldsymbol{\Gamma} d\mathbf{w} =$$

$$= \iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{h})}{q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta})} d\boldsymbol{\Gamma} d\mathbf{w} + \iint_{\mathbf{w},\boldsymbol{\Gamma}} q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \log \frac{q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta})}{p(\mathbf{w}, \boldsymbol{\Gamma} | \mathbf{X}, \mathbf{y}, \mathbf{h})} d\boldsymbol{\Gamma} d\mathbf{w} =$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma} \,|\, \boldsymbol{\theta})} \Big[\log \frac{p(\mathbf{y}, \mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{h})}{q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta})} \Big] + \mathcal{D}_{KL} \big(q(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \boldsymbol{\theta}) \;||\; p(\mathbf{w}, \boldsymbol{\Gamma} \,|\, \mathbf{X}, \mathbf{y}, \mathbf{h}) \big).$$

Во втором слагаемом полученного выражения присутствует апостериорное распределение параметров и структуры, которое требуется оценить. Перенесём его в левую часть уравнения (4.1):

$$\log p(\mathbf{y}|\mathbf{X},\mathbf{h}) - \mathcal{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})\;||\;p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{y},\mathbf{h})\big) = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log\frac{p(\mathbf{y},\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})}{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big] \tag{4.2}$$

Заметим, что минимизация дивергенции Кульбака-Лейблера между вариационным распределением и истинным апостериорным распределением эквивалентна максимизации матожидания, расположенного в правой части уравнения (4.2). Распишем его более подробно:

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log\frac{p(\mathbf{y},\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})}{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big] = \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log p(\mathbf{y},\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h}) - \log q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})\Big] =$$

$$= \mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log p(\mathbf{y}\,|\,\mathbf{w},\boldsymbol{\Gamma},\mathbf{X}) + \log p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h}) - \log q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})\Big] =$$

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log p(\mathbf{y}\,|\,\mathbf{w},\boldsymbol{\Gamma},\mathbf{X}) - \log\frac{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}{p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})}\Big] =$$

$$\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log p(\mathbf{y}\,|\,\mathbf{w},\boldsymbol{\Gamma},\mathbf{X})\Big] - \mathcal{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})\;||\;p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})\big).$$

В силу независимости совместные распределения $q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})$ и $p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})$ декомпозируются на маргинальные распределения по параметрам и структуре модели. Поэтому второе слагаемое можно переписать в виде

$$\mathcal{D}_{KL}\big(q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})\,||\,p(\mathbf{w},\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})\big) = \mathcal{D}_{KL}\big(q_{\mathbf{w}}(\mathbf{w}\,|\,\boldsymbol{\theta}_{\mathbf{w}})\cdot q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta}_{\boldsymbol{\Gamma}})\,||\,p(\mathbf{w}\,|\,\mathbf{X},\mathbf{h})\cdot p(\boldsymbol{\Gamma}\,|\,\mathbf{X},\mathbf{h})\big) =$$

$$= \mathcal{D}_{KL}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta}_{\boldsymbol{\Gamma}})\;||\;p(\boldsymbol{\Gamma}\,|\,\mathbf{h})\big) + \mathcal{D}_{KL}\big(p_{\mathbf{w}}(\mathbf{w}\,|\,\boldsymbol{\theta}_{\mathbf{w}})\;||\;p(\mathbf{w}\,|\,\mathbf{h})\big).$$

Собирая результаты вместе, получаем выражение для функции потерь:

$$\mathcal{L}(\boldsymbol{\theta}_{\mathbf{w}},\boldsymbol{\theta}_{\boldsymbol{\Gamma}}) = -\mathsf{E}_{q(\mathbf{w},\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta})}\Big[\log p(\mathbf{y}\,|\,\mathbf{w},\boldsymbol{\Gamma},\mathbf{X})\Big] + \\ + \mathcal{D}_{KL}\big(q_{\boldsymbol{\Gamma}}(\boldsymbol{\Gamma}\,|\,\boldsymbol{\theta}_{\boldsymbol{\Gamma}})\;||\;p(\boldsymbol{\Gamma}\,|\,\mathbf{h})\big) + \mathcal{D}_{KL}\big(p_{\mathbf{w}}(\mathbf{w}\,|\,\boldsymbol{\theta}_{\mathbf{w}})\;||\;p(\mathbf{w}\,|\,\mathbf{h})\big).$$

Анализируя полученное выражение для функции потерь, стоит отметить, что первое слагаемое представляет собой матожидание по вариационному распределению от логарифма правдоподобия модели. Учитывая этот факт, покажем, что предлагаемый метод повышает устойчивость модели. Для этого докажем следующие два утверждения, являющиеся обобщением утверждений, описанных в [2].

