Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

ФАКУЛЬТЕТ КОМПЬЮТЕРНЫХ НАУК

Попов Никита Сергеевич

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Автоматическое построение ансамбля представлений через метрики сравнения пространств

Automatic Ensemble of Representation Development based on Space Comparison Metrics

по направлению подготовки 01.04.02 Прикладная математика и информатика
образовательная программа «Финансовые технологии и анализ данных»

Выполнил студент:
Попов Никита Сергеевич

Научный руководитель:
д-р техн. наук, Савченко Андрей Владимирович

МОСКВА
2025

Аннотация

В эпоху широкого распространения нейросетевых моделей важной становится задача эффективного объединения представлений (эмбеддингов), полученных различными моделями или слоями одной модели. В данной магистерской диссертации разработаны подходы к автоматическому построению ансамблей представлений на основе метрик сравнения пространств. Рассмотрены и проанализированы метрики сходства эмбеддингов, такие как CCA, CKA, RSA, RTD и другие. Предложено несколько методов формирования ансамблей, включая квадратичное программирование с учетом качества и разнообразия представлений, кластеризацию слоёв и жадные алгоритмы их комбинации. Экспериментальное исследование выполнено на задачах обработки текста с использованием трансформерных моделей и на транзакционных данных (эмбеддинги клиентов). Показано, что предложенные методы обеспечивают прирост качества относительно базовых подходов, таких как усреднение или выбор одного лучшего набора. Полученные результаты демонстрируют перспективность применения формальных мер различия для построения эффективных ансамблей представлений.

Abstract

In an era of widespread neural network applications, effectively combining embeddings from different models or layers becomes increasingly important. This master's thesis develops approaches for automatically constructing embedding ensembles based on representation space comparison metrics. Various similarity metrics, including CCA, CKA, RSA, RTD, and others, are reviewed and analyzed. Several ensemble construction methods are proposed, including quadratic programming considering representation

quality and diversity, layer clustering, and greedy algorithms for embedding combination. Experimental validation was performed on text-processing tasks using transformer models and on transactional data (customer embeddings). The proposed methods demonstrate improvements over baseline aggregation strategies, such as averaging or selecting the single best embedding set. The results highlight the promise of using formal measures of representation differences to build effective embedding ensembles.

Оглавление

ВВЕДЕНИЕ

Современные методы машинного обучения широко используют представления (эмбеддинги) данных — векторные описания объектов, автоматически извлекаемые моделями. Зачастую представления одних и тех же объектов могут быть получены из различных источников или моделей, например: разные модели для кодирования одного изображения, внутренние слои больших языковых моделей (LLM), кодирующие одну входную последовательность, различные источники для получения графовых/транзакционных эмбеддингов клиентов и т.д. Возникает практическая задача объединения нескольких пространств представлений в единый ансамблевый признак, чтобы улучшить качество решаемой задачи за счёт объединения знаний из разных источников. Известно, что объединение нескольких моделей или признаков (ансамблирование) обычно повышает точность и устойчивость системы за счёт усреднения шумов отдельных моделей и усиления устойчивых закономерностей. Однако ручной подбор способов слияния представлений затруднён, особенно когда число возможных комбинаций велико. Поэтому актуальной является задача автоматического построения ансамбля представлений на основе формальных мер сходства/различия между пространствами — метрик сравнения представлений. Такие метрики позволяют количественно оценить, насколько схожи два множества эмбеддингов, и могут служить основанием для алгоритмического выбора оптимальных комбинаций.

Цель данной работы — разработать подход к автоматическому построению ансамбля представлений различных моделей, опирающийся на анализ их сходства и разнообразия с помощью метрик сравнения пространств представлений. Для достижения этой цели необходимо решить

следующие задачи: (1) изучить и формализовать существующие метрики сравнения представлений (CKA, RSA, RTD и др.), проанализировать их свойства и применимость; (2) предложить метод автоматического выбора или построения ансамбля на основе указанных метрик (например, выбирая максимально дополняющие друг друга представления); (3) экспериментально проверить эффективность предложенного ансамбля на задачах из различных доменов.

Научный вклад данной работы заключается в следующем:

- установлена эмпирическая связь между метриками различия и выигрышем от объединения представлений;
- предложены и реализованы методы построения ансамблей эмбеддингов с использованием метрик различия между пространствами представлений: оптимизационные методы, ищущие баланс между качеством и разнообразием, а также жадные алгоритмы, последовательно включающие наиболее полезные представления;
- проведено экспериментальное сравнение предложенных стратегий ансамблирования на задачах объединения информации с разных слоёв трансформеров и последовательностей транзакций, описывающих клиентов; показано преимущество предложенных методов над базовыми подходами (усреднение, выбор лучшего представления).

Практическая ценность работы состоит в разработке инструментария для мульти-модельного представления данных, что востребовано, например, при интеграции разнородных источников данных о пользователях (транзакции, поведение, тексты) в финтех-приложениях, при объединении знаний разных слоёв нейросети для интерпретируемости или повышения точности, при комбинировании нескольких языковых моделей для улучшения семантического понимания текста и т.д.

1. Метрики сравнения пространств представлений

Сравнение двух множеств представлений (эмбеддингов), полученных из разных моделей или слоёв, требует метрики, способной выявить степень их эквивалентности либо различия. Рассмотрим основные метрики, предложенные в литературе для этой цели, их математические основы, примеры применения и различия между ними.

1.1. Canonical Correlation Analysis (CCA).

Канонический корреляционный анализ (CCA), введённый Hotelling в 1936 году, представляет собой классическую линейную статистическую технику для измерения многомерного сходства между двумя наборами признаков $X \in \mathbb{R}^{n \times p}$ и $Y \in \mathbb{R}^{n \times q}$, полученными для одного и того же набора из n объектов (стимулов). В контексте представлений нейросетей CCA применяется для поиска тех направлений (линейных комбинаций признаков) в обоих пространствах, проекции на которые максимизируют корреляцию между X и Y.

Формально, i-й канонический корреляционный коэффициент ρ_i определяется как максимум корреляции между линейными проекциями $Xw_x^{(i)}$ и $Yw_y^{(i)}$:

$$\rho_i = \max_{w_x^{(i)}, w_y^{(i)}} \operatorname{corr}(Xw_x^{(i)}, Yw_y^{(i)}),$$

при условиях ортогональности:

$$Xw_x^{(i)} \perp Xw_x^{(j)}, \quad Yw_y^{(i)} \perp Yw_y^{(j)}, \quad \forall j < i.$$

Наборы $Xw_x^{(i)}$ и $Yw_y^{(i)}$ называются каноническими переменными (canonical variables), а веса $w_x^{(i)}$ и $w_y^{(i)}$ — каноническими векторами. Полученные коэффициенты $\rho_1, \rho_2, \ldots, \rho_{N'}$, где $N' = \min(p, q)$, отражают

степень линейной зависимости между проекциями данных. Для агрегированной оценки сходства между двумя представлениями часто используется среднее значение первых N' канонических корреляций:

$$\bar{\rho} = \frac{1}{N'} \sum_{i=1}^{N'} \rho_i.$$

Это значение можно выразить также через ядерную норму следующего произведения:

$$\bar{\rho} = \frac{\|Q_Y^\top Q_X\|_*}{N'},$$

где $Q_X = X(X^\top X)^{-\frac{1}{2}}$ и $Q_Y = Y(Y^\top Y)^{-\frac{1}{2}}$ — ортонормированные базисы столбцов X и Y, а $\|\cdot\|_*$ — ядерная норма (сумма сингулярных значений).

Важно отметить, что CCA инвариантен к любым обратимым линейным преобразованиям признаков, таким как масштабирование, вращение и линейное смешивание. Это означает, что метод чувствителен исключительно к линейной взаимосвязи между подпространствами, не учитывая различия в их внутренней структуре. В контексте анализа нейросетевых представлений это свойство может ограничивать чувствительность метрики: например, два набора эмбеддингов, связанные произвольным линейным преобразованием, будут оценены как схожие, даже если их геометрические или топологические свойства различаются существенно. Таким образом, CCA может быть недостаточно дискриминативен для задач, где важно учитывать внутреннюю организацию представлений.

Несмотря на это, CCA остаётся популярным инструментом благодаря своей математической строгости, интерпретируемости и устойчивости в линейных задачах. Более того, он служит основой для усовершенствованных методов, таких как SVCCA и PWCCA, которые используют сингулярное разложение и взвешивание компонент с учётом их вклада в общую дисперсию. Эти методы нашли применение в визуализации

динамики обучения нейросетей, анализе активности слоёв и сравнении представлений различных архитектур.

1.2. Centered Kernel Alignment (CKA).

Центрированное выравнивание ядер (CKA) — это метрика на уровне представлений, предложенная Kornblith et al. (2019) на основе более ранней работы Gretton et al. (2005), которая оценивает степень зависимости (или сходства) двух пространств признаков при фиксированном наборе объектов. CKA инвариантна к ортогональным преобразованиям представлений, но не ко всем линейным обратимым, в отличие от CCA. Это делает её более чувствительной к геометрии пространства: она сохраняет скалярные произведения и евклидовы расстояния между парами объектов, что важно для оценки структурного сходства представлений.

CKA формально определяется через критерий независимости Хилберта–Шмидта (HSIC), применённый к двум ядровым матрицам K и L:

$$\mathrm{CKA}(K, L) = \frac{\mathrm{HSIC}(K, L)}{\sqrt{\mathrm{HSIC}(K, K) \cdot \mathrm{HSIC}(L, L)}},$$

где K и L — это матрицы ядер (kernel matrices), построенные по представлениям двух моделей $X = [\varphi_1, \ldots, \varphi_n]^\top$ и $Y = [\psi_1, \ldots, \psi_n]^\top$ на одном и том же множестве из n объектов. Элементы матриц задаются как $K_{ij} = \kappa(\varphi_i, \varphi_j)$ и $L_{ij} = \kappa(\psi_i, \psi_j)$, где κ — функция ядра. В линейном случае ядро κ является обычным скалярным произведением, и тогда $K = XX^\top$, $L = YY^\top$.

