

doi: 10.17586/2226-1494-2026-26-3-544-555

УДК 004.852

Принцип максимальной энтропии в глубоком обучении: от теории к практическому регуляризатору для сложных данных

Андрей Владимирович Тимофеев✉

ООО «Центр ИИ для Стран ШОС+», Санкт-Петербург, 197341, Российская Федерация
 timofeev.andrey@gmail.com✉, tav@centeraiso.com, <https://orcid.org/0000-0001-7212-5230>

Аннотация

Введение. Представлена и экспериментально проверена гипотеза о том, что в машинном обучении максимизация энтропии распределения эмбедингов при заданных ограничениях ведет к улучшению обобщающей способности моделей. **Метод.** Для реализации гипотезы предложен новый метод — регуляризация методом максимизации энтропии на гиперсфере с разделением классов, предназначенный для задач классификации с высокой зашумленностью и дисбалансом классов (например, распознавание сигналов Orthogonal Frequency Division Multiplexing (OFDM)). Метод основан на расширении функции потерь классической модели за счет трех специализированных компонент: первый максимизирует энтропию через равномерность распределения на гиперсфере, второй контролирует ковариацию признаков и третий обеспечивает межклассовую разделимость. Динамическая регуляризация позволяет адаптивно балансировать влияние энтропийного члена в процессе обучения. **Основные результаты.** Показана связь между максимизацией энтропии распределения эмбедингов и сужением границ обобщения модели. Экспериментально на задаче классификации OFDM-сигналов продемонстрировано, что предложенный метод обеспечивает формирование высокоравномерного пространства признаков: среднее косинусное сходство снижается до 0,064, покрытие достигает 1,71, энтропия распределения сходств — 10,34 нат при сохранении отличной дискриминативности, что подтверждается точностью 0,97 и сбалансированным Macro F1-score 0,98. Метод эффективно устраняет переобучение и повышает устойчивость модели на миноритарных классах. **Обсуждение.** Разработанный подход, явно максимизирующий энтропию представлений при заданных ограничениях, предоставляет перспективный принцип для улучшения обобщающей способности в условиях недостаточности или зашумленности данных. Перспективы применения метода включают ослабление теоретических предположений и применение в смежных областях, где надежность классификации критически зависит от качества представлений.

Ключевые слова

эмбединг, машинное обучение, обобщающая способность, OFDM, энтропия, регуляризация

Ссылка для цитирования: Тимофеев А.В. Принцип максимальной энтропии в глубоком обучении: от теории к практическому регуляризатору для сложных данных // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 3. С. 544–555. doi: 10.17586/2226-1494-2026-26-3-544-555

Maximum entropy principle in deep learning: from theory to a practical regularizer for complex data

Andrey V. Timofeev✉

LLC “AI Center for SCO+ Countries”, Saint Petersburg, 197341, Russian Federation
 timofeev.andrey@gmail.com✉, tav@centeraiso.com, <https://orcid.org/0000-0001-7212-5230>

Abstract

We hypothesize and experimentally validate that maximizing the entropy of embedding distributions under specific constraints enhances the generalization ability of machine learning models. To test this hypothesis, we propose a novel regularization method — Hyperspherical Entropy Maximization with Class Separation — designed for noisy and imbalanced classification tasks (e.g., Orthogonal Frequency Division Multiplexing (OFDM) signal recognition).

The method is based on extending the loss function of the classical model with three specialized components: the first maximizes entropy by enforcing uniformity of distribution on the hypersphere, the second controls the covariance of features, and the third ensures inter-class separability. Dynamic regularization allows adaptively balancing the influence of the entropy term during training. We establish a link between maximizing the entropy of the embedding distribution and a reduction in the model generalization error bound. In experiments on an OFDM signal classification task, our method shapes a highly uniform feature space: mean cosine similarity decreases to 0.064, coverage reaches 1.71, and the entropy of the similarity distribution attains 10.34 nats, while preserving strong discriminative power, achieving 0.97 accuracy and a Macro F1-score of 0.98, outperforming standard regularization techniques. The approach effectively mitigates overfitting and improves robustness on minority classes. Our approach, which explicitly maximizes representation entropy under task-specific constraints, establishes a promising principle for enhancing generalization in conditions of data scarcity or high noise. Future work will focus on relaxing theoretical assumptions and applying the method to other domains where classification reliability critically depends on representation quality.

Keywords

embeddings, machine learning, generalization, OFDM, entropy, regularization

For citation: Timofeev A.V. Maximum entropy principle in deep learning: from theory to a practical regularizer for complex data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 3, pp. 544–555 (in Russian). doi: 10.17586/2226-1494-2026-26-3-544-555

Введение

В современном машинном обучении ключевой задачей остается повышение обобщающей способности моделей, особенно в условиях ограниченных, зашумленных или несбалансированных данных. Одним из перспективных направлений является анализ и целенаправленное конструирование пространств эмбедингов — сжатых векторных представлений данных, от качества которых напрямую зависит успех последующего обучения классификаторов или регрессоров.

Исследованию свойств эмбедингов и их влияния на обобщение посвящено значительное число научных работ. В работе [1] изложены фундаментальные основы обучения представлений, где подчеркивалась важность выделения инвариантных и дизангвированных факторов вариативности. В области контрастивного обучения (метод Simple Framework for Contrastive Learning of Visual Representations (SimCLR)) [2] продемонстрировано, что минимизация сходства между представлениями различных объектов (негативных пар) при одновременном притяжении аугментированных копий одного объекта (позитивных пар) позволяет получать высококачественные эмбединги без использования меток. В работе [3] был развит данный подход, предложены динамический словарь и очередь негативных примеров для масштабирования на большие объемы данных.

Ключевым теоретическим прорывом стала декомпозиция качества контрастивных представлений на две компоненты — выравнивание позитивных пар и равномерность распределения на гиперсфере, предложенная в работе [4]. В работе показано, что равномерность тесно связана с максимизацией энтропии распределения эмбедингов при фиксированной норме. Этот тезис получил дальнейшее развитие в [5], где был предложен эффективный критерий максимизации энтропии (E^2MC) на основе низкоразмерных статистик, позволяющий дообучать уже готовые модели Secure Sockets Layer с заметным улучшением downstream-задач. В [5] отмечено, что прямая оценка энтропии в пространствах высокой размерности затруднена, однако необходимые условия (максимальная энтропия (Maximum Entropy,

MaxEnt) маргинальных распределений и декорреляция признаков) могут быть эффективно реализованы на практике.