Теорема 1. Пусть заданы две структуры $\hat{\mathbf{\Gamma}_1}$ и $\hat{\mathbf{\Gamma}_2}$, функция ошибки на которых принимает одинаковое значение, то есть $\mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}_1}) = \mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}_2})$. Пусть также $||\nabla^2_{\mathbf{\Gamma}} \mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}_1})|| < ||\nabla^2_{\mathbf{\Gamma}} \mathcal{L}(\mathbf{w}, \hat{\mathbf{\Gamma}_2})||$. Тогда справедливо

$$||\hat{\mathbf{\Gamma}_1} - \mathbf{\Gamma}^*|| > ||\hat{\mathbf{\Gamma}_2} - \mathbf{\Gamma}^*||,$$

где $\mathbf{\Gamma^*} = \arg\min\limits_{\mathbf{\Gamma}} \mathcal{L}_{val}(\mathbf{w}, \mathbf{\Gamma})$ – оптимальная структура.

Доказательство.

Пусть $(\mathbf{w}^*, \mathbf{\Gamma}^*)$ – решение оптимизационной задачи (1), то есть $\mathcal{L}(\mathbf{w}^*, \mathbf{\Gamma}^*) = 0$. Обозначим через $\hat{\mathbf{\Gamma}}$ проекцию решения $\mathbf{\Gamma}^*$ на симплекс.

Разложим функцию потерь $\mathcal{L}(\mathbf{w}^*, \hat{\mathbf{\Gamma}})$ в ряд Тейлора в точке $(\mathbf{w}^*, \mathbf{\Gamma}^*)$:

$$\mathcal{L}(\mathbf{w}^*, \hat{\mathbf{\Gamma}}) = \mathcal{L}(\mathbf{w}^*, \mathbf{\Gamma}^*) + (\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)^{\mathrm{T}} \nabla_{\mathbf{\Gamma}} \mathcal{L}(\mathbf{w}^*, \mathbf{\Gamma}^*) + \frac{1}{2}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)^{\mathrm{T}} H (\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*) =$$

$$= \mathcal{L}(\mathbf{w}^*, \mathbf{\Gamma}^*) + \frac{1}{2}(\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*)^{\mathrm{T}} H (\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*),$$

где $H = \int_{\mathbf{\Gamma}^*}^{\hat{\mathbf{\Gamma}}} \nabla^2_{\mathbf{\Gamma}} \mathcal{L}(\mathbf{w}^*, \mathbf{\Gamma}) d\Gamma$.

Получаем, что величину потери точности после проекции характеризует величина $C = ||H|| \cdot ||\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}^*||$.

По условию $\exists\ \hat{\mathbf{\Gamma}_1}, \hat{\mathbf{\Gamma}_2} :\ \mathcal{L}(\mathbf{w}^*, \hat{\mathbf{\Gamma}_1}) = \mathcal{L}(\mathbf{w}^*, \hat{\mathbf{\Gamma}_2})$. Тогда

$$(\hat{\mathbf{\Gamma}_1} - \mathbf{\Gamma}^*)^{\mathrm{T}} H_1 (\hat{\mathbf{\Gamma}_1} - \mathbf{\Gamma}^*) = (\hat{\mathbf{\Gamma}_2} - \mathbf{\Gamma}^*)^{\mathrm{T}} H_2 (\hat{\mathbf{\Gamma}_2} - \mathbf{\Gamma}^*),$$

то есть $C_1 = C_2$. Без ограничения общности будем считать, что $||H_1|| < ||H_2||$.

В таком случае незамедлительно получаем $||\hat{\mathbf{\Gamma}_1} - \mathbf{\Gamma}^*|| > ||\hat{\mathbf{\Gamma}_2} - \mathbf{\Gamma}^*||$. А значит первая модель является более устойчивой. ∎

Теорема 2. Пусть задана функция $G(\mathbf{w}, \mathbf{\Gamma})$. Тогда для любого распределения $q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\delta})$ такого, что компоненты $\mathbf{\Gamma}$ – независимые случайные величины справедливо

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\delta})} \big[G(\mathbf{w}, \mathbf{\Gamma}) \big] \approx G(\mathbf{w}, \boldsymbol{\mu}) + \frac{\sigma^2}{2} \mathrm{Tr} \big(\nabla^2_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu}) \big),$$

где $\boldsymbol{\mu} = \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\delta})} \big[\mathbf{\Gamma} \big]$ и $\boldsymbol{\sigma}^2 = \mathsf{D}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\delta})} \big[\mathbf{\Gamma} \big]$.

Доказательство.

