Использование квантовых эмбеддингов при обработке данных

В. Л. Литвинов

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

vlad.litvinov61@gmail.com

Ф. В. Филиппов

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича

filippovfelix@gmail.com

И. В. Цыварев

Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича

cyvarev.ilya156@gmail.com

***Аннотация*. Фундаментальные принципы квантовых вычислений, в частности суперпозиция и запутанность, имеют решающее значение для применения квантовой обработки естественного языка. В работе исследуются вопросы использования запутанности при формировании квантовых эмбеддингов. Оцениваются особенности и возможности различных типов параметризованных квантовых схем. Рассматриваются вопросы использования SWAP тестирования для измерения вероятности квантовых состояний. Приводится пример использования ядер для классификации данных.**

Ключевые слова: машинное обучение; эмбеддинги; параметризованные квантовые схемы; оптимизация параметров; классификация данных

I. Введение

Все изучаемые объекты в классическом машинном обучении представляются с помощью векторов в многомерном евклидовом пространстве. Когда в качестве данных выступают изображения, слова и тексты – они представляются эмбеддингами, векторами «несущими» семантические признаки. Близость между классическими эмбеддингами измеряется с помощью метрик типа косинусного сходства или евклидова расстояния. Квантовый подход предполагает отображение данных в состояние квантовой системы, которое описывается волновой функцией в гильбертовом пространстве. Волновая функция представляется вектором состояния и этот квантовый эмбеддинг содержит комплексные амплитуды, квадраты модулей которых дают вероятности измерения соответствующих базисных состояний. Сходство квантовых эмбеддингов измеряется через скалярное произведение волновых функций, которое соответствует вероятности перехода из одного квантового состояния в другое. В подавляющем числе задач, связанных с обработкой естественного языка (NLP), очень важна интерпретация меры сходства эмбеддингов. В классическом варианте сходство – это геометрическая близость векторов. Квантовая близость – это вероятность «увидеть» одно состояние, когда приготовлено другое. Благодаря квантовой запутанности, эмбеддинги могут представлять связи между объектами нелокальным способом, который невозможно прямо выразить в классической геометрии.

II. Квантовый подход

В контексте NLP-задач и использования квантового подхода для работы с эмбеддингами можно выделить несколько практически интересных направлений. Они связаны с тем, как квантовые вычисления могут дополнить или улучшить классические подходы к обработке текста.

Например, использование параметризированных квантовых схем [1, 2, 3] для преобразования классических текстовых эмбеддингов в квантовые состояния позволит обогатить их за счет таких квантовых эффектов как суперпозиция и запутанность. Запутанность позволяет квантовым состояниям представлять сложные зависимости между признаками (например, словами в тексте), которые трудно или невозможно описать классическими методами. Квантовые состояния с запутанностью могут кодировать экспоненциально больше информации, чем классические векторы той же размерности. В задачах NLP (например, классификации текстов) запутанность может помочь лучше разделить классы за счёт использования неклассических корреляций.

Для практической реализации подобного преобразования можно применять амплитудное кодирование и параметризованные квантовые схемы. Пусть у нас есть классический вектор эмбеддинга $x = (x_1, x_2, \dots, x_n)$. Его можно закодировать в амплитуды квантового состояния:

$$|\psi\rangle = \frac{1}{\|x\|}\sum_{i=1}^{n} x_i\,|i\rangle,$$

где $\|x\|$ – норма вектора, $|i\rangle$ – базисные состояния. Далее, для получения квантового эмбеддинга подбирается оптимальная параметризованная схема. Здесь, мы исследуем известные схемы кодирования, такие как ZZFeatureMap, EfficientSU2 и TwoLocal. Из-за различий в их структуре и количестве параметров они генерируют разные квантовые эмбеддинги. Рассмотрим основные особенности этих схем.

В ZZFeatureMap ZZ-взаимодействия применяются между всеми парами кубитов, что создаёт полносвязную запутанность. Это означает, что каждый кубит запутан со всеми другими кубитами, что позволяет кодировать глобальные зависимости в данных. Формула ZZ-взаимодействия:

$$U_{ZZ}(\theta) = e^{-i\theta Z\otimes Z}.$$

Эти взаимодействия изменяют фазу состояния системы в зависимости от состояний обоих кубитов,

создавая корреляции между ними. Они играют ключевую роль в создании запутанности в квантовых схемах.