С геометрической точки зрения, равномерное распределение точек на гиперсфере исследовалось в контексте моделирования ориентаций и случайных вращений [6]. В машинном обучении этот принцип нашел отражение в методах, использующих гиперсферические эмбединги для повышения робастности, например, в состязательном обучении [7]. В работе [8] предложено рассматривать процесс понимания и представления данных через минимизацию «репрезентационной энтропии», что согласуется с идеями настоящей работы.

Теоретические гарантии обобщения для методов, основанных на равномерности распределения, могут быть получены с использованием радиемахеровской сложности. В работе [9] применен такой подход для анализа графовых нейронных сетей, показав, что более равномерное распределение представлений ведет к более узким границам обобщения. Отметим, что в прикладной области классификации сигналов, которая лежит в основе эксперимента в настоящей работе, задача автоматического распознавания модуляции Orthogonal Frequency Division Multiplexing (OFDM) остается актуальной, особенно в условиях аппаратных искажений и низкого отношения сигнал/шум [10].

Настоящая работа базируется и развивает перечисленные идеи. В работе не только неявно использованы равномерности через контрастивные механизмы, но и введена явная регуляризация, максимизирующая энтропию распределения эмбедингов при заданных структурных ограничениях (ковариация, межклассовая разделимость). Предложена и исследована фундаментальная гипотеза (H1) о том, что максимизация энтропии распределения таких эмбедингов при заданных структурных ограничениях ведет к сужению границ обобщения и, как следствие, к улучшению итоговой производительности модели. Этот принцип, основанный на теории информации, формализует интуитивное представление о том, что наиболее «беспристрастное» (максимально энтропийное) распределение признаков минимизирует априорные предположения и способствует лучшему покрытию пространства возможных

данных. Для практической проверки гипотезы H1 предлагается новый метод регуляризации — регуляризация методом максимизации энтропии (MaxEnt) на гиперсфере с разделением классов (Hyperspherical Entropy Maximization with Class Separation (HEMS), MaxEnt). Метод явным образом дополняет стандартную функцию потерь компонентами, ответственными за равномерность распределения на сфере, контроль ковариации признаков и гарантированное расстояние между классами. Теоретическая часть работы содержит доказательство ключевой теоремы, связывающей высокую энтропию распределения эмбеддингов с уменьшением радиемахеровской сложности и улучшением гарантий обобщения. Экспериментальная проверка выполнена на сложной практической задаче классификации зашумленных OFDM-сигналов с выраженным дисбалансом классов, где предложенный метод продемонстрировал формирование высокоравномерного и хорошо разделимого пространства признаков, что привело к улучшению всех ключевых метрик по сравнению с базовыми подходами.

Постановка задачи

Пусть: $\mathcal{X}$ — пространство объектов (например, тексты, изображения); $\mathcal{Z}$ — пространство эмбеддингов, где $\mathcal{Z} \subseteq R^d$ (здесь R^d — евклидово пространство размерности d) и $\mathcal{Z}$ компактно (например, единичная сфера S^{d-1}); $f_\theta: \mathcal{X} \rightarrow \mathcal{Z}$ — модель эмбеддингов с параметрами θ ; p_{data} — исходное распределение данных на $\mathcal{X}$; q_θ — индуцированное распределение эмбеддингов на $\mathcal{Z}$, полученное как $z = f_\theta(x) + \epsilon$, где $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$, $\mathcal{N}(0, \sigma^2 I)$ — плотность многомерного нормального распределения с нулевым средним и ковариацией $\sigma^2 I$.

Чтобы гарантировать существование дифференциальной энтропии (которая определена только для абсолютно непрерывных распределений), рассмотрим сглаженное распределение

$$q_\theta(z) = \int_{\mathcal{X}} p_{data}(x) \mathcal{N}(z - f_\theta(x); \sigma^2 I) dx.$$

Добавление малого гауссовского шума делает распределение абсолютно непрерывным относительно меры Лебега на R^d , кроме того, при $\sigma \rightarrow 0$ оно стремится к исходному (возможно, сингулярному) распределению эмбеддингов. Это позволяет использовать классическое определение дифференциальной энтропии распределения q_θ :

$$H[q_\theta] = - \int_{\mathcal{Z}} q_\theta(z) \ln q_\theta(z) dz.$$

Все дальнейшие теоретические рассуждения опираются именно на это сглаженное распределение; при достаточно малом σ полученные границы обобщения остаются релевантными для исходной модели (без шума).

Гипотеза H1. Среди всех θ модель f_θ , максимизирующая $H[q_\theta]$ при заданных ограничениях (например, фиксированной ковариации), приводит к лучшей обобщающей способности для моделей, обучаемых на этих эмбеддингах.

Вспомогательные леммы

Лемма 1. MaxEnt как минимальное предположение. Пусть: $\mathcal{Z}$ — компактное подмножество R^d с конечной мерой Лебега $|\mathcal{Z}|$. $\mathcal{C} = \{q: \mathcal{Z} \rightarrow R_{\geq 0} \mid \int_{\mathcal{Z}} q(z) dz = 1, \mathbb{E}_q[g_i(z)] = c_i, i = 1, \dots, m\}$, где $g_i: \mathcal{Z} \rightarrow R$ — заданные функции, определяющие моментные ограничения (например, для фиксации ковариации используются $g_{ij}(z) = z_i z_j$ при условии нулевого среднего); $R_{\geq 0}$ — множество неотрицательных вещественных чисел, т. е. множество плотностей распределений, удовлетворяющих заданным ограничениям на моменты.