Разложим функцию $G(\mathbf{w}, \mathbf{\Gamma})$ в ряд Тейлора до второго порядка в точке $(\mathbf{w}, \boldsymbol{\mu})$:

$$\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}\,|\,\boldsymbol{\delta})}\big[G(\mathbf{w}, \mathbf{\Gamma})\big] \approx \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}\,|\,\boldsymbol{\delta})}\big[G(\mathbf{w}, \boldsymbol{\mu}) + (\mathbf{\Gamma} - \boldsymbol{\mu})^{\mathrm{T}} \nabla_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu}) + \frac{1}{2}(\mathbf{\Gamma} - \boldsymbol{\mu})^{\mathrm{T}} \nabla^2_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu})(\mathbf{\Gamma} - \boldsymbol{\mu})\big] =$$

$$= G(\mathbf{w}, \boldsymbol{\mu}) + \mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}\,|\,\boldsymbol{\delta})}\big[(\mathbf{\Gamma} - \boldsymbol{\mu})^{\mathrm{T}}\big] \nabla_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu}) + \frac{1}{2}\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}\,|\,\boldsymbol{\delta})}\big[(\mathbf{\Gamma} - \boldsymbol{\mu})^{\mathrm{T}} \nabla^2_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu})(\mathbf{\Gamma} - \boldsymbol{\mu})\big] =$$

$$= G(\mathbf{w}, \boldsymbol{\mu}) + \frac{\sigma^2}{2} \mathrm{Tr}\big(\nabla^2_{\mathbf{\Gamma}} G(\mathbf{w}, \boldsymbol{\mu})\big),$$

так как $\mathsf{E}_{q_{\mathbf{\Gamma}}(\mathbf{\Gamma}\,|\,\boldsymbol{\delta})}\big[(\mathbf{\Gamma} - \boldsymbol{\mu})^{\mathrm{T}}\big] = 0$, а диагональные элементы матрицы ковариации зануляются в силу независимости элементов $\mathbf{\Gamma}$. ∎

Таким образом, получается, что проведение байесовского вывода в задаче ПАНМ регуляризует норму гессиана по структуре. Кроме того, модель с меньшей нормой такого гессиана является более устойчивой. Наконец, можно сделать вывод, что нейросетевые модели с архитектурами, полученными с помощью предлагаемого метода, имеют повышенную устойчивость к внешним воздействиям.

4.2 Выбор априорных распределений

В качестве априорного распределения на параметры модели предлагается взять нормальное распределение с нулевым вектором средних и единичной матрицей ковариации, то есть

$$\mathbf{p}(\mathbf{w}|\mathbf{h}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Априорное распределение на вариационное распределение по параметрам предлагается также выбрать в качестве нормального:

$$\mathbf{q}_{\mathbf{w}}(\mathbf{w}|\theta_{\mathbf{w}}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}^{-1}).$$

Перейдем к выбору априорного распределения для структуры модели. Стоит отметить, что так как структура модели представляет из себя категориальное распределение вероятностей нмд операциями для каждого ребра графа модели, поэтому для обучения с помощью градиентных методов необходимо провести его релаксацию

на непрерывное множество. Для соблюдения этого требования на структуру модели задаётся распределение Gumbel-Softmax [13]. Такое распределение позволяет перевести категориальное распределение в непрерывное с помощью следующей репараметризации:

$$y_i = \frac{exp((g_i + log(\gamma_i)/\tau)}{\sum_i exp((g_i + log(\gamma_i)/\tau)},$$

где $G_i \sim Gumbel(0,1) = -\ln(-\ln(\mathcal{U}(0,1)))$ – случайная величина из стандартного распределения Гумбеля. Параметр температуры τ регулирует энтропию распределения, сохраняя при этом относительные ранги каждого события.

Предлагается в качестве априорного распределения на структуру выбрать конкретное распределение (Gumbel-Softmax):

$$\mathbf{q}_{\mathbf{\Gamma}}(\mathbf{\Gamma}|\theta_{\mathbf{\Gamma}}) \sim \text{Gumbel-Softmax}(\alpha_1, \dots \alpha_n).$$

Отметим, что мы не имеем никакой априорной информации о распределении $\mathbf{p}(\mathbf{\Gamma}|\mathbf{h})$. В таком случае предлагается в качестве априорного выбрать так называемое несобственное распределение [21], при котором реализация случайной величины равновероятна на всей области определения, то есть

$$\mathbf{p}(\mathbf{\Gamma}|\mathbf{h}) \sim \frac{1}{|\mathbf{\Gamma}|}.$$

Таким образом, процесс работы предлагаемого метода выглядит следующим образом:

1. Инициализация:
   1.1. параметры распределений $q_{\mathbf{w}}(\mathbf{w} \,|\, \boldsymbol{\theta}_{\mathbf{w}})$
   1.2. параметры распределений $q_{\mathbf{\Gamma}}(\mathbf{\Gamma} \,|\, \boldsymbol{\theta}_{\mathbf{\Gamma}})$.
   1.3. пространство поиска
2. Пока нет сходимости:
   2.1. Обновить структуру $\mathbf{\Gamma}$ с помощью шага градиента $\nabla_{\mathbf{\Gamma}} \mathcal{L}_{val}(\mathbf{w}, \mathbf{\Gamma})$.
   2.2. Обновить параметры $\mathbf{w}$ с помощью шага градиента $\nabla_{\mathbf{w}} \mathcal{L}_{train}(\mathbf{w}, \mathbf{\Gamma})$.
3. Сгенерировать архитектуру, основываясь на обученных структурных параметрах.