HSIC вычисляется как:

$$\mathrm{HSIC}(K, L) = \frac{1}{(n-1)^2} \operatorname{tr}(K_c L_c),$$

где K_c и L_c — центрированные ядровые матрицы, полученные вычитанием среднего по строкам и столбцам. Центрирование необходимо,

чтобы устранить влияние среднего положения точек и сделать метрику более стабильной.

CKA принимает значения от 0 до 1: значение близкое к 1 говорит о сильной зависимости (сходстве) структур представлений, тогда как значения около 0 означают независимость. Важно отметить, что CKA — это не просто корреляция признаков, а мера согласованности всех парных скалярных произведений между объектами в каждом пространстве. Благодаря этому она устойчива к произвольной перестановке признаков внутри каждого пространства и масштабным искажениям, если они одинаковы для всех объектов.

Именно эта структура делает CKA особенно полезной для сравнения скрытых представлений нейросетей, например, между слоями одной модели или между различными моделями, обученными на одних и тех же данных. В работе Kornblith et al. (2019) было показано, что CKA даёт более стабильные и интерпретируемые оценки, чем CCA, особенно при сравнении представлений в глубоких архитектурах.

1.3. Representational Similarity Analysis (RSA).

Анализ сходства представлений — методика, пришедшая из нейронауки, часто используется для сравнения структур различных репрезентаций (например, активности разных участков мозга или активаций разных моделей) [4, 5].

В основе RSA лежит идея сравнивать не сами координаты представлений, а матрицы (не)сходства между всеми парами рассматриваемых объектов в каждом пространстве. Для каждого пространства вычисляется матрица расстояний или представленческая матрица неоднородности (Representational Dissimilarity Matrix, RDM) размером $n \times n$, элементом $D(i, j)$ которой может служить, например, евклидово расстояние между эмбеддингами объекта i и j, или 1 минус коэффициент корреляции между активациями этих объектов [4].

Затем между двумя пространствами представлений вычисляется мера согласованности их RDM, чаще всего — корреляция (линейная

Пирсона или ранговая Спирмена) между соответствующими элементами двух матриц расстояний [6, 7].

Проще говоря, RSA отвечает на вопрос: одинаково ли две разные модели “чувствуют”, какие объекты похожи или различны? Если объекты, которые близки (или далеки) друг от друга в первом пространстве, также близки (или далеки) во втором, то корреляция между расстояниями будет высокой — это и указывает на структурную схожесть представлений.

В нейронауке RSA применяли для сравнения, насколько похожим образом структурированы стимулы в представлениях мозга и в активациях нейросети [4]. В машинном обучении RSA использовали для сопоставления слоёв моделей через RDM и наблюдали, что ранние слои разных CNN показывают высокое сходство, тогда как более глубокие слои — существенно различаются [1].

Важно отметить, что RSA, в отличие от CKA, чувствителен к монотонным преобразованиям расстояний — обычно применяется ранговая корреляция, нормализующая нелинейные и масштабные эффекты. Однако отсутствие строгой нормировки, как у CKA, делает RSA менее надёжной при сравнении признаков с различными распределениями. С другой стороны, RSA более универсален: можно подставлять любую меру расстояния (евклидову, косинусную, корреляционную и т.п.) в RDM, выбирая ту, что лучше отражает семантическую близость для задачи.

Частным случаем RSA является статистический тест Мантела, который измеряет коэффициент корреляции Пирсона между двумя матрицами расстояний и оценивает значимость этой корреляции методом перестановок [8].

1.4. Representation Topology Divergence (RTD)

Representation Topology Divergence (RTD) — метрика различия между пространствами представлений, основанная на методах топологического анализа данных (Topological Data Analysis, TDA). В отличие от геометрических метрик (CKA, RSA), RTD опирается на устойчивые топологические признаки графов, построенных по

эмбеддингам двух моделей, и тем самым фиксирует более сложные взаимозависимости между ними [10].

Построение графов.

Пусть заданы два массива эмбеддингов одинакового множества объектов:

$$X = \{x_1, \ldots, x_n\}, \qquad Y = \{y_1, \ldots, y_n\}, \qquad x_i, y_i \in \mathbb{R}^d.$$

Для каждого пространства формируем взвешенный граф близостей $G_1 = (V, E_1, w_1)$ и $G_2 = (V, E_2, w_2)$, где $V = \{1, \ldots, n\}$, а вес ребра (i, j) задаётся некоторой функцией близости (например, $w_1(i, j) = \|x_i - x_j\|_2$ или косинусное расстояние). Тем самым вся геометрия эмбеддингов переводится в дискретный граф.

Бар-коды (barcodes).

Для анализа графа вводится фильтрация по порогу τ: удаляем рёбра, вес которых превышает τ, получая семейство графов G^τ. По мере роста τ топологические свойства графа (количество компонент, циклов и т.д.) меняются. Каждый топологический признак фиксируется интервалом существования $[\tau_{\text{birth}}, \tau_{\text{death}}]$. Совокупность интервалов образует barcode и отражает устойчивость признаков при изменении масштаба [11].

R-Cross-Barcode.

Чтобы сравнить два графа, строят объединённый граф

$$G_{\text{union}} = (V, E_{\text{union}}, w_{\text{union}}), \qquad w_{\text{union}}(i, j) = \min\big(w_1(i, j),\, w_2(i, j)\big).$$

При фиксированном τ рассматривают три графа G_1^τ, G_2^τ и G_{union}^τ. Если две компоненты связности, раздельные в G_1^τ, объединяются в G_{union}^τ из-за рёбер, отсутствовавших в G_1^τ, фиксируется интервал $[\tau_{\text{union}}, \tau_1]$, где τ_1 — порог слияния в G_1, а τ_{union} — порог слияния в объединённом графе. Все такие интервалы образуют R-cross-barcode, который количественно

описывает «сколько» топологии одного графа приходится «догонять», чтобы совпасть с объединением.

RTD-оценка.

RTD вычисляют как среднее суммарной длины интервалов R-cross-barcode, усреднённое по нескольким случайным подмножествам вершин; такая агрегация уменьшает влияние локального шума. Значение RTD тем больше, чем сильнее расходятся топологические структуры двух представлений. Метрика инвариантна к размерности эмбеддингов: важна лишь идентичность множеств объектов, а не совпадение размерностей пространств.

Преимущества и ограничения.

RTD успешно выявляет расхождения представлений, которые не фиксируются линейными метриками, и показывает корреляцию с разницей в предсказаниях ансамблей моделей [10]. Ограничения — вычислительная затратность (расчёт persistent homology) и необходимость визуализации бар-кодов для интуитивной интерпретации численного значения.

1.5. Прочие метрики.

Помимо вышеперечисленных, в литературе предложен ряд других подходов к сравнению представлений. Исторически одними из первых стали методы на основе канонического корреляционного анализа (CCA). Например, метод SVCCA (Singular Vector CCA) использует усреднённую каноническую корреляцию между линейными проекциями двух наборов признаков [12], а метод PWCCA — взвешенную по дисперсии проекций корреляцию [13]. Такие методы способны выявлять линейно соответствующие подпространства в представлениях.

Однако их существенный недостаток — инвариантность к любому невырожденному линейному преобразованию: даже совершенно разные по сути представления могут дать максимально высокий коэффициент CCA, если между ними существует произвольное

обратимое линейное отображение [?]. Это делает CCA-метрики слишком «слабым» критерием сходства, поскольку они не различают, например, ситуацию, когда два представления связаны простой перестановкой координат (что действительно несущественно), и ситуацию, когда представление Y получается из X через сложную смешивающую матрицу, потенциально уничтожающую семантическую интерпретируемость. Именно поэтому более современные меры (CKA, RTD) стремятся зафиксировать структуру данных и игнорировать только тривиальные преобразования (ортогональные, изотропные масштабирования), не допуская произвольного вырождения соответствия.

Ещё один показатель — коэффициент Жаккара, широко применяемый для оценки сходства множеств. Его можно применять и к представлениям, если определить множество характерных признаков или соседей объекта в каждом пространстве. Например, можно для каждого объекта выбрать множество его k ближайших соседей (top-k) в первом пространстве и во втором; тогда коэффициент Жаккара J между этими двумя множествами соседей, усреднённый по объектам, даст меру того, насколько модели сохраняют соседние отношения между точками. Коэффициент Жаккара определяется как отношение размера пересечения множеств к размеру их объединения [14]:

$$J(A, B) \;=\; \frac{|A \cap B|}{|A \cup B|}\,,$$

где A и B — множества (например, соседей одного и того же объекта в двух разных эмбеддинг-пространствах). Значение $J = 1$ означает полное совпадение, $J = 0$ — полное различие. Данный коэффициент удобен для дискретизированных представлений или для оценки локальной структуры (сохранение соседств), однако менее информативен для глобальной структуры данных.

Помимо этого, можно отметить метрику дистанционной корреляции (distance correlation) — статистическую меру зависимости, которая обобщает понятие корреляции на случай векторов произвольной размерности, основываясь на расстояниях между точками. Дистанционная

корреляция равна нулю тогда и только тогда, когда представления статистически независимы. Её можно использовать для сравнения представлений, трактуя одно множество эмбеддингов как преобразование случайной переменной X, а другое — Y, и вычисляя степень зависимости X и Y через их попарные расстояния. Однако прямая интерпретация distance correlation для фиксированного набора детерминированных точек затруднена; чаще её применяют, когда рассматривают распределения точек. В задаче же сравнения двух заданных облаков данных более осмысленным оказывается именно коэффициент корреляции между матрицами расстояний (что эквивалентно RSA или тесту Мантела).