Тогда распределение q^* , имеющее максимальную дифференциальную энтропию

$$H[q] = - \int_{\mathcal{Z}} q(z) \ln q(z) dz$$

в классе $\mathcal{C}$, существует, является единственным и минимизирует расхождение Кульбака–Лейблера (Kullback–Leibler (KL) дивергенция) $D_{KL}(q||u)$ относительно равномерного распределения $u(z) = 1/|\mathcal{Z}|$ на $\mathcal{Z}$.

Доказательство. Докажем существование решения. Класс $\mathcal{C}$ является выпуклым замкнутым подмножеством $L^1(\mathcal{Z})$, ограниченным по норме. По теореме Данфорда–Петтиса [11] он слабо компактен.

Функционал энтропии $H[q]$ строго вогнут и полунепрерывен сверху относительно слабой сходимости. Действительно, для $q_n \rightarrow q$ слабо в L^1 , тогда из неравенства Йенсена имеем:

$$\limsup_{n \rightarrow \infty} H[q_n] \leq H[q].$$

По теореме Вейерштрасса [12] о достижении максимума полунепрерывного сверху функционала на компакте существует $q^* \in \mathcal{C}$, на котором достигается максимум:

$$H[q^*] = \max_{q \in \mathcal{C}} H[q].$$

Рассмотрим задачу максимизации $H[q]$ при ограничениях. Введем функционал Лагранжа:

$$\mathcal{L}[q] = -H[q] + \lambda_0 \left(\int_{\mathcal{Z}} q(z) dz - 1 \right) + \sum_{i=1}^m \lambda_i \left(\int_{\mathcal{Z}} g_i(z) q(z) dz - c_i \right),$$

где λ_i — множители Лагранжа, соответствующие ограничениям $\mathbb{E}_q[g_i(z)] = c_i$.

Получим вариационную производную:

$$\frac{\delta \mathcal{L}}{\delta q(z)} = \ln q(z) + 1 + \lambda_0 + \sum_{i=1}^m \lambda_i g_i(z).$$

Приравнявая вариационную производную к нулю, получаем:

$$q^*(z) = \exp \left(-1 - \lambda_0 - \sum_{i=1}^m \lambda_i g_i(z) \right) = \frac{1}{Z} \exp \left(- \sum_{i=1}^m \lambda_i g_i(z) \right),$$

$$\text{где } Z = \exp(1 + \lambda_0) = \int_{\mathcal{Z}} \exp \left(- \sum_{i=1}^m \lambda_i g_i(z) \right) dz.$$

Докажем единственность решения. Пусть существуют два различных $q_1^*, q_2^* \in \mathcal{C}$, максимизирующих энтропию. Рассмотрим $q' = \frac{1}{2}q_1^* + \frac{1}{2}q_2^*$. По строгой вогнутости энтропии:

$$H[q'] > \frac{1}{2}H[q_1^*] + \frac{1}{2}H[q_2^*] = H_{\max},$$

что противоречит максимальности $H_{\max}$.

Таким образом, решение единственное.

Докажем связь с KL-дивергенцией. Для этого вычислим:

$$D_{KL}(q||u) = \int_{\mathcal{Z}} q(z) \ln \frac{q(z)}{1/|\mathcal{Z}|} dz = -H[q] + \ln|\mathcal{Z}|.$$

Поскольку $\ln|\mathcal{Z}|$ — константа, минимизация $D_{KL}(q||u)$ эквивалентна максимизации $H[q]$ и поэтому максимизация энтропии минимизирует KL-дивергенцию. Таким образом, доказаны все утверждения леммы. ■

Замечание 1. Лемма 1 устанавливает, что MaxEnt-распределение является наиболее «беспристрастным» — оно минимально отклоняется от равномерного распределения в смысле KL-дивергенции.

Лемма 2. Концентрация меры для распределений с фиксированными моментами. Пусть:

Л2.1. $\mathcal{Z} \subset \mathbb{R}^d$ — выпуклое компактное множество, симметричное относительно нуля;

Л2.2. q_1, q_2 — два распределения на $\mathcal{Z}$ с нулевым средним и одинаковой ковариацией Σ ;

Л2.3. $H[q_1] > H[q_2]$, причем q_1 — MaxEnt-распределение в классе распределений с нулевым средним и ковариацией Σ на $\mathcal{Z}$.

Тогда для любого выпуклого симметричного множества $A \subset \mathcal{Z}$:

$$P_{\mathbf{z}-q_1}(\mathbf{z} \in A) \leq P_{\mathbf{z}-q_2}(\mathbf{z} \in A).$$

Здесь запись $P_{\mathbf{z}-q_1}(\mathbf{z} \in A)$ обозначает вероятность того, что случайный вектор $\mathbf{z}$, распределенный согласно распределению q_1 , принадлежит множеству A . Иными словами, это вероятность события $\{\mathbf{z} \in A\}$ при условии, что $\mathbf{z}$ подчиняется закону распределения q_1 . Аналогично, $P_{\mathbf{z}-q_2}(\mathbf{z} \in A)$ — вероятность того же события для распределения q_2 .

В частности, для любого шара $B(0, r) = \{\mathbf{z} \in \mathcal{Z} : \|\mathbf{z}\| \leq r\}$ получим:

$$P_{\mathbf{z}-q_1}(\|\mathbf{z}\| \leq r) \leq P_{\mathbf{z}-q_2}(\|\mathbf{z}\| \leq r).$$

Доказательство. MaxEnt-распределение при фиксированных первых двух моментах является гауссовским (или усеченным гауссовским на $\mathcal{Z}$) [13]. Для $\mathbb{R}^d$ без ограничений это многомерное нормальное распределение $\mathcal{N}(0, \Sigma)$.

По теореме Андерсона [13], для любого выпуклого симметричного множества $A \subset \mathbb{R}^d$ и любого центрального распределения q с ковариацией Σ :

$$P_{\mathbf{z}-\mathcal{N}(0, \Sigma)}(\mathbf{z} \in A) \leq P_{\mathbf{z}-q}(\mathbf{z} \in A),$$

т. е. гауссовское распределение минимизирует вероятность попадания в любое выпуклое симметрич-

ное множество при фиксированных первых двух моментах.