Видно, что описанный процесс представляет из себя итеративное поочередное обновление структуры и параметров модели до сходимости.

5 Вычислительный эксперимент

В данном разделе описаны используемые в работе наборы данных, процесс проведения эксперимента, используемые параметры обучения, а также анализ полученных результатов.

5.1 Данные

В качестве набора данных, на котором будет проходить поиск архитектуры, рассматривается набор изображений Fashion-MNIST. Он представляет из себя набор изображений для задачи классификации. Пример изображений представлен на рис. 7. Детальная описательная статистика представлена в табл. 1.

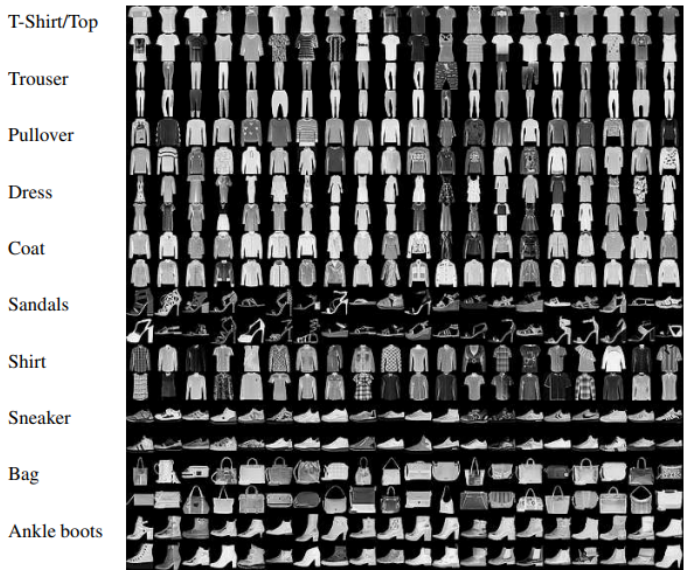


Рис. 7: Пример изображений набора данных Fashion-MNIST.

5.2 Эксперимент

В настоящей работе предлагается проводить поиск архитектуры свёрточной нейронной сети специального вида, называемой *клетка* 8. На вход такая архитектура

Таблица 1: Описательная статистика набора данных Fashion-MNIST

Размер изображения	Число классов	Размер $\mathfrak{D}_{train}$	Размер $\mathfrak{D}_{val}$
$1 \times 28 \times 28$	10	60000	10000

получает выходные объекты с двух предыдущих клеток. Выходом является конкатенация фильтров со внутренних промежуточных слоёв.

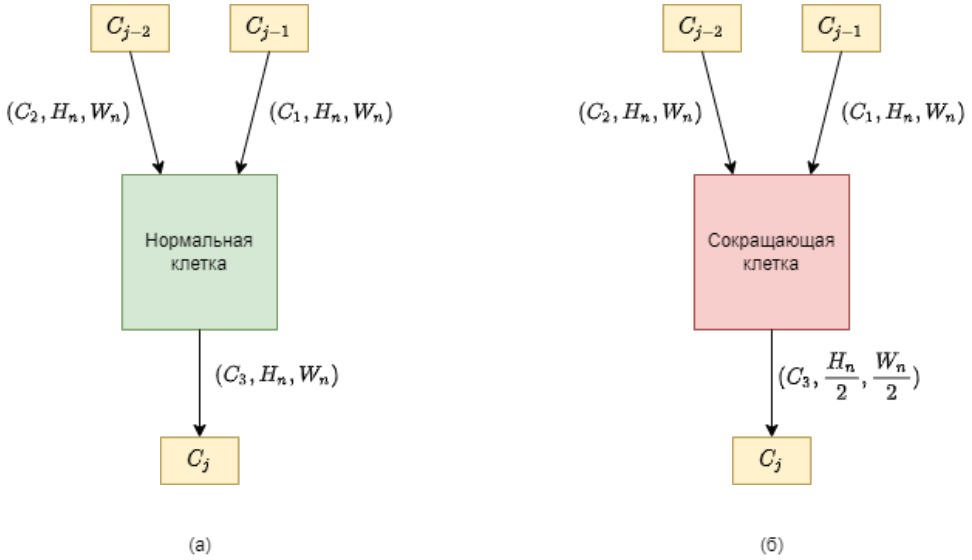


Рис. 8: Схематичная визуализация нормальной и сокращающей клеток, расположенных на (а), (б), соответственно. На вершинах указаны индексы предыдущих клеток. На рёбрах указаны промежуточные размеры входных изображений.