Таким образом, арсенал метрик сравнения представлений достаточно широк. CKA и RTD предоставляют более сложные, но инвариантные и надёжные оценки сходства, учитывающие нелинейную структуру данных; RSA и попарная корреляция расстояний просты и интерпретируемы, особенно при наличии априорных представлений о значимости расстояния; коэффициент Жаккара и другие специальные меры фокусируются на частных аспектах (например, сохранение соседей).

В каждом конкретном случае выбор метрики зависит от того, какие свойства представлений считаются существенными для задачи (глобальная геометрия или локальная структура, линейные соответствия или нелинейные). Для построения ансамблей представлений метрики сходства играют ключевую роль: они позволяют формально измерять разнообразие моделей. Например, низкое значение CKA между двумя наборами эмбеддингов указывает, что эти модели дополняют друг друга (их представления различаются), что потенциально означает выгоду от их объединения в ансамбле; наоборот, если CKA очень высок, модели избыточно похожи, и ансамбль может не дать выигрыша.

2. Объединение слоёв в трансформерах

2.1. Описание моделей

В экспериментах использовались три модели для получения эмбеддингов: BERT, RoBERTa и SimCSE. Все они относятся к классу трансформеров, но различаются архитектурными и обучающими особенностями, что делает их подходящими кандидатами для анализа различий представлений.

BERT (Bidirectional Encoder Representations from Transformers) — одна из первых двунаправленных трансформерных моделей, предложенная в 2018 году. Она предобучается на задачах маскированного моделирования языка и предсказания следующего предложения. Благодаря двунаправленному вниманию, BERT эффективно улавливает контекст как слева, так и справа от текущего токена, что позволяет формировать содержательные эмбеддинги. Мы используем базовую версию модели (BERT-base), включающую 12 слоёв и 110 млн параметров. Эмбеддинги извлекаются с каждого слоя и усредняются по токенам.

RoBERTa (Robustly Optimized BERT Approach) является модификацией BERT с переработанной стратегией обучения. Модель обучается дольше, на большем объёме данных, без задачи предсказания следующего предложения, а маскировка токенов выполняется динамически. Архитектура остаётся идентичной BERT, однако за счёт оптимизации предобучения RoBERTa демонстрирует более высокие результаты на большинстве задач. В наших экспериментах используется версия RoBERTa-base с 12 слоями.

SimCSE (Simple Contrastive Learning of Sentence Embeddings) — это метод дообучения моделей BERT и RoBERTa для получения качественных эмбеддингов предложений. Он основан на контрастивном обучении: модель обучается так, чтобы эмбеддинги идентичных (или близких по смыслу) предложений были ближе друг к другу, а различных — дальше. Существуют две версии SimCSE: без учителя (unsupervised), где положительные пары формируются за счёт dropout-аугментации, и с учителем (supervised), где используются аннотированные пары. SimCSE существенно повышает качество эмбеддингов по сравнению с исходными моделями. Мы используем версию, обученную с учителем, построенную на базе BERT.

2.2. Процедура оценки похожести слоев

Для оценки попарного сходства слоёв метрики рассчитывались независимо для всех трёх моделей, участвующих в экспериментах: BERT, RoBERTa и SimCSE (supervised версия). Для каждой модели были извлечены sentence embeddings с каждого слоя по случайной подвыборке из 30 000 предложений из корпуса C4. Представления получались путём усреднения скрытых состояний по токенам с учётом attention-маски:

$$\mathbf{v}_i = \frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \cdot h_t^{(i)},$$

где m_t — маска токена, а $h_t^{(i)}$ — вектор скрытого состояния на i-м слое.

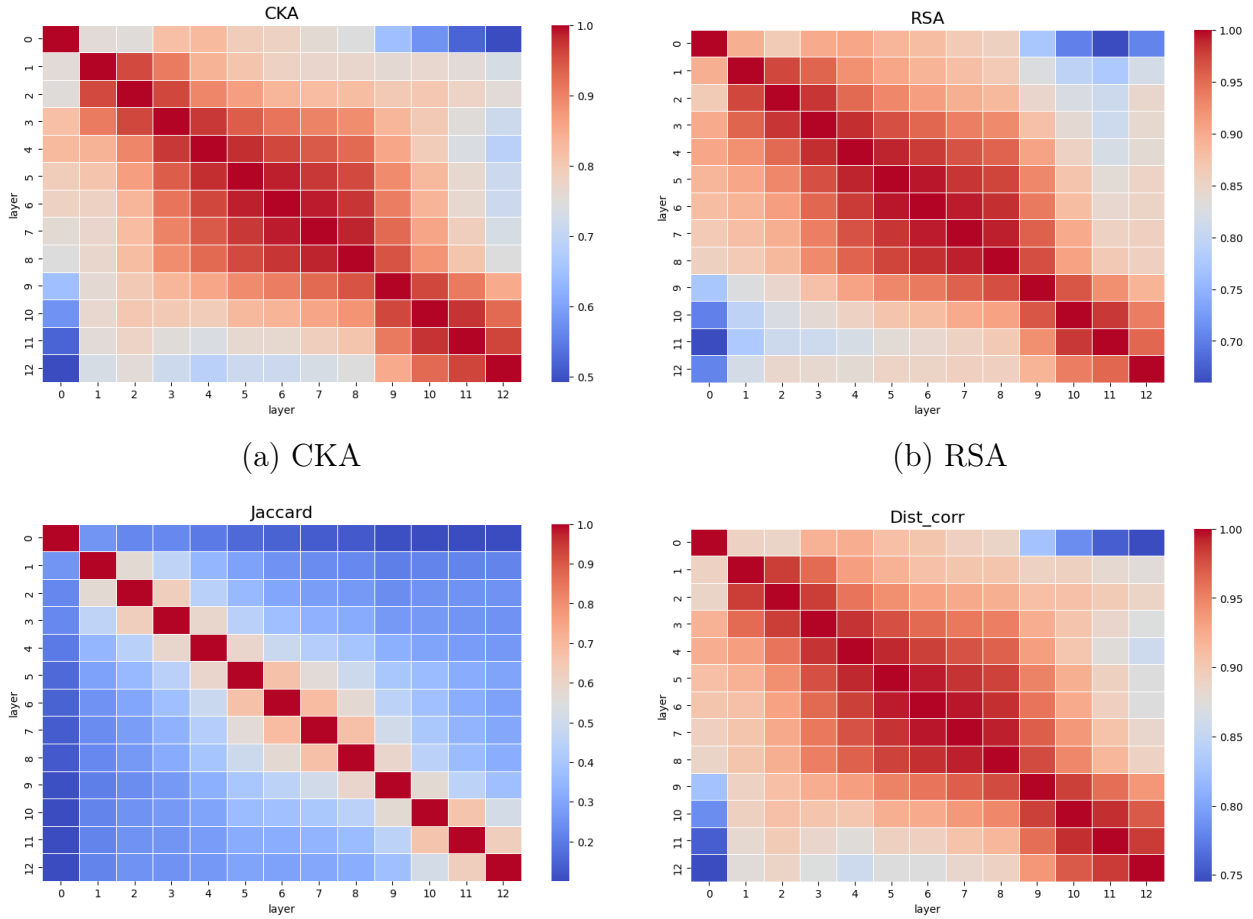
Для модели SimCSE помимо среднего эмббеддинга брался pooling output - эмбеддинг, специально дообученный кодировать предложение целиком.

Для каждой пары слоёв (i, j) последовательно вычислялись значения метрик CKA, RSA, Jaccard, Distance Correlation(метрика RTD не считалась, из-за очень больших вычислительных затрат). Расчёты проводились на случайных подмножествах данных, а итоговые значения усреднялись по нескольким запускам. Таким образом, для каждой модели

была построена полная матрица взаимного сходства между слоями размером num_layers×num_layers.

2.3. Выбор метрики

Визуальный анализ полученных тепловых карт (см. рис. 2.1) показывает, что все четыре метрики выявляют схожие структуры: выделяются группы центральных слоёв с высокой взаимной согласованностью, в то время как embedding-слой и финальный слой демонстрируют существенные отличия от остальных. Аналогичная картина наблюдалась в модели RoBERTa.



(a) CKA

(b) RSA

(c) Jaccard Similarity

(d) Distance Correlation

Рис. 2.1: Тепловые карты попарного сходства между слоями модели BERT по различным метрикам. Структура аналогична для RoBERTa.

Также была рассчитана корреляция между матрицами различных метрик (по Спирмену), которая во всех моделях превышала 0.9, что подтверждает согласованность оценок. Это позволяет сделать вывод о

надёжности и взаимной согласованности разных способов измерения структурной близости представлений.

Учитывая:

- высокую согласованность результатов между метриками;
- теоретические достоинства CKA (инвариантность к ортогональным преобразованиям, устойчивость к масштабированию);
- её широкое распространение и интерпретируемость в исследованиях представлений нейросетей [1],

в качестве основной метрики для всех последующих экспериментов была выбрана CKA. Она демонстрирует стабильные результаты, обладает теоретическим обоснованием и практической надёжностью при сравнении многомерных пространств представлений.

2.4. Методы построения ансамблей

В данной работе рассматриваются и сравниваются несколько стратегий объединения эмбеддингов, полученных с различных слоёв трансформерной модели. Все методы реализованы на базе библиотеки HuggingFace Transformers и предполагают использование mean- или cls-пулинга по токенам, а также микро-батчинг по входным предложениям для эффективного использования памяти.