Поскольку q_1 совпадает с $\mathcal{N}(0, \Sigma)$ на $\mathbb{R}^d$ или является его усечением на $\mathcal{Z}$. В любом случае для усеченного гауссовского распределения свойство минимальной концентрации сохраняется [14]. Следовательно, для любого выпуклого симметричного A имеем:

$$P_{\mathbf{z}-q_1}(\mathbf{z} \in A) \leq P_{\mathbf{z}-q_2}(\mathbf{z} \in A).$$

Так как шар $B(0, r)$ выпуклый и симметричный, утверждение также выполняется. ■

Лемма 2 показывает, что MaxEnt-распределения минимизируют концентрацию меры в любых симметричных выпуклых множествах (следствие теоремы Андерсона).

Лемма 3. Вероятность образования «пустот» в выборке. Пусть выполнены условия Л2.2. Рассмотрим выборку $S_n = \{\mathbf{z}_i\}_{i=1}^n \sim q^n$ из распределения q . Тогда для любого $r > 0$:

$$\begin{aligned} P_{S_n \sim q_1^n}(\exists \mathbf{z} \in \mathcal{Z} : B(\mathbf{z}, r) \cap S_n = \emptyset) &\leq \\ &\leq P_{S_n \sim q_2^n}(\exists \mathbf{z} \in \mathcal{Z} : B(\mathbf{z}, r) \cap S_n = \emptyset), \end{aligned}$$

где $B(\mathbf{z}, r)$ — шар радиуса r с центром в $\mathbf{z}$. Здесь и далее символ $\emptyset$ обозначает пустое множество.

Доказательство. Для фиксированного $\mathbf{z}$ получим:

$$P(B(\mathbf{z}, r) \cap S_n = \emptyset) = \left(1 - \int_{B(\mathbf{z}, r)} q(s) ds\right)^n.$$

По свойству сдвига и симметрии распределений (все распределения имеют нулевое среднее и одинаковую ковариацию) для любого $\mathbf{z}$:

$$\int_{B(\mathbf{z}, r)} q(s) ds \geq \int_{B(0, r)} q(s) ds,$$

поскольку плотность q убывает по мере удаления от нуля для MaxEnt-распределений с и им подобных. Из Л2.2. следует, что:

$$\int_{B(0, r)} q_1(s) ds \geq \int_{B(0, r)} q_2(s) ds,$$

так как интеграл по шару равен вероятности попадания в шар.

Следовательно, для любого $\mathbf{z}$:

$$\int_{B(\mathbf{z}, r)} q_1(s) ds \geq \int_{B(\mathbf{z}, r)} q_2(s) ds$$

и значит:

$$\left(1 - \int_{B(\mathbf{z}, r)} q_1(s) ds\right)^n \leq \left(1 - \int_{B(\mathbf{z}, r)} q_2(s) ds\right)^n. \quad (1)$$

Оценим вероятность существования пустого шара. Рассмотрим $r/2$ -сеть $\mathcal{N}$ пространства $\mathcal{Z}$, т. е. множество точек $\{\mathbf{z}_j\}_{j=1}^M$ такое, что $\mathcal{Z} \subseteq \bigcup_{j=1}^M B(\mathbf{z}_j, r/2)$. Если существует такой $\mathbf{z} \in \mathcal{Z}$, что $B(\mathbf{z}, r)$ не содержит точек выборки, то $B(\mathbf{z}_j, r/2)$ для ближайшего $\mathbf{z}_j \in \mathcal{N}$ также не содержит точек выборки.

Таким образом:

$$\begin{aligned} P(\exists \mathbf{z} \in \mathcal{Z}: B(\mathbf{z}, r) \cap S_N = \emptyset) &\leq \\ &\leq \sum_{j=1}^M P(B(\mathbf{z}_j, r/2) \cap S_N = \emptyset). \end{aligned} \quad (2)$$

Из выражений (1) и (2) имеем:

$$\begin{aligned} \sum_{j=1}^M P_{S_n \sim q_1^n} \left(B\left(\mathbf{z}_j, \frac{r}{2}\right) \cap S_n = \emptyset \right) &\leq \\ &\leq \sum_{j=1}^M P_{S_n \sim q_2^n} \left(B\left(\mathbf{z}_j, \frac{r}{2}\right) \cap S_n = \emptyset \right), \end{aligned}$$

что завершает доказательство. ■

Замечание 2. Лемма 3 доказывает, что для MaxEnt-распределений вероятность образования больших «пустот» в пространстве эмбедингов меньше, что улучшает покрытие данных.

Лемма 4. Радиемахеровская сложность для липшицевых функций. Пусть:

1. $\mathcal{H} = \{h: \mathcal{Z} \rightarrow [-M, M]\}$ — класс L -липшицевых функций, т. е. L — константа Липшица для функций класса $\mathcal{H}$;

2. q_1, q_2 — два распределения на $\mathcal{Z}$ с нулевым средним и одинаковой ковариацией Σ .

3. $H[q_1] > H[q_2]$.

Тогда:

$$\mathbb{E}_{S \sim q_1^n} [\hat{\mathfrak{N}}_S(\mathcal{H})] \leq \mathbb{E}_{S \sim q_2^n} [\hat{\mathfrak{N}}_S(\mathcal{H})],$$

где $\hat{\mathfrak{N}}_S(\mathcal{H})$ — эмпирическая радиемахеровская сложность.

Доказательство. Оценим радиемахеровскую сложность через нормы данных. Для класса липшицевых функций известна оценка [15]:

$$\hat{\mathfrak{N}}_S(\mathcal{H}) \leq \frac{2L}{n} \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\| \right],$$

где σ_i — независимые случайные величины с равномерным распределением на $\{-1, 1\}$.