Структура EfficientSU2 состоит из чередующихся слоёв однокубитовых вращений и двухкубитовых CNOT-вентилей. Каждый слой включает вращения $R_Y(\theta)$ $R_Z(\phi)$ и запутывающие CNOT-вентили между соседними кубитами. Общая структура слоя:

$$U_{\text{слой}} = \prod_{i=1}^{n} R_Y(\theta_i) R_Z(\phi_i) \text{CNOT}_{i,i+1},$$

где θ_i и ϕ_i – параметры, которые оптимизируются в процессе обучения. Эта структура позволяет эффективно исследовать пространство состояний благодаря большому количеству параметров и создаёт запутанность между соседними кубитами, что позволяет захватывать сложные зависимости.

Наконец, структура `TwoLocal` состоит из локальных вращений и запутывающих вентилей, причем можно использовать разные типы запутывающих вентилей, такие как CNOT, CZ, CY. Общая структура слоя:

$$U_{\text{слой}} = \prod_{i=1}^{n} R\,(\theta_i) \text{Entanglement}_{i,i+1},$$

где $R\,(\theta_i)-$ локальные вращения, а Entanglement – запутывающие вентили. Эта структура позволяет задавать разные типы вращений, запутанность между кубитами позволяет захватывать сложные зависимости.

Таким образом, рассматриваемые структуры представляют собой мощные инструменты для создания параметризованных квантовых схем. Схема ZZFeatureMap играет ключевую роль в создании запутанности, EfficientSU2 позволяет эффективно исследовать пространство состояний, а TwoLocal, в свою очередь, предоставляет высокую гибкость.

Запутанность измеряется с помощью различных метрик, наиболее распространённой мерой является энтропия фон Неймана для подсистемы. Если у нас есть квантовая система, состоящая из двух подсистем A и B, то запутанность можно измерить, вычислив энтропию фон Неймана для одной из подсистем как:

$$S(\rho_A) = -\text{Tr}(\rho_A \log_2 \rho_A),$$

где $\rho_A -$ матрица плотности подсистемы A.

Матрица плотности ρ описывает состояние квантовой системы. Если система находится в чистом состоянии $|\psi\rangle$, то матрица плотности равна $\rho = |\psi\rangle\langle\psi|$. Матрица плотности для частично запутанных состояний описывает квантовую систему, которая находится в смешанном состоянии, состоящем из запутанных и сепарабельных компонент. Например, если система с вероятностью p находится в состоянии Белла $|\phi^+\rangle = \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, и с вероятностью $(1-p)$ – в сепарабельном состоянии $|00\rangle$, то её матрица плотности будет выпуклой комбинацией матриц плотности этих состояний:

$$\rho = p|\phi^+\rangle\langle\phi^+| + (1-p)|00\rangle\langle 00|.$$

Исследование влияния запутанности при использовании квантовых эмбеддингов выполнялось на наборе данных IMDB [4], содержащим 50 000 отзывов о фильмах. Для извлечения признаков из текстовых данных использовался метод TF-IDF (Term Frequency-Inverse Document Frequency). Был предпринят следующий вариант формирования классических эмбеддингов. В методе TF-IDF значение параметра *max_features* устанавливалась равным 100 и из полученных 100-компонентных векторов выделялось 16 главных компонент с помощью PCA. Преобразование в квантовые эмбеддинги выполнялось с помощью трех параметризированных схем ZZFeatureMap, EfficientSU2 и TwoLocal. Для оценки эффективности эмбеддингов использовался метод логистической регрессии **(**Logistic Regression) и метод опорных векторов (SVM). Результаты исследования приведены в табл. 1, где представлены оценки точности бинарной классификации отзывов, выполненной на основе классических/квантовых эмбеддингов.

ТАБЛИЦА I. ТОЧНОСТЬ КЛАССИФИКАЦИИ ОТЗЫВОВ

Квантовая схема	**Logistic Regression**	**SVM**
ZZFeatureMap	0.55/0.75	0.60/0.75
EfficientSu2, TwoLocal	0.55/0.70	0.60/0.70

Как видно из табл. 1, во всех случаях точность классификации при использовании квантовых векторов была выше.