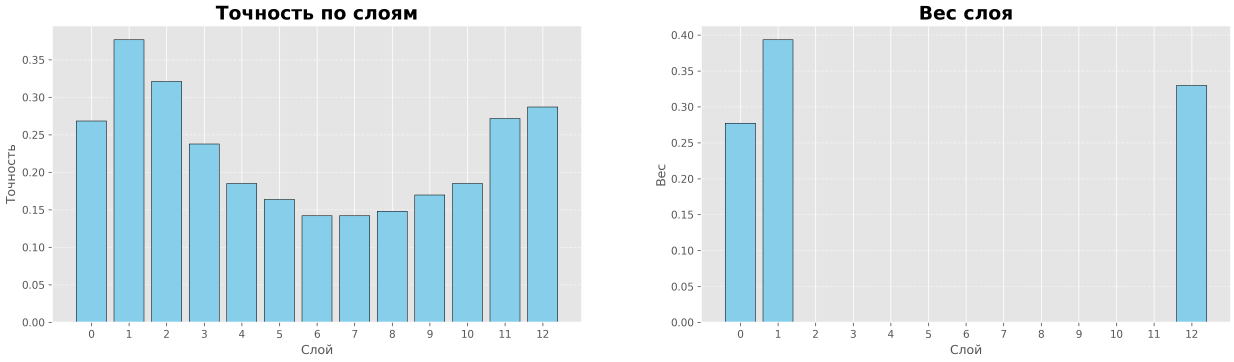
1. Усреднение последнего слоя (Last layer). Наиболее распространённый подход к извлечению sentence embeddings — использование последнего слоя трансформера. Представление предложения получается путём усреднения токенов (mean-pooling). Этот метод служит простой и широко применяемой базовой точкой отсчёта.

2. Выбор лучшего слоя (Best layer). Из всех слоёв модели выбирается один — тот, который даёт наилучший результат на валидационной выборке по целевой метрике. Только он и используется при формировании представлений. Такой подход позволяет использовать наиболее информативный слой, не смешивая остальные.

3. Взвешенное объединение слоёв через квадратичное программирование (QP weighted). Представления с разных слоёв объединяются с весами, полученными в результате решения задачи квадратичного программирования. Оптимизация направлена на баланс между качеством слоёв и их разнообразием:

$$\max_{w \in \mathbb{R}^L} \quad q^\top w - \lambda w^\top S w \quad \text{при } w \geq 0, \sum w_i = 1, \tag{2.1}$$

где q — вектор качества слоёв, S — матрица их сходства (например, по CKA), а λ — коэффициент баланса между точностью и разнообразием.



(a) Качество каждого слоя модели

(b) Оптимальные веса слоёв (многие равны нулю)

Рис. 2.2: Пример расчёта весов слоёв с помощью QP.

Как видно на рис. 2.2, многие слои получают нулевые или почти нулевые веса, что позволяет трактовать QP также как механизм отбора наиболее полезных слоёв. Это приводит к следующей модификации метода:

4. QP + отбор слоёв + PCA. Используется та же постановка задачи QP, однако после получения оптимальных весов зануляются все значения ниже порогового уровня $\varepsilon = 10^{-3}$:

$$w_i^{\text{final}} = \begin{cases} w_i, & \text{если } w_i > \varepsilon, \\ 0, & \text{иначе.} \end{cases}$$

Векторы представлений отобранных слоёв далее агрегируются с их весами, причём каждый вектор предварительно домножается на $\sqrt{w_i}$. Такое преобразование увеличивает дисперсию более значимых слоёв, тем самым

усиливая их влияние на последующую PCA-проекцию. Это позволяет учитывать относительную важность слоёв уже на этапе формирования пространства признаков:

$$\mathbf{v}_{\text{agg}} = \text{PCA}\left(\sum_{i \in \mathcal{S}} \sqrt{w_i} \cdot \mathbf{v}_i\right),$$

где $\mathcal{S}$ — множество выбранных слоёв. Подобная нормализация обеспечивает совместимость весов с геометрическим смыслом проекции и увеличивает устойчивость к включению слабо информативных признаков.

5. Кластеризация слоёв и объединение кластеров (cluster-mean + PCA). Слои группируются в K кластеров на основе матрицы попарной похожести (например, CKA). Из каждого кластера извлекается среднее представление (mean-pooling по слоям), которое далее домножается на $\sqrt{w_c}$, где w_c — средний вес слоёв внутри данного кластера. Это обеспечивает согласование влияния кластеров в итоговом пространстве:

$$\mathbf{v}_{\text{cluster},k} = \sqrt{w_c} \cdot \frac{1}{|C_k|} \sum_{i \in C_k} \mathbf{v}_i,$$

где C_k — множество индексов слоёв в k-м кластере. Полученные K векторов конкатенируются и проецируются в фиксированное пространство с помощью PCA. Такой подход позволяет учитывать как разнообразие представлений, так и их относительную значимость, сохраняя при этом компактность и интерпретируемость итоговых эмбеддингов.

Пример кластеризации слоев BERT по CKA можно увидеть на рисунке 2.3. В данном случае выделилось 4 кластера и довольно интуитивно: близкие слои объединились, нулевой слой в отдельной группе: (1) [9, 10, 11, 12], (2) [4 – 8], (3) [1 – 3], (4) [0].

6. Жадный отбор слоёв (greedy selection). Итеративный метод, который на каждом шаге выбирает следующий слой с максимальным приростом полезности:

$$\text{gain}_i = q_i - \lambda \cdot \frac{1}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} S_{ij},$$

Рис. 2.3: Дендрограмма кластеризации слоёв BERT по CKA. Видно четыре устойчивых кластера: (1) [9, 10, 11, 12], (2) [4 – 8], (3) [1 – 3], (4) [0].

где $\mathcal{S}$ — уже выбранные слои. После отбора k слоёв их представления агрегируются взвешенным образом (вес обратно пропорционален порядку выбора). Метод обеспечивает баланс между точностью и разнообразием, избегая включения избыточно похожих слоёв.

7. QP с маской похожести (masked QP). Данный метод представляет собой развитие классического подхода QP (см. формулу (2.1)) и сочетает идеи оптимизационного ансамблирования и жадного отбора. Используется маскированная матрица A, в которой сохраняются только элементы, отражающие схожесть с более качественными слоями:

$$A_{ij} = \begin{cases} S_{ij}, & \text{если } q_i < q_j, \\ 0, & \text{иначе,} \end{cases}$$

где q_i — значение валидационного качества i-го слоя, а S_{ij} — мера сходства между слоями i и j (например, CKA). Такое преобразование означает, что в функцию потерь включается только схожесть с более сильными слоями, игнорируя влияние менее качественных представлений.

Задача оптимизации формулируется в том же виде, что и в обычном QP, но с новой матрицей:

$$\max_{w \in \mathbb{R}^L} \quad q^\top w - \lambda w^\top A w \quad \text{при } w \geq 0, \ \sum w_i = 1.$$

Решение получается с помощью стандартных оптимизаторов

(например, SLSQP), а итоговые веса используются для построения взвешенного представления.

Метод обеспечивает более гибкий компромисс между разнообразием и точностью, выступая как промежуточный вариант между QP и жадным отбором: от первого он унаследовал непрерывную оптимизацию весов, а от второго — направленность на использование только «сильных» слоёв без штрафа за их схожесть с менее полезными.

2.5. Описание данных

Для оценки универсальности и эффективности методов построения ансамблей эмбеддингов были проведены эксперименты на задачах из открытого бенчмарка MTEB (Massive Text Embedding Benchmark) [15]. Этот бенчмарк включает широкий спектр задач, охватывающих различные аспекты семантического представления текста, и является де-факто стандартом для оценки sentence encoders.

В экспериментах были использованы задачи следующих типов:

- Классификация (classification) — определение категории для одного предложения;
- Парная классификация (pair classification) — бинарная классификация пар предложений (например, дубликат/не дубликат);
- Оценка семантической близости (STS) — предсказание степени смыслового сходства между двумя предложениями;
- Извлечение информации (retrieval) — поиск релевантных документов или вопросов по текстовому запросу.

В таблице 2.1 приведён список использованных датасетов и соответствующих им типов задач.

Название датасета	Тип задачи
Banking77Classification	Классификация
TweetSentimentExtraction	Классификация
STSBenchmark	Semantic Textual Similarity (STS)
SprintDuplicateQuestions	Парная классификация
QuoraRetrieval	Извлечение информации (retrieval)
NFCorpus	Извлечение информации (retrieval)

Таблица 2.1: Используемые датасеты из MTEB и типы соответствующих задач

2.6. Результаты экспериментов

Для каждой из трёх моделей (BERT, RoBERTa, SimCSE) отдельно отображаются значения качества по шести задачам из разных доменов и типов: классификация, парная классификация, семантическое сравнение и извлечение информации. Каждая строка таблицы соответствует конкретной задаче, а столбцы — методам объединения слоёв.