Оценим математическое ожидание функции Радемахера. Используя неравенство Йенсена:

$$\begin{aligned} \mathbb{E}_S [\hat{\mathfrak{N}}_S(\mathcal{H})] &\leq \frac{2L}{n} \mathbb{E}_S \mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\| \right] \leq \\ &\leq \frac{2L}{n} \mathbb{E}_S \left[\left(\mathbb{E}_\sigma \left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\|^2 \right)^{1/2} \right]. \end{aligned}$$

Для любых векторов $\mathbf{z}_i \in \mathbb{R}^d$ и случайных знаков $\sigma_i \in \{-1, 1\}$ получим:

$$\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\|^2 = \left(\sum_{i=1}^n \sigma_i \mathbf{z}_i \right)^\top \left(\sum_{j=1}^n \sigma_j \mathbf{z}_j \right).$$

Так как квадрат евклидовой нормы вектора $\mathbf{v}$: $\|\mathbf{v}\|^2 = \mathbf{v}^\top \mathbf{v}$, где $\mathbf{v} = \sum_{i=1}^n \sigma_i \mathbf{z}_i$.

Скалярное произведение билинейно, поэтому:

$$\left(\sum_{i=1}^n \sigma_i \mathbf{z}_i \right)^\top \left(\sum_{j=1}^n \sigma_j \mathbf{z}_j \right) = \sum_{i=1}^n \sum_{j=1}^n \sigma_i \mathbf{z}_i^\top (\sigma_j \mathbf{z}_j).$$

Так как $(\sigma_i \mathbf{z}_i)^\top (\sigma_j \mathbf{z}_j) = \sigma_i \sigma_j (\mathbf{z}_i^\top \mathbf{z}_j)$, имеем:

$$\sum_{i=1}^n \sum_{j=1}^n (\sigma_i \mathbf{z}_i)^\top (\sigma_j \mathbf{z}_j) = \sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j (\mathbf{z}_i^\top \mathbf{z}_j).$$

Рассчитаем математическое ожидание по σ :

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\|^2 \right] = \mathbb{E}_\sigma \left[\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j (\mathbf{z}_i^\top \mathbf{z}_j) \right].$$

Линейность математического ожидания позволяет вынести сумму за знак математического ожидания, поэтому

$$\mathbb{E}_\sigma \left[\sum_{i=1}^n \sum_{j=1}^n \sigma_i \sigma_j (\mathbf{z}_i^\top \mathbf{z}_j) \right] = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}_\sigma [\sigma_i \sigma_j] (\mathbf{z}_i^\top \mathbf{z}_j).$$

Теперь проанализируем компонент $\mathbb{E}_\sigma [\sigma_i \sigma_j]$.

Если $i = j$, то

$$\mathbb{E}_\sigma [\sigma_i^2] = \mathbb{E}_\sigma [1] = 1,$$

так как $\sigma_i^2 = 1$.

При $i \neq j$: так как σ_i и σ_j независимы и $\mathbb{E}[\sigma_i] = 0$, то

$$\mathbb{E}_\sigma [\sigma_i \sigma_j] = \mathbb{E}_\sigma [\sigma_i] \mathbb{E}_\sigma [\sigma_j] = 0 \cdot 0 = 0.$$

В результате получим, что каждое произведение $\sigma_i \sigma_j$ принимает значения 1 и -1 с вероятностью 1/2, поэтому математическое ожидание равно 0. После расчетов в сумме $\sum_{i=1}^n \sum_{j=1}^n \dots$ останутся только члены с $i = j$. Тогда,

$$\mathbb{E}_\sigma \left[\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\|^2 \right] = \sum_{i=1}^n 1 (\mathbf{z}_i^\top \mathbf{z}_i) = \sum_{i=1}^n \|\mathbf{z}_i\|^2,$$

так как $\mathbf{z}_i^\top \mathbf{z}_i = \|\mathbf{z}_i\|^2$.

Таким образом:

$$\left\| \sum_{i=1}^n \sigma_i \mathbf{z}_i \right\|^2 = \sum_{i=1}^n \|\mathbf{z}_i\|^2.$$

Следовательно:

$$\mathbb{E}_S [\hat{\mathfrak{N}}_S(\mathcal{H})] \leq \frac{2L}{n} \mathbb{E}_S \left[\sqrt{\sum_{i=1}^n \|\mathbf{z}_i\|^2} \right].$$

Используя неравенство Йенсена, имеем:

$$\mathbb{E}_S \left[\sqrt{\sum_{i=1}^n \|\mathbf{z}_i\|^2} \right] \leq \sqrt{\mathbb{E}_S \left[\sum_{i=1}^n \|\mathbf{z}_i\|^2 \right]} = \sqrt{n \mathbb{E}_{z \sim q} [\|\mathbf{z}\|^2]}.$$

В итоге получим:

$$\mathbb{E}_S [\hat{\mathfrak{N}}_S(\mathcal{H})] \leq 2L \sqrt{\frac{\mathbb{E}_{z \sim q} [\|\mathbf{z}\|^2]}{n}}.$$

Используя более точную оценку [16], имеем:

$$\mathbb{E}_S [\hat{\mathfrak{N}}_S(\mathcal{H})] \leq \frac{2L}{n} \mathbb{E}_{z \sim q} [\|\mathbf{z}\|].$$

По теореме Андерсона [13], для любого выпуклого четного функционала Φ , гауссовское распределение минимизирует $\mathbb{E}_{z \sim q} \Phi(z)$ среди всех распределений с фиксированной ковариацией.

В частности, для $\Phi(z) = \|\mathbf{z}\|$ (выпуклый и четный):

$$\mathbb{E}_{z \sim q_1} [\|\mathbf{z}\|] \leq \mathbb{E}_{z \sim q_2} [\|\mathbf{z}\|].$$

Следовательно:

$$\begin{aligned} \mathbb{E}_{S \sim q_1^n}[\hat{\mathcal{H}}_S(\mathcal{H})] &\leq \frac{2L}{\sqrt{n}} \mathbb{E}_{z \sim q_1}[\|z\|] \leq \frac{2L}{\sqrt{n}} \mathbb{E}_{z \sim q_2}[\|z\|] \leq \\ &\leq \mathbb{E}_{S \sim q_2^n}[\hat{\mathcal{H}}_S(\mathcal{H})]. \quad \blacksquare \end{aligned}$$

Замечание 3. Лемма 4 связывает высокую энтропию с меньшей радиемахеровской сложностью для классов липшицевых функций, что ведет к более узким доверительным интервалам обобщения.