Клетки делятся на два типа – нормальная и сокращающая. Нормальная клетка возвращает тот же размер внутреннего состояния изображения, что был получен на вход. Сокращающая клетка уменьшает этот размер в два раза, но увеличивает число фильров в два раза. Каждая из клеток принимает на вход выходы с двух предыдущих клеток, имеет четыре внутренних состояния.

Поиск ведётся на структуре, состоящей из двух подряд идущих нормальных клеток и сокращающей (рис. 9).

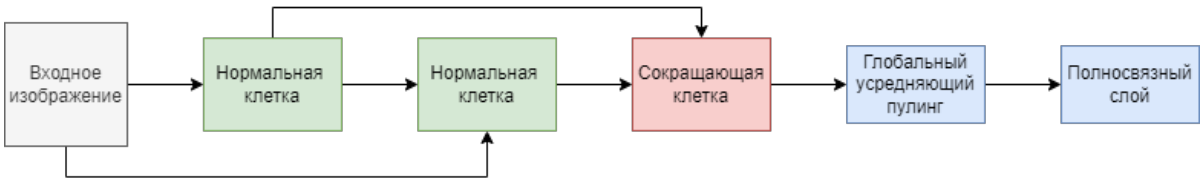


Рис. 9: Общая схема архитектуры, на которой ведётся поиск.

Используется пространство поиска, описанное в [16]. Оно состоит из семи операций, включающих в себя разные виды свёрток, пулинги и идентичной операции, равносильной отстутствию операции на ребре графа модели.

Конкретно пространство составлено из следующих операций:

- *максимальный пулинг 3x3*;
- *усреднённый пулинг 3x3*;
- *расширенная свёртка 3x3*;
- *расширенная свёртка 5x5*;
- *раздельная свёртка 3x3*;
- *раздельная свёртка 5x5*;
- *отсутствие операции.*

Для каждого слоя для сохранения размеров внутреннего представления изображения добавляется паддинг, если нужно. Для нормальных клеток шаг свёртки составляет 1, для сокращающих - 2.

После того, как метод завершит поиск архитектуры, модель с полученной структурой обучается заново до сходимости.

Для проведения эксперимента в качестве основного вычислительного ресурса используется GPU Nvidia Tesla T4. Эксперимент описан на ЯП Python 3.8 с помощью специализированных библиотек глубокого обучения PyTorch и NNI.

5.3 Анализ результатов

В текущей секции предоставлены результаты вычислительного эксперимента, и проводится сравнительный анализ с другими методами.

На рис. 10 приводится график зависимости значения функции ошибки от числа эпох обучения. Видно, что процесс обучения является стабильным и сходится к почти нулевому значению ошибки.

На рис. 11 визуализирована зависимость точности обученных архитектур от размера вносимого с помощью состязательной атаки на знак градиента шума в каждое входное изображение валидационной выборки.

На рис. 12 схематично показаны обученные нормальная и сокращающая клетки. Анализируя полученные структуры, стоит отметить следующие выводы. В нормальной клетке все внутренние состояния, кроме третьего, связаны исключительно со входными объектами. В сокращающей клетке, наоборот, вход с предпредыдущей

клетки почти никак не учитывается, а каждое из внутренних состояний зависит от предыдущих состояний и входа с предыдущей клетки.

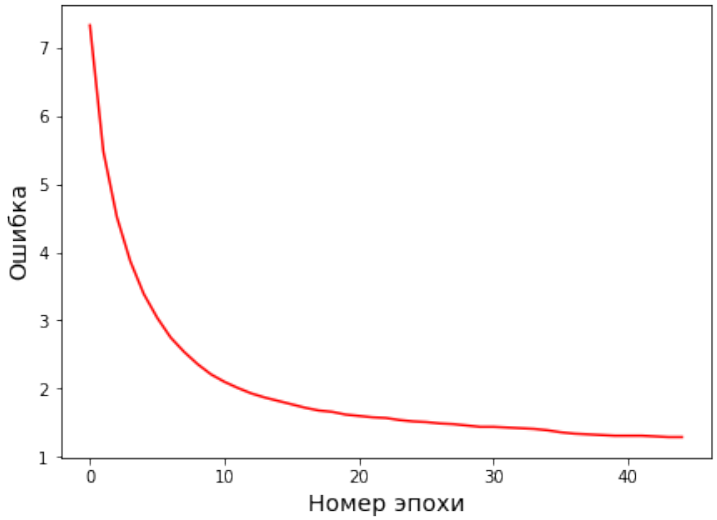


Рис. 10: График зависимости функции ошибки от числа эпох обучения.