	Last Layer	Best Layer	QP Weighted	Cluster + PCA	QP + PCA	Greedy	QP masked
Banking77Classification	0.635	0.684	0.698	0.689	0.677	0.718	0.712
TweetSentimentExtractionClassification	0.518	0.528	0.523	0.499	0.526	0.535	0.535
NFCorpus	0.052	0.127	0.142	0.107	0.086	0.118	0.125
QuoraRetrieval	0.61	0.712	0.704	0.664	0.676	0.715	0.71
STSBenchmark	0.473	0.581	0.594	0.605	0.591	0.595	0.593
SprintDuplicateQuestions	0.368	0.785	0.782	0.807	0.826	0.782	0.772

Таблица 2.2: Результаты для модели BERT

	Last Layer	Best Layer	QP Weighted	Cluster + PCA	QP + PCA	Greedy	QP masked
Banking77Classification	0.63	0.673	0.701	0.627	0.708	0.697	0.694
TweetSentimentExtractionClassification	0.519	0.528	0.53	0.514	0.541	0.548	0.547
NFCorpus	0.013	0.039	0.039	0.026	0.038	0.03	0.022
QuoraRetrieval	0.542	0.624	0.634	0.642	0.637	0.645	0.637
STSBenchmark	0.544	0.551	0.558	0.552	0.549	0.571	0.571
SprintDuplicateQuestions	0.495	0.577	0.577	0.719	0.719	0.565	0.569

Таблица 2.3: Результаты для модели RoBERTa

	Pooling	QP Weighted	Cluster + PCA	QP + PCA	QP masked	Greedy
Banking77Classification	0.758	0.763	0.759	0.76	0.759	0.707
TweetSentimentExtractionClassification	0.597	0.597	0.601	0.602	0.584	0.592
NFCorpus	0.13	0.144	0.136	0.117	0.134	0.145
QuoraRetrieval	0.796	0.806	0.799	0.808	0.788	0.781
STSBenchmark	0.842	0.833	0.842	0.831	0.817	0.825
SprintDuplicateQuestions	0.73	0.826	0.814	0.832	0.833	0.827

Таблица 2.4: Результаты для модели SimCSE

Для сводного сравнения методов мы дополнительно смотрим:

- Средний ранг метода — среднее место, которое метод занимает среди всех на каждой задаче (чем ниже, тем лучше);
- Среднее качество метода — усреднённое значение целевой метрики по всем задачам и моделям (чем выше, тем лучше).

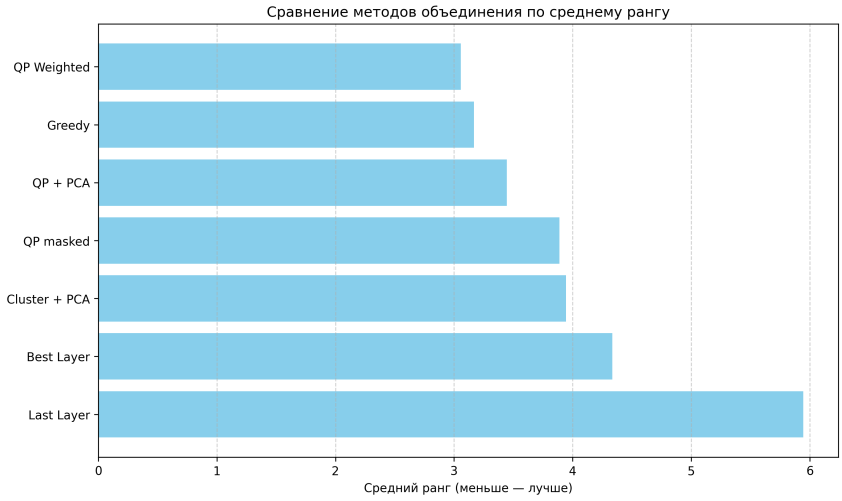


Рис. 2.4: Средний ранг метода (чем ниже — тем лучше).

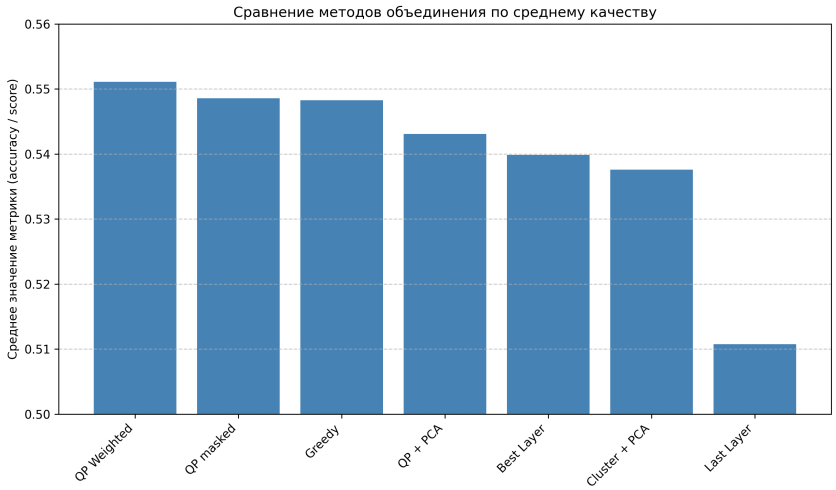


Рис. 2.5: Среднее качество по всем задачам (чем выше — тем лучше).

Обсуждение результатов для RoBERTa и BERT

- В подавляющем большинстве случаев любой из исследуемых методов работает лучше, чем наивное усреднение токенов с последнего слоя (Last Layer). Это ожидаемо, поскольку последний слой может быть переориентирован на предобученную задачу и терять семантическую универсальность.
- Метод Best Layer показывает относительно стабильные, но ограниченные результаты. С оптимальным слоем оказывается довольно тяжело соревноваться, но все же по среднему рангу он уступает всем предложенным методам.
- Метод QP Weighted показал лучшие результаты по всем сводным метрикам: средний ранг 3.06 и наивысшее среднее качество 0.551. Это подтверждает, что одновременно учитывать как качество слоёв, так и их различие — наиболее эффективная стратегия.
- Метод QP masked показал сопоставимое качество (0.549) при чуть большем среднем ранге (3.89). Это говорит о том, что исключение влияния слабых слоёв в процессе взвешивания может быть разумным компромиссом между разнообразием и надёжностью.
- Метод Greedy показал средний ранг 3.17 и качество 0.548. Он успешно отбирает дополняющие слои, но чувствителен к качеству

кандидатов, особенно в задачах с выраженным неравенством между слоями (например, retrieval). В таких задачах он зачастую включает слои с плохим качесвом, что портит итоговый ансамбль.

- QP + PCA и Cluster + PCA находятся в середине рейтинга: они немного уступают по качеству (0.543 и 0.538 соответственно) и имеют более высокие ранги (3.44 и 3.94). Это подтверждает, что дополнительная проекция не всегда оправдана, особенно если взвешивание уже эффективно выделяет информативные компоненты.
- Необычный случай наблюдается в задаче SprintDuplicateQuestions на модели RoBERTa: метод Cluster + PCA дал наибольший прирост, несмотря на то, что по результатам QP был выбран только один слой. Это может объясняться тем, что в данной задаче используется косинусное расстояние как основная метрика, и PCA, выполняющая ортонормализацию пространства, способствует устранению анизотропии и нормализации эмбеддингов, тем самым улучшая их сопоставимость в терминах косинусной близости.

Обсуждение результатов для SimCSE

- Для большинства задач удаётся улучшить качество относительно финального pooling-слоя, несмотря на то, что он явно обучен. Это говорит о наличии ценной информации в других слоях, не захваченной в итоговом представлении.
- Наилучшие результаты показали методы QP Weighted, QP + PCA и QP masked, что свидетельствует об универсальности этих подходов и их применимости даже в моделях с явно выделенным агрегатором.
- Метод Cluster + PCA также оказался полезным. Pooling-слой был выделен в отдельный кластер, что позволило сохранить его специфику и дополнить её более усреднёнными представлениями из других кластеров.
- Метод Greedy оказался менее успешен. Он не учитывает специфику архитектуры SimCSE, где информация уже агрегирована, и может

нарушать баланс, дополняя представление менее релевантными векторами.

Общие выводы

- QP Weighted лидирует как по среднему рангу, так и по качеству. Это делает его основным кандидатом для применения в прикладных задачах.
- Greedy и QP masked показали также высокие результаты, но требуют дополнительной настройки: первый — по критерию остановки при формировании комбинации, второй — по механизму маскирования.
- Cluster + PCA и QP + PCA работают надёжно, но редко превосходят QP без PCA. Однако они могут быть особенно полезны в задачах с сильно разнородными слоями.
- Простые методы (Last Layer, Best Layer) значительно проигрывают: последний — почти на 8% по средней метрике, а его ранг — худший. Это подчёркивает необходимость более осмысленного подхода к агрегации слоёв.

Дополнительно стоит отметить, что во всех рассмотренных методах параметр λ, регулирующий баланс между точностью и разнообразием, может быть предметом настройки. Более тщательная калибровка этого параметра потенциально способна улучшить результаты, особенно в задачах с различной чувствительностью к схожести между слоями.

3. Объединение эмбеддингов клиентов

В дополнение к слоёвым представлениям в трансформерах, важным направлением является сравнение и объединение различных эмбеддингов клиентов. В реальных приложениях это могут быть различные виды эмбеддингов, например, графовые, транзакционные, текстовые, эмбеддинги, полученные по историям покупок и тд. В данной работе мы пытаемся аггрегировать транзакционные представления клиентов, полученные разными моделями.

3.1. Источник данных и задача

В качестве источника использовался датасет транзакций клиентов из открытого набора данных по задаче предсказания возрастной группы: age-group-prediction. Каждая последовательность включает дату и категорию транзакции, а также сумму операции. Целевая переменная — возрастная группа клиента (бинарная классификация).

Для генерации разнообразных эмбеддингов использовалась библиотека pytorch-lifestream [16], предназначенная для обучения моделей последовательных транзакций. В каждом эксперименте (всего 102 прогона) обучалась отдельная модель с рандомизированными параметрами, включая:

- выбор подмножества признаков (категориальных и числовых);
- размерность эмбеддингов категориальных признаков;
- тип рекуррентного слоя (LSTM / GRU);
- наличие линейного проектора и его размерность;
- коэффициент шума и другие гиперпараметры модели CoLES;

- параметры сэмплирования: число срезов, минимальная/максимальная длина последовательности;
- параметры обучения (learning rate, число эпох и т.д.).