Для компактности изложения введем следующие обозначения:

- математическое ожидание радиемахеровской сложности: $\mathcal{A}(\theta) := \mathbb{E}_{S \sim q_\theta^n}[\hat{\mathcal{H}}_S(\mathcal{H})]$;
- константа концентрации: для любого $\delta > 0$ и размера выборки n :

$$\mathcal{B}(\delta, n) = 3 \sqrt{\frac{\ln(2/\delta)}{2n}}.$$

Утверждение L1. Из леммы 3: вероятность образования «пустот» меньше для распределения с большей энтропией.

Утверждение L2. Из леммы 4: если $H[q_{\theta_1}] > H[q_{\theta_2}]$, то $\mathcal{A}(\theta_1) \leq \mathcal{A}(\theta_2)$.

Теорема 1. Обобщающая способность MaxEnt-эмбедингов. Пусть для распределения q_θ выполнены все условия лемм 1–4, и задана функция потерь $\ell: H \times Z \rightarrow R$, L_ℓ -липшицева по второму аргументу.

Рассмотрим два набора параметров θ_1 и θ_2 таких, что:

1. $H[q_{\theta_1}] > H[q_{\theta_2}]$;
2. распределения q_{θ_1} и q_{θ_2} имеют одинаковые первые два момента (нулевое среднее и ковариацию Σ).

Тогда существует константа $C > 0$, зависящая от d , L_h , L_ℓ , D , M и геометрических параметров Z , такая что для любого скаляра $\delta > 0$ и достаточно больших n , с вероятностью не менее $1 - \delta$ по независимым выборкам $S_1 \sim q_{\theta_1}^n$ и $S_2 \sim q_{\theta_2}^n$, выполняется неравенство:

$$\begin{aligned} \sup_{h \in \mathcal{H}} [R_{D_{\theta_1}}(h) - \hat{R}_{S_1}(h)] &\leq \sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] + \\ &+ C \sqrt{\frac{\ln(1/\delta)}{n}}, \end{aligned}$$

где d — размерность пространства эмбедингов ($Z \subseteq R^d$); L_h — константа Липшица для функций класса $\mathcal{H}$ (указана в лемме 4); L_ℓ — константа Липшица для функции потерь ℓ по второму аргументу; D — диаметр множества Z (максимальное расстояние между точками); M — верхняя граница абсолютного значения функций из $\mathcal{H}$ (в условии леммы 4: $h: Z \rightarrow [-M, M]$).

Доказательство. Из теории статистического обучения [15, Theorem 3.1] следует то, что для любого распределения q_θ и любого $\delta > 0$, с вероятностью не менее $1 - \delta$ по выборке $S \sim q_\theta^n$, для всех $h \in \mathcal{H}$:

$$R_{D_\theta}(h) \leq \hat{R}_S(h) + 2\mathcal{A}(\theta) + \mathcal{B}(\delta, n).$$

По условию доказываемой теоремы, параметры θ_1 , θ_2 таковы, что $H[q_{\theta_1}] > H[q_{\theta_2}]$. Тогда для θ_1 , с вероятностью не менее $1 - \delta$, по выборке $S_1 \sim q_{\theta_1}^n$:

$$\sup_{h \in \mathcal{H}} [R_{D_{\theta_1}}(h) - \hat{R}_{S_1}(h)] \leq 2\mathcal{A}(\theta_1) + \mathcal{B}(\delta_1, n). \quad (3)$$

С другой стороны, для θ_2 с вероятностью не менее $1 - \delta_2$ по выборке $S_2 \sim q_{\theta_2}^n$:

$$\sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] \leq 2\mathcal{A}(\theta_2) + \mathcal{B}(\delta_2, n).$$

Из утверждения L2 следует, что $\mathcal{A}(\theta_1) \leq \mathcal{A}(\theta_2)$. Подставим утверждение в (3), получим:

$$\sup_{h \in \mathcal{H}} [R_{D_{\theta_1}}(h) - \hat{R}_{S_1}(h)] \leq 2\mathcal{A}(\theta_2) + \mathcal{B}(\delta_1, n). \quad (4)$$

Из свойств концентрации радиемахеровской сложности [16, Theorem 2.3] следует, что с вероятностью не менее $1 - \delta_3$ по выборке $S_2 \sim q_{\theta_2}^n$:

$$\mathcal{A}(\theta_2) \leq R_{S_2}(\mathcal{H}) + \sqrt{\frac{\ln(1/\delta_3)}{2n}}.$$

Отметим, что вспомогательный скаляр δ_3 используется для управления вероятностью события концентрации радиемахеровской сложности. Далее, с вероятностью не менее $1 - \delta_3$:

$$2\mathcal{A}(\theta_2) \leq R_{S_2}(\mathcal{H}) + 2 \sqrt{\frac{\ln(1/\delta_3)}{2n}}. \quad (5)$$

Из стандартной симметризации (например, [16, Lemma 3.2]) известно, что для любой выборки S_2 :

$$\begin{aligned} \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i \ell(h, z_i) \right] &\leq \sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] + \\ &+ \mathcal{B}(\delta_3, n). \end{aligned} \quad (6)$$

Напомним, что симметризация — метод оценки математического ожидания обобщающей ошибки, который использует введение второй независимой выборки и случайных знаков для перехода к радиемахеровской сложности.