III. КВАНТОВАЯ БЛИЗОСТЬ

Квантовую близость, то есть вероятность «увидеть» одно состояние, когда приготовлено другое, можно измерить физически с помощью квантовой схемы SWAP-теста (рис. 1).

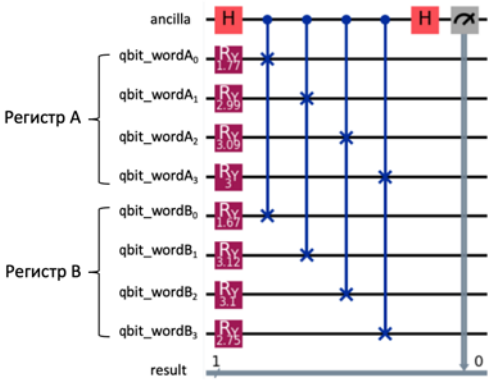

Рис. 1. Квантовая схема SWAP-теста для 16-разрядных эмбеддингов

Схема состоит из: кубита ancilla (верхний), который используется для измерения перекрытия слов, и двух регистров: регистр A (средние кубиты) – кодирует первое слово и регистр B (нижние кубиты) – кодирует второе слово. Управляемые CSWAP операции выполняются, когда ancilla в состоянии $|1\rangle$.

Измерение вероятности состояния $|0\rangle$ кубита ancilla после интерференции связано с перекрытием $|\langle\psi_A|\psi_B\rangle|^2$ слов по формуле:

$$p(0) = \frac{1 + |\langle\psi_A|\psi_B\rangle|^2}{2},$$

где ψ_A и ψ_B – это квантовые состояния, кодирующее первое и второе слово. Из этой формулы получаем косинусное расстояние между словами, как $d = \arccos(\sqrt{2p(0)-1})$. Для демонстрации квантовой близости было взято 10 классических эмбеддингов из модели GloVe [5], которые отражают важные линейные подструктуры векторного пространства слов. Для формирования квантовых эмбеддингов 100-размерные оригиналы были сокращены до 16 компонент. Далее, для случайного набора слов: sea, *peace*, *dog*, *mouse*, *bird*, *fish*, *morning*, *flower*, *red* и *moscow* с использованием параметризованной схемы ZZFeatureMap были получены квантовые эмбеддинги. На основе последних, с помощью схемы SWAP-теста (рис. 1) были определены косинусные расстояния d между словом *bird* и остальными словами. Результаты представлены на рис. 2, слева. Для сравнения, на рис. 2, справа представлены аналогичные данные, полученные на основе классических эмбеддингов.

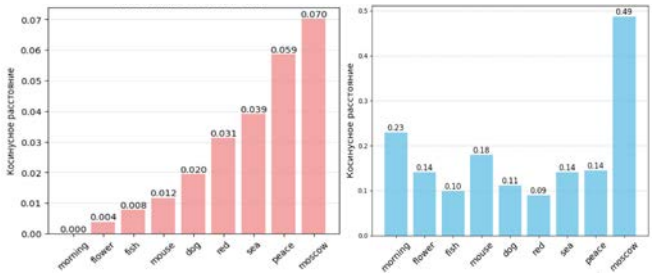

Рис. 2. Косинусные расстояния, рассчитанные по квантовым и классическим эмбеддингам

Исследование особенностей, обусловленных такими явлениями, как квантовая суперпозиция и запутанность, открывает доступ к изучению более сложных, нелинейных зависимостей между компонентами как классических, так и квантовых эмбеддингов. Примеры экспериментов с демонстрацией квантовых эмбеддингов доступны на colab ресурсе по ссылке [7].

IV. Квантовый эмбеддинг

Многие алгоритмы машинного обучения основаны на вычислении ядер – меры сходства между сложными объектами. Существует гипотеза, что квантовые компьютеры могут эффективно вычислять классы ядер, которые классическим компьютерам из-за экспоненциальных затрат недоступны. Квантовый эмбеддинг – это способ «поднять» данные в это высокоразмерное гильбертово пространство, где разделение классов становится тривиальным. Так можно расширить возможности линейного классификатора $y = sign(\langle w, x\rangle + b)$ за счет отображения $\phi(x)$ анализируемых точек x в пространство большей размерности и выполнять линейную классификацию в этом пространстве [6]. Если выбрать $\phi(x)$ так, чтобы образы точек стали линейно разделимы, то линейный классификатор в новом пространстве будет эквивалентен нелинейному классификатору в исходном.