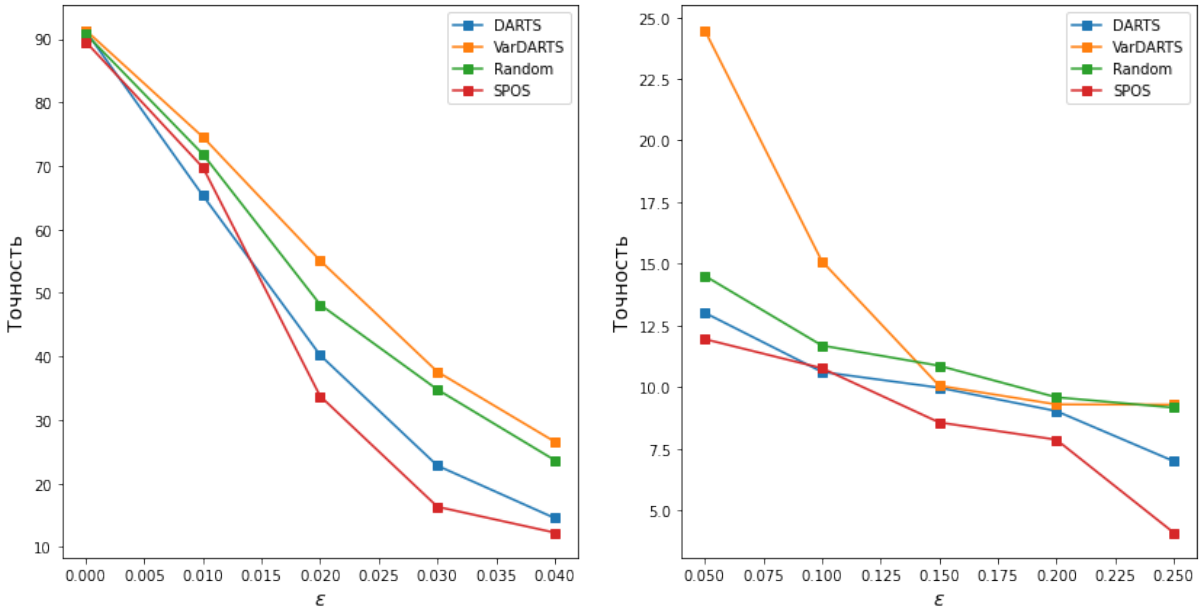
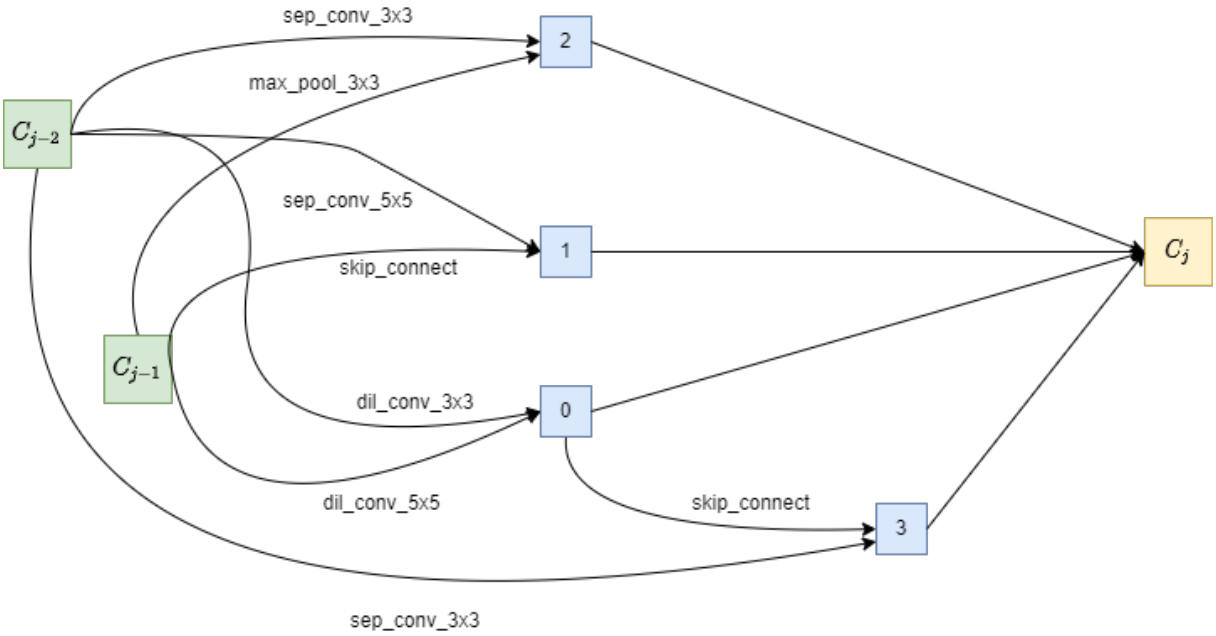
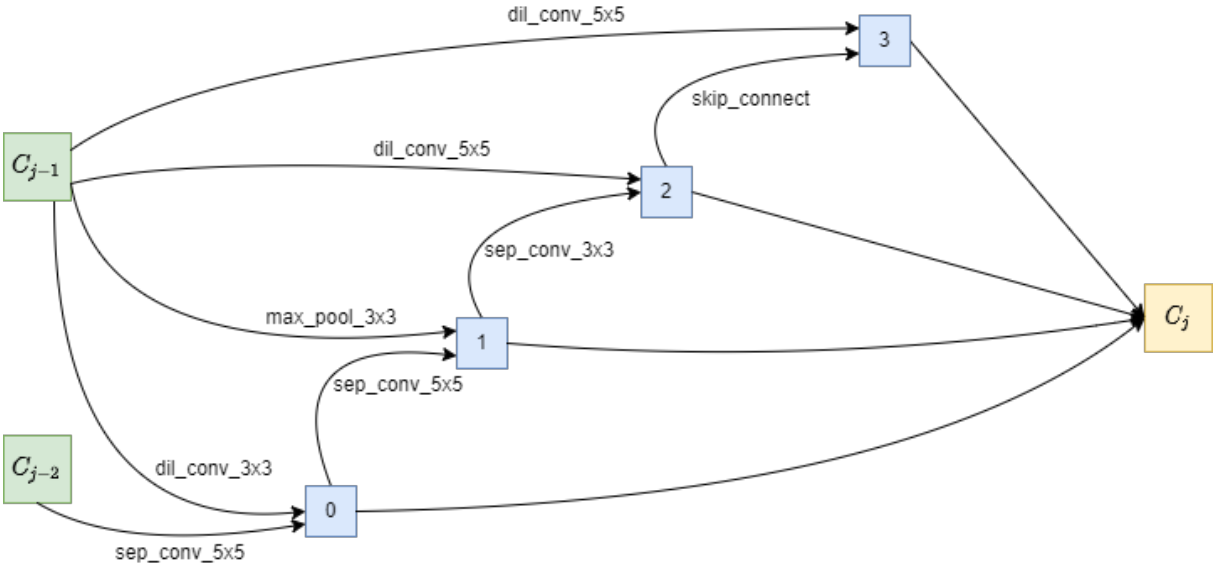