Комбинируя выражения (5) и (6), получим, что с вероятностью не менее $1 - \delta_3$:

$$\begin{aligned} 2\mathcal{A}(\theta_2) &\leq \sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] + \mathcal{B}(\delta_3, n) + \\ &+ 2 \sqrt{\frac{\ln(1/\delta_3)}{2n}}. \end{aligned} \quad (7)$$

Подставим (7) в (4):

$$\begin{aligned} \sup_{h \in \mathcal{H}} [R_{D_{\theta_1}}(h) - \hat{R}_{S_1}(h)] &\leq \sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] + \\ &+ \mathcal{B}(\delta_1, n) + \mathcal{B}(\delta_3, n) + 2 \sqrt{\frac{\ln(1/\delta)}{n}}. \end{aligned} \quad (8)$$

Выберем $\delta_1 = \delta_3 = \delta/2$. Тогда по неравенству объединения (неравенство Буля), с вероятностью не менее $1 - \delta$ одновременно выполняются все необходимые условия для обеих выборок S_1 и S_2 . Подставим $\delta/2$ вместо δ_1 и δ_3 в (8):

$$\mathcal{B}(\delta/2, n) = 3 \sqrt{\frac{\ln(4/\delta)}{2n}} = 3 \sqrt{\frac{\ln 4 + \ln(1/\delta)}{2n}} \leq$$

$$\leq 3 \sqrt{\frac{\ln(1/\delta)}{n}} + 3 \sqrt{\frac{2\ln 2}{n}}.$$

Аналогично:

$$2 \sqrt{\frac{\ln(2/\delta)}{2n}} = \sqrt{\frac{2\ln(2/\delta)}{n}} \leq \sqrt{\frac{2\ln(1/\delta)}{n}} + \sqrt{\frac{2\ln 2}{n}}.$$

Объединяя все константы (стандартный метод «свертка констант»), получаем существование единой константы $C > 0$ такой, что:

$$\sup_{h \in \mathcal{H}} [R_{D_{\theta_1}}(h) - \hat{R}_{S_1}(h)] \leq \sup_{h \in \mathcal{H}} [R_{D_{\theta_2}}(h) - \hat{R}_{S_2}(h)] + C \sqrt{\frac{\ln(1/\delta)}{n}}, \quad (9)$$

где C зависит от d, L_h, L_ℓ, D . ■

Интерпретация и выводы по результатам доказательства теоремы 1

Неравенство (9) показывает, что максимальное отклонение между истинным и эмпирическим рисками контролируется меньшей (в смысле порядка по N) верхней границей для распределения с большей энтропией.

Таким образом, обосновано: **дисперсионная компонента обобщающей способности улучшается** при максимизации энтропии распределения эмбеддингов.

Выводы по другим доказываемым утверждениям:

- систематическая ошибка: лемма 1 устанавливает, что MaxEnt-распределение ближе к равномерному в смысле KL-дивергенции. Это минимизирует априорные предположения о данных, что в сочетании с покрытием пространства (см. лемма 3, посвященная вероятности обнаружения «пустот» и покрытия пространства эмбеддингов) способствует снижению систематической ошибки;
- риск наихудшего случая: леммы 2 и 3 (Утверждение L1) показывают, что MaxEnt-распределение обеспечивает *лучшее* покрытие пространства эмбеддингов и уменьшает вероятность образования больших «пустот», где модель может допускать катастрофические ошибки.

Таким образом, доказанная теорема 1 показывает: **дисперсионная компонента обобщающей способности улучшается** для MaxEnt-распределения. В сочетании с результатами лемм 1–3, которые показывают улучшение покрытия пространства признаков и снижение априорных предположений, получаем обоснование гипотезы H1: *максимизация энтропии распределения эмбеддингов способствует улучшению ключевых аспектов обобщающей способности*, включая контроль дисперсии, снижение систематической ошибки, связанной с неудачным представлением, и **уменьшение вероятности больших ошибок на тестовых примерах, удаленных от обучающей выборки**.

Следствия и выводы

Следствие 1. Практическая регуляризация. Добавление к функции потерь слагаемого, MaxEnt-распределения эмбеддингов:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{task}}(\theta) - \lambda H[q_\theta]$$

улучшает обобщающую способность моделей, обучаемых на этих эмбеддингах.

Следствие 2. Объяснение успеха контрастивно-го обучения. Методы контрастивного обучения неявно максимизируют равномерность распределения эмбеддингов на единичной сфере, что соответствует максимизации энтропии при фиксированной норме (фиксированном втором моменте). Действительно, методы контрастивного обучения, такие как SimCLR или Information Noise Contrastive Estimation (InfoNCE), которые эмпирически показали свою высокую эффективность в обучении качественных представлений, причем, эта эффективность может быть объяснена через фундаментальный принцип максимизации энтропии. В основе лежит метод использования L2-нормализации эмбеддингов, который проецирует эмбеддинги на поверхность единичной сферы, что, в свою очередь, фиксирует их норму (второй момент). В этих условиях, согласно принципу MaxEnt, наиболее «беспристрастным» распределением, делающим минимальное количество априорных предположений, является равномерное распределение на сфере. Контрастивное обучение неявно способствует достижению этого состояния через свой механизм работы: его функция потерь, например InfoNCE, явно притягивает эмбеддинги схожих объектов (позитивные пары), но, что критически важно, неявно и сильно отталкивает эмбеддинги всех остальных объектов (негативные примеры) в батче. Чтобы минимизировать потерю, модель должна минимизировать среднее попарное косинусное сходство каждого вектора со всеми остальными, и на компактном многообразии, таком как сфера, единственная устойчивая конфигурация, минимизирующая это попарное сходство для всей системы векторов, — это их максимально равномерное распределение по поверхности. Таким образом, механизм отталкивания негативов действует как сила, эффективно распределяющая точки по сфере, что является практически способом MaxEnt-распределения при заданном ограничении на норму. Эта идея согласуется как с математической логикой, так и с эмпирическими наблюдениями: визуализации успешно обученных эмбеддингов демонстрируют их хорошее разделение и покрытие сферы, а в теоретических работах выделяются два компонента качества представлений — выравнивание позитивных пар и равномерность распределения, причем последняя связывается с энтропией. Таким образом, успех контрастивного обучения может быть интерпретирован в рамках предложенной теории как результат неявной MaxEnt-распределения эмбеддингов на единичной сфере, что способствует улучшению их обобщающей способности.

Следствие 3. Оптимальность гауссовских эмбеддингов. При фиксированной ковариации гауссовское

распределение эмбеддингов является оптимальным для последующего обучения моделей, так как максимизирует энтропию [17, теорема 8.6.5, С. 254–256], и, следовательно, обобщающую способность.