Таким образом, каждый запуск создавал уникальный набор эмбеддингов, обученный на одной и той же задаче, но на разных признаковых представлениях и с различными конфигурациями модели.

3.2. Анализ связи между различием представлений и выигрышем от объединения

Одной из ключевых гипотез данной работы является следующая: чем более различны два набора эмбеддингов, тем выше потенциальная выгода от их объединения. Чтобы количественно проверить эту гипотезу, был проведён специальный эксперимент, в котором сравнивались пары эмбеддингов клиентов, полученных из различных моделей, и оценивалось, насколько их объединение (конкатенация или суммирование) улучшает качество предсказания.

3.2.1. Описание эксперимента

Были собраны эмбеддинги клиентов, полученные в транзакционном кейсе, описанном ранее, и для каждой пары эмбеддингов $(\text{emb}_1, \text{emb}_2)$ были вычислены:

- индивидуальные качества acc_1, acc_2 на задаче предсказания возрастной группы;
- качество объединения: $\text{acc}_{\text{concat}}$ и acc_{sum} — для конкатенации и суммы соответственно;
- метрики различия между двумя наборами эмбеддингов (CKA, RSA, RTD, Jaccard и др.).

Качество модели оценивалось с помощью библиотеки LightAutoML, автоматически подбирающей оптимальный классификатор на основе переданных признаков.

3.2.2. Целевая переменная

В качестве целевой величины рассматривалась величина прироста от объединения:

$$\Delta_{\text{concat}} = \text{acc}_{\text{concat}} - \max(\text{acc}_1, \text{acc}_2),$$

$$\Delta_{\text{sum}} = \text{acc}_{\text{sum}} - \max(\text{acc}_1, \text{acc}_2).$$

Таким образом, нас интересовало, насколько новое объединённое представление превосходит лучшие из исходных.

3.2.3. Регрессия на остатки

Так как абсолютные значения точности могут зависеть от сложности задачи или самой модели, то для устранения влияния масштаба использовался следующий подход: сначала на Δ обучалась линейная регрессия от $\min(\text{acc}_1, \text{acc}_2)$ и $\max(\text{acc}_1, \text{acc}_2)$, и далее анализу подвергались остатки этой регрессии. То есть анализировалась не сама Δ, а её часть, необъяснимая качеством отдельных представлений.

3.2.4. Корреляционный анализ

Для каждой пары эмбеддингов были также заранее рассчитаны значения всех метрик различия между ними. Затем была вычислена корреляция между значением метрики (например, CKA-различие) и соответствующим значением очищенной Δ. Таким образом, каждая метрика получает скалярное значение — степень её согласованности с приростом от объединения.

В таблице 3.1 приведены значения коэффициентов корреляции между каждой из исследуемых метрик и приростом ROC-AUC при объединении представлений.

Таблица 3.1: Корреляция метрик различия с приростом качества от объединения (ROC-AUC)

Метрика	Корреляция
OrthogonalAngularShapeMetricCentered	+0.362
PermutationProcrustes	+0.358
OrthogonalProcrustesCenteredAndNormalized	+0.357
RSMNormDifference	+0.326
...	...
JaccardSimilarity	−0.300
RSA	−0.303
SoftCorrelationMatch	−0.304
RankSimilarity	−0.320
HardCorrelationMatch	−0.322
CKA	−0.339
DistanceCorrelation	**−0.390**

Положительное значения корреляции получились у метрик, которые показывают разницу, а отрицательные у тех, которые показывают схожесть. Видно, что гипотеза о том, что объединять имеет смысл наиболее отличающиеся наборы, подтверждается.

Аналогичные таблицы были построены для анализа прироста от суммирования, конкатенации и произведения. Наиболее стабильную сильную отрицательную корреляцию показала метрика CKA.

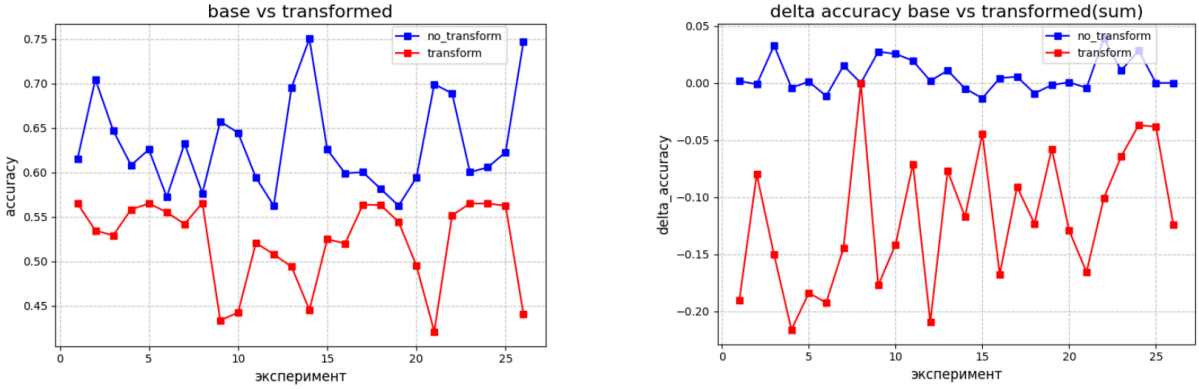
3.3. Минимизация корреляции через обучаемое преобразование

Было показано, что некоторые метрики, в частности корреляционные (например, CKA, RSA, Distance Correlation), демонстрируют сильную отрицательную связь с приростом точности при объединении эмбеддингов. Это означает, что высокая корреляция между представлениями может свидетельствовать о дублировании информации, и наоборот — различие

между эмбеддингами может способствовать более эффективному ансамблированию.

В данной секции была предпринята попытка обучить преобразование одного из эмбеддингов таким образом, чтобы оно минимизировало корреляцию (в частности, модифицированную версию distance correlation) с другим эмбеддингом. В качестве базовой модели использовалась небольшая MLP-сеть, которая по эмбеддингу X училась выдавать трансформированное представление X', минимизирующее корреляцию с эмбеддингом Y.

Несмотря на успешную оптимизацию метрики расхождения, качество объединения X' и Y оказывалось ниже, чем при использовании оригинального X. Более того, само трансформированное представление X' демонстрировало понижение качества по сравнению с оригинальным X, что говорит о потере полезной информации при агрессивной декорреляции.



(a) Точность исходного и трансформированного эмбеддинга

(b) Прирост при объединении (исходный/трансформированный)

Рис. 3.1: Сравнение исходного и трансформированного представлений.

Результаты показывают, что простое снижение корреляции между наборами эмбеддингов не гарантирует рост качества ансамбля. Необходим более сложный подход, который бы одновременно:

- снижал избыточную зависимость между представлениями,
- но при этом сохранял (или даже усиливал) информативность самих эмбеддингов.

Будущая работа может быть направлена на обучение таких преобразований в многокритериальной постановке, балансирующей между разнообразием и сохранением качества.

3.4. Итоговый эксперимент: объединение транзакционных эмбеддингов

На заключительном этапе была проведена серия из 102 запусков, разбитых на 17 независимых экспериментов. В каждом эксперименте обучались 6 различных моделей для извлечения эмбеддингов клиентов из транзакционных данных. Эти модели отличались по используемым признакам, архитектуре и гиперпараметрам (см. секцию 3.1.). Основная задача — объединить 6 представлений в итоговое векторное описание, обеспечивающее наилучшее качество в downstream-задаче классификации.

3.4.1. Сравнение базовых стратегий объединения

Были рассмотрены три базовых способа построения итогового эмбеддинга:

1. Конкатенация (Concat) всех 6 наборов эмбеддингов по признаковому пространству.
2. Максимум (Best) — выбор одного самого качественного набора.
3. Усреднение (Mean) — поэлементное среднее всех 6 представлений.

Результаты приведены на рис. 3.2.

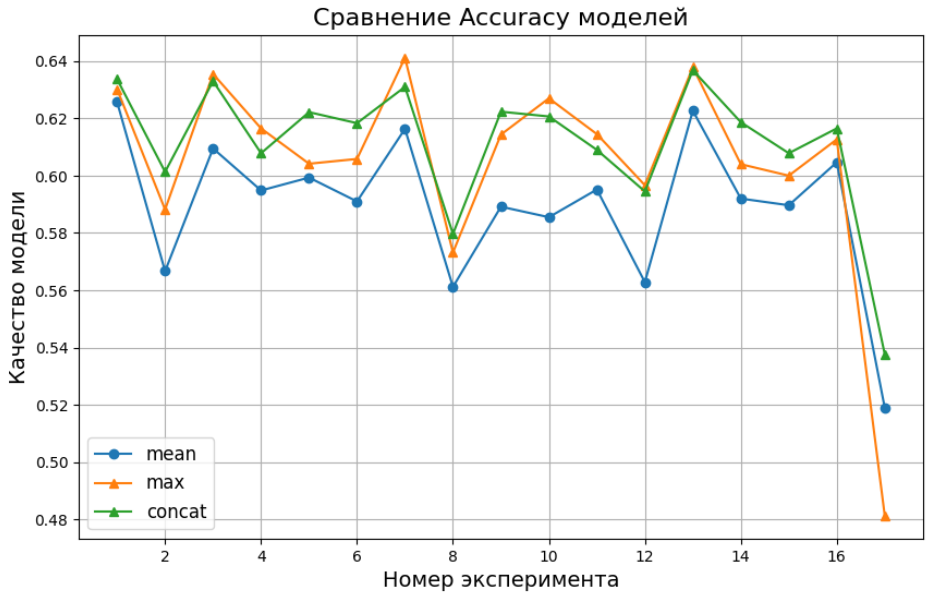


Рис. 3.2: Сравнение базовых методов объединения эмбеддингов: конкатенация, лучший отдельный и средний.