Классический трюк ядер состоит в том, чтобы заменить скалярное произведение в исходном пространстве признаков на скалярное произведение в некотором пространстве большей размерности:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle.$$

В квантовом случае указанное отображение строится как $x \to |\Phi(x)\rangle$ — квантовое состояние, подготовленное на n кубитах параметризованной схемы. Тогда ядро между двумя точками определяется как квадрат модуля перекрытия состояний:

$$k(x_i, x_j) = \left|\langle \Phi(x_i)|\Phi(x_j)\rangle\right|^2.$$

Суть такой замены заключается в том, что вместо явного вычисления $\phi(x)$, используем ядро — функцию, которая вычисляет скалярное произведение напрямую, не вычисляя сами $\phi(x_i)$ и $\phi(x_j)$. Эти значения можно оценить на квантовом компьютере с помощью схемы SWAP-теста.

Рассмотрим простой пример, демонстрирующий описанные особенности использования ядер. В качестве исходных данных возьмем два линейно неразделимых класса объектов (рис. 3). Для построения квантового отображения используем параметризованную схему ZZFeatureMap(dim=2, reps=2, entanglement="linear"). Увеличение числа повторений reps повышает выразительность квантового ядра, но требует больше квантовых ресурсов. В эксперименте reps=2 обеспечивает баланс между точностью и сложностью. Линейная запутанность (entanglement="linear") подходит для данных с локальными корреляциями, в то время, как полная запутанность (entanglement="full") может улучшить точность для сложных данных, но требует больше кубитов.

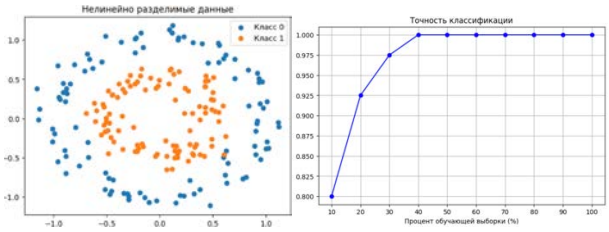

Рис. 3. Исходные нелинейно разделимые данные и точность классификации

Для двух признаков x_1, x_2 выбранная схема с двумя повторениями для кодирования признаков описывается как:

$$U_{ZZFeatureMap}(x) = \exp\left(-i\sum_{i,j} Z_i Z_j x_i x_j\right) H^{\otimes 2},$$

где Z – оператор Паули на i-ом кубите, H — оператор Адамара.

Квантовое ядро вычисляется как перекрытие между квантовыми состояниями, полученными после применения ZZFeatureMap к входным данным:

$$k(x_i, x_j) = |\langle 0|U^{\dagger}_{ZZFutureMap}(x_i) U_{ZZFutureMap}(x_j)|0\rangle|^2,$$

где $|0\rangle$ – начальное состояние кубитов. Далее используется классический SVM с ядром $k(x_i, x_j)$, вычисленным квантовым методом.

Квантовое ядро на основе ZZFeatureMap демонстрирует высокую точность классификации нелинейно разделимых данных даже при малом размере обучающей выборки (рис. 4). Высокая точность достигается уже при 40% обучающей выборки.

В табл. 2 приведены данные для сравнения точности различных методов классификации данных с использование ядер. Из данных таблицы следует, что квантовое ядро на основе ZZFeatureMap превосходит классические методы по точности для нелинейно разделимых данных.

ТАБЛИЦА II. ТОЧНОСТЬ «ЯДЕРНЫХ» МЕТОДОВ КЛАССИФИКАЦИИ

Метод	**Точность (100% обучающей выборки)**	**Зависимость точности от размера выборки**
Квантовое ядро (ZZ)	0.95-1.00	Быстрый рост
Линейное ядро SVM	0.50-0.60	Низкая точность
RBF-ядро SVM	0.85-0.90	Медленный рост
Полиномиальное ядро SVM	0.80-0.85	Умеренный рост

При попытке вместо ZZFeatureMap использовать схему PauliFeatureMap была получена более низкая точность результатов. Известно, что PauliFeatureMap выполняет тензорные произведения операторов Паули X, Y и Z для кодирования данных. Формула для одного блока:

$$U_{Pauli}(x) = \prod_j \exp(-ix_j P_j),$$

где P_j — оператор Паули.