Рис. 11: Графики зависимости точности классификации от размера вносимого шума. Левый график соответствует малому шуму, правый – большому.



а)



б)

Рис. 12: Визуализация обученных клеток. (а): нормальная клетка, (б): сокращающая клетка.

В таблице 2 приводится сравнение предлагаемого метода VarDARTS с существующими подходами ПАНМ. Видно, что метод VarDARTS имеет наибольшую точность классификации, однако он также имеет наибольшее время сходимости.

В таблице 3 описаны результаты зависимости точности классификации от размера шума, вносимого с помощью состязательной атаки на знак градиента [8], для каждой обученной архитектуры. Анализируя полученные значения, можно заклю-

Таблица 2: Точность классификации найденных архитектур

Метод	Accuracy top-1, %	Время обучения, ч
DARTS	91.22	2
Random	90.79	1.5
SPOS	89.54	**0.5**
VarDARTS	**91.36**	5

Таблица 3: Устойчивость найденных архитектур в зависимости от вносимого шума FGSM-атаки.

Метод	ϵ=0.01	ϵ=0.02	ϵ=0.03	ϵ=0.04	ϵ=0.05	ϵ=0.1	ϵ=0.15	ϵ=0.2	ϵ=0.25
DARTS	65.32	40.20	22.81	14.56	13.01	10.61	9.97	9.02	6.99
Random	71.83	48.16	34.79	23.68	14.51	11.68	**10.86**	**9.58**	9.16
SPOS	69.74	33.75	16.34	12.27	11.94	10.76	8.56	7.86	4.09
VarDARTS	**74.54**	**55.09**	**37.59**	**26.59**	**24.47**	**15.10**	10.05	9.29	**9.28**

чить прирост устойчивости предлагаемого метода по отношению к другим подходам.

6 Заключение

В настоящей работе был предложен градиентный метод поиска архитектуры нейросетевой модели, оценивающий апостериорное совместное распределение структуры и параметров модели с помощью байесовского вывода. Была предложен метод внеения равномерно распределенного шума в параметры структуры. Была получена теоретическая интерпретация такого зашумления. Вычислительный эксперимент свидетельствует о повышении робастности предложенного метода относительно состязательных атак относительно базового метода.