Как видно из графика, конкатенация демонстрирует прирост примерно на 1.2% по сравнению с максимумом. При этом простое усреднение, наоборот, приводит к деградации качества. Такая динамика отличается от поведения в задачах NLP, где усреднение слоёв часто повышает точность.

Одним из возможных объяснений является отсутствие согласованности между пространствами эмбеддингов: обученные на разных признаках и с разными параметрами, представления могут иметь различную ориентацию, масштаб и семантику. Усреднение векторов из таких пространств может не сохранять полезную информацию, в то время как конкатенация позволяет сохранить все аспекты разнородных эмбеддингов, хоть и увеличивает размерность. Кроме того, следует учитывать, что качество отдельных наборов эмбеддингов может существенно различаться: среди них встречаются как сильно информативные, так и слабо информативные представления. В условиях такой неоднородности простое усреднение или выбор случайной комбинации может не дать прироста, поскольку слабые эмбеддинги могут «размывать» вклад сильных. Именно поэтому даже конкатенация даёт лишь умеренное улучшение по сравнению с наилучшим индивидуальным представлением.

3.4.2. Жадный алгоритм объединения эмбеддингов

Для более гибкого и адаптивного объединения эмбеддингов был реализован жадный алгоритм на основе двух идей:

- отбор кандидатов, максимально дополняющих текущее представление по метрике CKA;
- поэтапное обновление текущего эмбеддинга путём замены наиболее «слабых» компонент на наиболее информативные из кандидата.

Алгоритм состоит из следующих шагов:

1. В качестве начального эмбеддинга выбирается тот, у которого максимальная энтропия сингулярных значений (метрика RankMe).
2. Итеративно к текущему представлению подбирается кандидат, дающий наибольший информационный прирост (наименьшее CKA).
3. Объединение выполняется через частичную замену компонент с низкой дисперсией на высокоинформативные признаки нового кандидата. Количество замен зависит от величины прироста (см. формулу (3.1)).

$$n_{\text{replace}} = \lfloor N \cdot p \cdot g \rfloor \tag{3.1}$$

где N — общее число признаков, $p = 0.2$ — максимальная доля заменяемых компонент, g — прирост (обратное значение CKA).

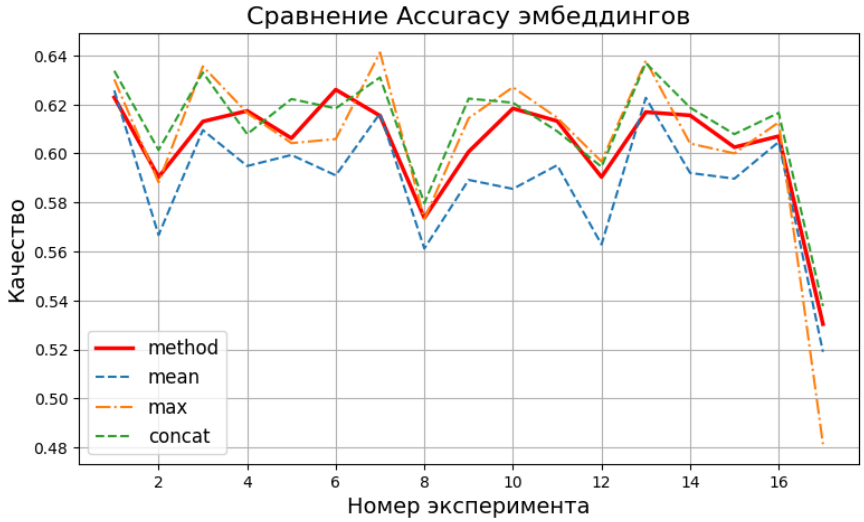


Рис. 3.3: Сравнение качества жадного объединения и лучшего индивидуального набора эмбеддингов.

Результаты жадного метода представлены на рис. 3.3. В половине случаев достигается выигрыш по сравнению с наилучшим индивидуальным представлением (baseline maximum), хотя информация о качестве не используется в алгоритме. При этом итоговая размерность представления остаётся неизменной, в отличие от конкатенации. Это подчёркивает способность жадного алгоритма извлекать дополняющую информацию из разных источников, эффективно балансируя между разнообразием и качеством. Метод показывает стабильные результаты и может применяться в системах с ограничениями по памяти или времени работы, где важно избегать лишнего увеличения размерности входа.

3.5. Модификация жадного алгоритма с учётом качества

В предыдущем варианте жадного алгоритма выбор и объединение эмбеддингов происходили исключительно на основе метрики различия между пространствами (например, CKA). Однако это не всегда гарантирует улучшение итогового качества, особенно при добавлении представлений с низкой прогностической способностью. Для улучшения стратегии был разработан комбинированный метод, учитывающий как разнообразие эмбеддингов, так и качество каждого из них на валидации.

Основные идеи

- В качестве первого набора эмбеддингов выбирается тот, который показывает наилучшее качество классификации на валидационной части данных.
- При добавлении новых кандидатов используется комбинированная метрика:

$$\text{score} = \alpha \cdot \text{gain}_{\text{CKA}} + (1 - \alpha) \cdot \text{accuracy}_{\text{val}} \,,$$

  где α — гиперпараметр, определяющий баланс между разнообразием и качеством.
- Число координат, подлежащих замене, динамически определяется на основе значения score и текущего шага алгоритма, что позволяет контролировать агрессивность изменений.
- Замена координат осуществляется с учётом важности признаков, вычисленной по важностям модели LightGBM, обученной на текущем представлении.

Таким образом, стратегия направлена на то, чтобы избежать включения в ансамбль слабых эмбеддингов и сохранить сильные признаки от уже выбранных источников.

Результаты

Сравнение результатов работы модифицированного алгоритма с максимумом по отдельным представлениям показано на рис. 3.4. Алгоритм демонстрирует устойчивое преимущество над наивным подходом, опирающимся только на качество одного лучшего набора.

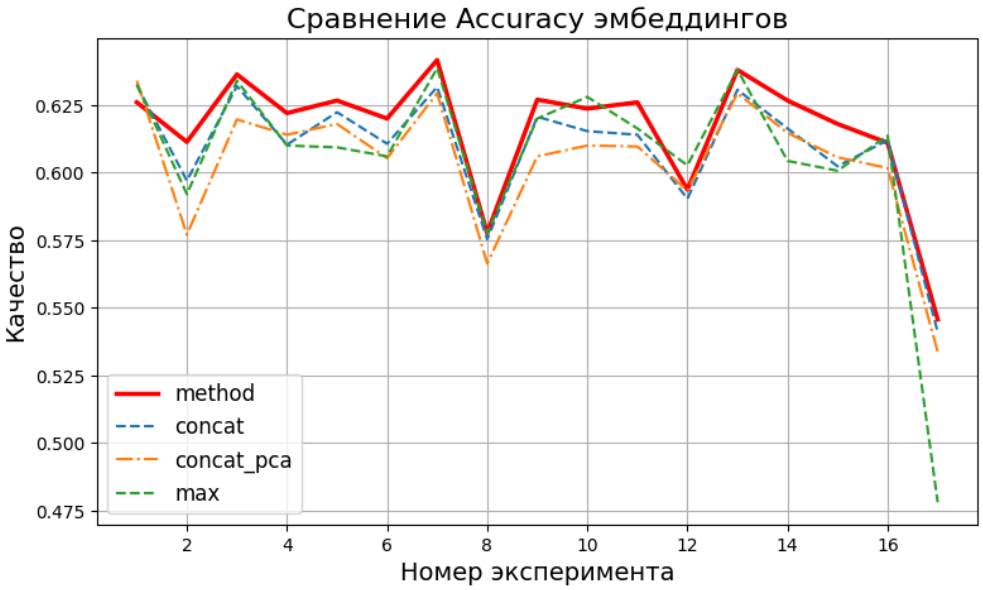


Рис. 3.4: Сравнение точности модифицированного жадного алгоритма с максимумом по отдельным эмбеддингам, конкатенацией, конкатенацией с PCA

Комбинирование информации о разнообразии и предсказательном качестве позволяет формировать более информативное представление при сохранении фиксированной размерности. Такой подход может быть полезен в случаях, когда необходимо учитывать различную полезность источников признаков и при этом избегать избыточности.

Таким образом, жадный алгоритм демонстрирует как интерпретируемость (понятная логика добавления слоёв), так и эффективность, показывая прирост по качеству без увеличения размерности.

Следует отметить, что в 88% экспериментов итоговое качество, полученное с помощью модифицированного жадного алгоритма, оказалось выше, чем при наивной конкатенации всех эмбеддингов. Это подтверждает эффективность учёта как валидируемого качества, так и разнообразия в процессе построения представления.

3.6. Оптимизация ансамбля с использованием валидационного качества и метрик различия

В данной секции рассматривается подход к построению ансамбля эмбеддингов на основе одновременно двух факторов: качества представлений на валидационной выборке и степени их различия. Метод вдохновлён идеями из работы [17], где предлагалось формировать ансамбль на основе метрики RTD. Однако в нашей работе вместо RTD использовались метрики, показавшие наибольшую отрицательную корреляцию с приростом качества при объединении эмбеддингов (см. секцию 3.2.): CKA, RSA, Distance Correlation и Orthogonal Angular Shape Metric Centered.

Метод абсолютно аналогичен тому, который использовался при объединении слоёв в трансформерах (см. описание в разделе 2.4.).

Сравнение с простым усреднением (baseline) показано в таблице 3.2. Во всех случаях наблюдается устойчивое преимущество по качеству (в среднем на 1.3–1.5%) и в 70–82% запусков ансамбль превосходит обычное среднее:

Таблица 3.2: Преимущество над усреднением.