Преимущества этой схемы состоит в гибкости при выборе конкретных операторов, что проще для интерпретации, если данные имеют простую структуру. Однако, это приводит к снижению эффективности для сложных нелинейных зависимостей, если не подобраны оптимальные операторы. Поэтому может потребоваться больше повторений (reps) для достижения той же выразительности, что и у ZZFeatureMap. Результаты эксперимента по выбору операторов и числу повторений для схемы PauliFeatureMap отражены на графиках рис. 4. Наилучшие результаты были получены при выборе оператора Z или X и числе повторений схемы reps = 2. При этом в обоих случаях была достигнута максимальная точность всего при 20% обучающей выборки. Варианты с оператором Y и XY оказались вовсе неприемлемыми.

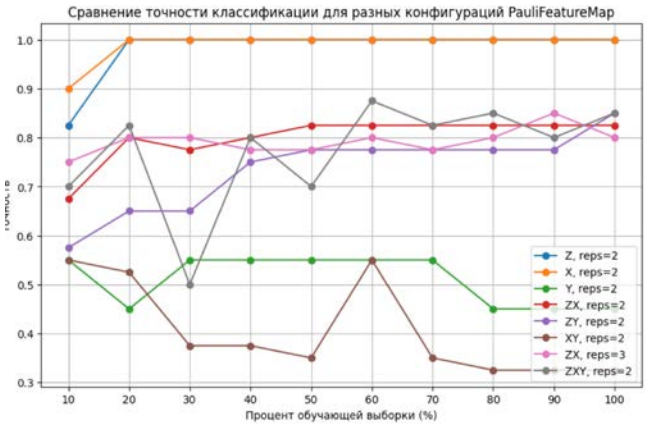

Рис. 4. Сравнение точности классификации

Примеры экспериментов с демонстрацией использования квантовых ядер доступны на colab ресурсе по ссылке [8].

V. ЗАКЛЮЧЕНИЕ

В ходе исследования подтверждена принципиальная возможность применения квантовых эмбеддингов, формируемых параметризованными схемами с использованием суперпозиции и запутанности, для задач обработки естественного языка. С помощью SWAP-теста удалось измерить квантовую близость между словами, получив интерпретируемые косинусные расстояния, согласующиеся с классическими аналогами. Продемонстрирована эффективность использования квантовых ядер для отображения анализируемых данных в пространство большей размерности, где разделение классов становится тривиальным. Полученные результаты показывают перспективы для дальнейшего изучения квантовых ядер, способных моделировать сложные нелинейные зависимости, недоступные классическим методам.

СПИСОК ЛИТЕРАТУРЫ

[1] Mina Abbaszade, Vahid Salari, Seyed Shahin Mousavi, Mariam Zomorodi, Xujuan Zhou. Application of quantum natural language processing for language translation. URL: https://ieeexplore.ieee.org/abstract/document/9525075. (Дата обращения 03.03.2026)

[2] Губин А.Н., Литвинов В.Л., Филиппов Ф.В. Градиентные методы обучения параметризованных квантовых схем. Актуальные проблемы инфотелекоммуникаций в науке и образовании (АПИНО 2024): Материалы XIII Международной научно-технической и научно-методической конференции. Санкт-Петербург, 2024. С. 268-273.

[3] Филиппов Ф.В., Жаранова А.О. Программное средство для вычисления градиента параметризованной квантовой схемы в гибридных квантово-классических алгоритмах. Свидетельство о регистрации программы для ЭВМ RU 2024616592, 21.03.2024. Заявка № 2024615156 От 07.03.2024.

[4] IMDB Dataset of 50K Movie Reviews. URL: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews. (Дата обращения 03.03.2026).

[5] GloVe model for distributed word representation. URL: https://sourceforge.net/projects/glove.mirror/ (Дата обращения 03.03.2026).

[6] B. Scholkopf and A. J. Smola, Learning with kernels: support vector machines, regularization, optimization, and beyond, Adaptive computation and machine learning (MIT Press, 2002).

[7] https://colab.research.google.com/drive/1e__xrsYK4Wv6qT4pantOAIPF_QVcr0IJ?usp=sharing.

[8] https://colab.research.google.com/drive/10ZeCDyILvQonbhBu0Haq3JpWTZIucNeM?usp=sharing