В дальнейших исследованиях планируется:

1. добавить зависимость распределения параметров модели от её структуры;
2. добавить распределения на гиперпараметры модели;
3. заменить априорные предположения на другие распределения.

Список литературы

[1] Bayesnas: A bayesian approach for neural architecture search / H. Zhou, M. Yang, J. Wang, W. Pan // International conference on machine learning / PMLR. — 2019. — Pp. 7603–7613.

[2] *Chen X., Hsieh C.-J.* Stabilizing differentiable architecture search via perturbation-based regularization // International conference on machine learning / PMLR. — 2020. — Pp. 1554–1565.

[3] Drnas: Dirichlet neural architecture search / X. Chen, R. Wang, M. Cheng et al. // *arXiv preprint arXiv:2006.10355.* — 2020.

[4] Efficient neural architecture search via parameters sharing / H. Pham, M. Guan, B. Zoph et al. // International conference on machine learning / PMLR. — 2018. — Pp. 4095–4104.

[5] *Elsken T., Metzen J. H., Hutter F.* Neural architecture search: A survey // *The Journal of Machine Learning Research.* — 2019. — Vol. 20, no. 1. — Pp. 1997–2017.

[6] An empirical study on the robustness of nas based architectures / C. Devaguptapu, D. Agarwal, G. Mittal, V. N. Balasubramanian // *CoRR.* — 2020.

[7] *Ferianc M., Fan H., Rodrigues M.* Vinnas: Variational inference-based neural network architecture search // *arXiv preprint arXiv:2007.06103.* — 2020.

[8] *Goodfellow I. J., Shlens J., Szegedy C.* Explaining and harnessing adversarial examples // *arXiv preprint arXiv:1412.6572.* — 2014.

[9] Gp-nas: Gaussian process based neural architecture search / Z. Li, T. Xi, J. Deng et al. // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2020. — Pp. 11933–11942.

[10] *Graves A.* Practical variational inference for neural networks // *Advances in neural information processing systems.* — 2011. — Vol. 24.

[11] *Hochreiter S., Schmidhuber J.* Long short-term memory // *Neural computation.* — 1997. — Vol. 9, no. 8. — Pp. 1735–1780.

[12] *Hosseini R., Yang X., Xie P.* Dsrna: Differentiable search of robust neural architectures // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. — 2021. — Pp. 6196–6205.

[13] *Jang E., Gu S., Poole B.* Categorical reparameterization with gumbel-softmax // *arXiv preprint arXiv:1611.01144*. — 2016.

[14] *Jing K., Xu J.* A survey on neural network language models // *arXiv preprint arXiv:1906.03591*. — 2019.

[15] Learning transferable architectures for scalable image recognition / B. Zoph, V. Vasudevan, J. Shlens, Q. V. Le // Proceedings of the IEEE conference on computer vision and pattern recognition. — 2018. — Pp. 8697–8710.

[16] *Liu H., Simonyan K., Yang Y.* Darts: Differentiable architecture search // *arXiv preprint arXiv:1806.09055*. — 2018.

[17] *Lopes V., Alexandre L. A.* Hmcnas: Neural architecture search using hidden markov chains and bayesian optimization // *arXiv preprint arXiv:2007.16149*. — 2020.

[18] Object detection in 20 years: A survey / Z. Zou, Z. Shi, Y. Guo, J. Ye // *arXiv preprint arXiv:1905.05055*. — 2019.

[19] Practical black-box attacks against deep learning systems using adversarial examples / N. Papernot, P. McDaniel, I. Goodfellow et al. // *arXiv preprint arXiv:1602.02697*. — 2016. — Vol. 1, no. 2. — P. 3.

[20] Snas: stochastic neural architecture search / S. Xie, H. Zheng, C. Liu, L. Lin // *arXiv preprint arXiv:1812.09926*. — 2018.

[21] *Stone M., Dawid A.* Un-bayesian implications of improper bayes inference in routine statistical problems // *Biometrika*. — 1972. — Vol. 59, no. 2. — Pp. 369–375.

[22] Understanding and robustifying differentiable architecture search / T. E. Arber Zela, T. Saikia, Y. Marrakchi et al. // International Conference on Learning Representations. — Vol. 2. — 2020.

[23] *White C., Neiswanger W., Savani Y.* Bananas: Bayesian optimization with neural architectures for neural architecture search // Proceedings of the AAAI Conference on Artificial Intelligence. — Vol. 35. — 2021. — Pp. 10293–10301.

[24] *Wu R., Saxena N., Jain R.* Poisoning the search space in neural architecture search // *arXiv preprint arXiv:2106.14406.* — 2021.