Метрика	Mean (%)	Std (%)	Win (%)
CKA	1.39	1.45	70.6
RSA	1.36	1.63	76.5
MyCorrelation	1.31	1.62	76.5
Orth. Angular Shape	1.52	1.32	82.4

При сравнении с максимумом по качеству (т.е. лучшим индивидуальным набором в эксперименте), метод, напротив, показывает отрицательную дельту (таблица 3.3). Это объясняется тем, что объединение может «размывать» сильные представления, особенно без предварительного согласования пространств:

Таблица 3.3: Сравнение с максимумом (отрицательные значения — проигрыш).

Метрика	Mean (%)	Std (%)	Win (%)
CKA	-2.74	3.08	5.9
RSA	-2.76	3.27	17.6
MyCorrelation	-2.81	3.08	11.8
Orth. Angular Shape	-2.62	2.82	5.9

Выводы

Предложенный подход демонстрирует стабильное преимущество по сравнению с простым средним объединением эмбеддингов. Однако он пока не превосходит лучшие индивидуальные представления, что указывает на потенциал дальнейших улучшений — в частности, через методы согласования пространств или более сложные формы взвешенного объединения.

3.6.1. Попытка полностью unsupervised-оптимизации

В дополнение к ранее рассмотренным методам, в которых использовалась информация о валидационном качестве, была исследована возможность полностью unsupervised-оптимизации. Идея заключается в использовании той же задачи квадратичного программирования, что и ранее (см. формулу (2.1)), но с заменой вектора q — оценок качества на валидации — на различные метрики оценки качества эмбеддингов без разметки.

Рассматривались следующие unsupervised-метрики:

- Pseudo Condition Number — отношение максимального и минимального сингулярных значений эмбеддингов.
- RankMe [18] — оценка ранговой разнообразности представлений.
- Coherence — средняя косинусная близость между эмбеддингами.

- Clustering Quality — внутрикластерное расстояние на k-means (без использования меток).
- Persistent Homology (H0) — топологическая характеристика структуры данных.

Результаты объединения с весами, полученными на основе этих метрик, представлены в таблице 3.4. В качестве базовой линии использовано простое среднее. Как видно, ни одна из метрик не привела к систематическому улучшению — наблюдается либо ухудшение, либо отсутствие стабильной тенденции.

Таблица 3.4: Сравнение unsupervised-метрик с baseline (mean).

Метрика	Mean (%)	Std (%)	Win (%)
Pseudo Condition Number	-0.43	6.11	52.9
RankMe	-3.07	4.93	35.3
Coherence	-2.32	4.81	47.1
Clustering	-0.03	2.41	47.1
H0 (Persistent Homology)	-2.17	2.77	23.5

Вывод

Полученные результаты указывают на то, что использование unsupervised-метрик качества эмбеддингов в качестве замены валидационных оценок в формуле (2.1) не приводит к улучшению, и не может служить полноценной альтернативой. В частности, RankMe, Coherence и Persistent Homology, несмотря на интерпретируемость, не обладают достаточной связью с downstream-качеством модели, что согласуется с результатами предыдущей секции о корреляции метрик и улучшения от объединения.

ЗАКЛЮЧЕНИЕ

В работе предложены и экспериментально исследованы подходы к автоматическому построению ансамбля представлений на основе метрик сходства в двух кейсах: объединение слоев трансформеров и различные представления клиентов. Основные выводы следующие:

1. Объединение слоёв трансформеров. Предложены и реализованы стратегии агрегирования слоёв: Best Layer, QP Weighted, QP + PCA, Cluster + PCA, Greedy, QP masked. Эксперименты на шести задачах показали:

   - Все методы заметно превосходят наивную базовую стратегию — усреднение последнего слоя. Среднее качество последнего слоя отстаёт от лучшего метода (QP) почти на 8 процентных пунктов.
   - Метод QP Weighted показывает лучшие результаты как по средней метрике, так и по среднему рангу. Он доказал устойчивую эффективность и может рассматриваться как универсальный базовый метод для прикладного применения.
   - Жадная стратегия (Greedy) и её модификация с маскированием (QP masked) демонстрируют сопоставимое качество, но требуют более тонкой настройки: Greedy чувствителен к стратегии останова, а QP masked — к механизму обнуления весов в матрице сходства.
   - Методы, использующие конкатенацию слоёв с последующей PCA-проекцией (QP + PCA и Cluster + PCA), менее универсальны, но эффективны на задачах с выраженной структурной неоднородностью слоёв или в объединении представлений

различной природы, например, CLS-токенов и усредненных эмбеддингов слоя.

- Стратегия выбора одного лучшего слоя (Best Layer) показывает умеренное качество, а наихудшие результаты демонстрирует Last Layer — это подтверждает ограниченность тривиальных подходов к агрегации.

2. Транзакционные эмбеддинги клиентов. На задаче агрегации представлений из 102 трансформеров, обученных на пользовательских последовательностях, установлено:

   - Простое усреднение работает плохо — из-за несогласованности пространств. Аналогично, взвешивание без предварительного выравнивания даёт нестабильный результат.
   - Жадный отбор слоёв с учётом качества и разнообразия демонстрирует выигрыш в 88% запусков — без увеличения размерности итогового эмбеддинга.
   - Метрики различия (CKA, DistCorr) обратно коррелируют с эффектом от ансамблирования, что подтверждает гипотезу: чем сильнее различие между представлениями, тем выше возможный выигрыш при их объединении.

Перспективы развития:

- обучение преобразований, минимизирующих зависимость между слоями (например, по CKA), при сохранении task-loss;
- внедрение предварительного нелинейного выравнивания пространств (например, через автоэнкодеры);
- адаптивный выбор параметра λ и критерия останова в Greedy на основе валидационной кривой;
- расширение подхода на мультимодальные данные (текст, графы, транзакции).

Таким образом, работа демонстрирует, что метрики структурного сходства между представлениями могут служить надёжным сигналом для автоматического ансамблирования. Независимо от природы источников (слои одной модели или разнородные модели), анализ взаимной информативности и разнообразия позволяет формировать более выразительные и устойчивые представления без существенного увеличения размерности или переобучения. Предложенные методы обладают высокой переносимостью, не требуют специального дообучения и могут быть применены в широком спектре прикладных задач.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kornblith S., Norouzi M., Lee H., Hinton G. Similarity of neural network representations revisited // Proceedings of the International Conference on Machine Learning. – 2019. – P. 3519–3529.

2. Gretton A., Bousquet O., Smola A., Schölkopf B. Measuring statistical dependence with Hilbert-Schmidt norms // Proceedings of the International Conference on Algorithmic Learning Theory. – Springer, 2005. – P. 63–77.

3. Hotelling H. Relations between two sets of variates // Biometrika. – 1936. – Vol. 28, no. 3/4. – P. 321–377.

4. Kriegeskorte N., Mur M., Bandettini P. A. Representational similarity analysis – connecting the branches of systems neuroscience // Frontiers in Systems Neuroscience. – 2008.

5. Kriegeskorte N., Kievit R. A. Representational geometry: integrating cognition, computation, and the brain // Trends in Cognitive Sciences. – 2013.

6. Diedrichsen J., Provost S., Zareamoghaddam H. On the distribution of cross-validated Mahalanobis distances // arXiv preprint arXiv:1108.4126. – 2011.

7. Walther A., Nili H., Ejaz N., Alink A., Kriegeskorte N., Diedrichsen J. Reliability of dissimilarity measures for multi-voxel pattern analysis // NeuroImage. – 2016. – Vol. 137. – P. 188–200.

8. Mantel N. The detection of disease clustering and a generalized regression approach // Cancer Research. – 1967. – Vol. 27, no. 2, part 1. – P. 209–220.

9. Barannikov S., Trofimov I., Balabin N., Burnaev E. Representation topology divergence: A method for comparing neural network representations // Proceedings of the International Conference on Machine Learning. – PMLR, 2022. – P. 1607–1626.

10. Barannikov S., Koryakin D., Popov S., Struminsky K. Representation Topology Divergence: A Topological Measure of Representation Dissimilarity // Proceedings of the 39th International Conference on Machine Learning. – 2022.

11. Edelsbrunner H., Harer J. Persistent homology – a survey // Contemporary Mathematics. – 2008. – Vol. 453. – P. 257–282.

12. Raghu M., Gilmer J., Yosinski J., Sohl-Dickstein J. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability // Advances in Neural Information Processing Systems. – 2017.

13. Morcos A. S., Raghu M., Bengio S. Insights on representational similarity in neural networks with canonical correlation // Advances in Neural Information Processing Systems. – 2018.

14. Jaccard P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura // Bulletin de la Société Vaudoise des Sciences Naturelles. – 1901. – Vol. 37. – P. 547–579.

15. Muennighoff N., Tazi A., Magne L., Karamcheti S., Fan A., Ainslie J., Riedel S. MTEB: Massive Text Embedding Benchmark // arXiv preprint arXiv:2210.07316. – 2022.

16. Burnaev E., Struminsky K., Popov S., et al. pytorch-lifestream: An open framework for sequential behavioral modeling [Электронный ресурс]. – URL: https://github.com/sberbank-ai-lab/pytorch-lifestream (дата обращения: 27.05.2025).

17. Proskura P., Zaytsev A. Beyond Simple Averaging: Improving NLP Ensemble Performance with Topological-Data-Analysis-Based Weighting // arXiv preprint arXiv:2402.14184. – 2024.

18. Ding G., Shi Y., Gao J., Gimpel K. Understanding and Evaluating Representational Quality Without Labels // Advances in Neural Information Processing Systems. – 2022.