Практический пример

С целью иллюстрации эффективности принципа MaxEnt в задачах машинного обучения, была рассмотрена задача классификации источников широкополосных OFDM-сигналов [18] в диапазоне 2–6 ГГц. В этом случае, при помощи группы методов обработки сигналов, спектры OFDM-радиосигналов сжимаются в вещественные вектора ограниченной длины N и формируются обучающие множества целевых сигналов (Class_ATL, Class_DJ, Class_F_ANLG) и класс фоновых сигналов (BackGround). Состав корпуса: класс 'Class_ATL' (код 0): 17 858 элементов, класс 'BackGround' (код 1): 60 986 элементов, класс 'Class_DJ' (код 2): 21 208 элементов, класс 'Class_F_ANLG' (код 3): 18 732 элементов.

Регуляризация методом MaxEnt на гиперсфере с разделением классов. Для практической реализации MaxEnt была использована комплексная система регуляризаций с акцентом на равномерное распределение эмбеддингов на гиперсфере. В этом случае общая функция потерь (HEMS, MaxEnt) имеет следующий вид:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{entropy}}(t)\mathcal{L}_{\text{uniformity}} + \lambda_{\text{diversity}}\mathcal{L}_{\text{diversity}} + \lambda_{\text{sep}}\mathcal{L}_{\text{separation}}. \quad (10)$$

Рассмотрим подробнее компоненты выражения (10).

$\mathcal{L}_{\text{CE}}$ — основная функция потерь — кросс-энтропия (Cross-Entropy, CE), которая отвечает за :

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n} \sum_{i=1}^n \ln \left(\frac{\exp(\mathbf{z}_i^T \mathbf{w}_{y_i} + b_{y_i})}{\sum_{c=1}^C \exp(\mathbf{z}_i^T \mathbf{w}_c + b_c)} \right), \quad (11)$$

где n — размер батча (количество примеров); C — число классов; $\mathbf{z}_i \in R^d$ — эмбеддинг i -го образца (нормированный); $\mathbf{w}_c \in R^d$, $b_c \in R$ — параметры линейного классификатора для класса c ; $y_i \in \{1, \dots, C\}$ — истинная метка i -го примера; T — символ транспонирования.

Выражение (11) представляет собой среднее по батчу от отрицательного логарифма вероятности истинного класса, получаемой после применения функции softmax к логитам. Эта стандартная форма CE используется во всех экспериментах.

$$\mathcal{L}_{\text{uniformity}} = \ln \left(\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} e^{2 \frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}} \right).$$

$\mathcal{L}_{\text{uniformity}}$ — компонент Hypersphere Uniformity Loss, который оценивает равномерность распределения на гиперсфере через косинусное сходство. Гиперпараметр t (температура) контролирует распределение, оптимизируя баланс между структурой и разнообразием эмбеддингов. $\mathbf{z}_k$ — векторы эмбеддингов k -го образца, $\|\mathbf{z}_k\|$ — норма вектора $\mathbf{z}_k$). Минимизация $\mathcal{L}_{\text{uniformity}}$ при-

водит к уменьшению среднего попарного косинусного сходства, т. е. способствует равномерному распределению векторов на единичной сфере. Это соответствует MaxEnt при фиксированной норме (второй момент фиксирован нормировкой).

$$\lambda_{\text{diversity}} = \underbrace{\frac{\sigma(\text{diag}(\Sigma))}{\mu(\text{diag}(\Sigma))}}_{\text{выравнивание дисперсий}} + \underbrace{\gamma \|\Sigma - I\|_F^2}_{\text{подавление корреляций}}. \quad (12)$$

$\lambda_{\text{diversity}}$ — компонент Feature Diversity Loss, который измеряет относительную вариативность (разнообразие) дисперсий признаков, чем меньше значение, тем более сбалансированы дисперсии — ни один признак не доминирует, что соответствует условию фиксированной ковариации в гипотезе H1; Σ — ковариационная матрица эмбеддингов; $\text{diag}(\Sigma)$ — ее диагональ (дисперсии каждого признака); σ — стандартное отклонение дисперсий; μ — среднее значение дисперсий; γ — весовой коэффициент. Первое слагаемое выражения (12) стимулирует равномерную дисперсию по всем признакам, второе — приближает корреляционную матрицу к единичной, уменьшая внедиагональные элементы.

$$\mathcal{L}_{\text{separation}} = \max(0, (1 - \min_{i \neq j} \|c_i - c_j\|_2)),$$

$\mathcal{L}_{\text{separation}}$ — компонент Class Separation Loss, который стимулирует увеличение расстояния между ближайшими центроидами классов, применяя пороговое ограничение в 1,0. Параметр регуляризации $\mathcal{L}_{\text{sep}}$ определяет вклад компонента разделения классов $\mathcal{L}_{\text{separation}}$ в значение общей функции потерь (10). Значение $\mathcal{L}_{\text{sep}}$ определяется методом подбора на валидационной выборке.

Центроиды вычисляются как скользящие средние:

$$c_k^{(t)} = \alpha c_k^{(t-1)} + (1 - \alpha) \frac{1}{|B_k|} \sum_{i \in B_k} \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|},$$

где B_k — множество образцов класса k в текущем батче, $\alpha = 0,99$. Это обеспечивает явное разделение кластеров разных классов в пространстве эмбеддингов, что согласуется с ограничениями на структуру. Иначе говоря, даже при равномерном распределении центры классов не слипаются.

Параметры регуляризации $\lambda_{\text{entropy}}(t)$ и $\lambda_{\text{diversity}}$ определяются в процессе обучения модели, причем, параметр $\lambda_{\text{entropy}}(t)$ динамически зависит от номера эпохи t :

$$\lambda_{\text{entropy}}(t) = \lambda_{\text{max}} \min \left(1, \frac{t}{T_{\text{warmup}}} \right),$$

где $T_{\text{warmup}} = 10$, $\lambda_{\text{max}} = 0,1$.

Выбор функции потерь (10) оптимален для демонстрации гипотезы H1, поскольку она непосредственно максимизирует энтропию распределения эмбеддингов через $\mathcal{L}_{\text{uniformity}}$, которая, минимизируя логарифм среднего экспоненцированного косинусного сходства, равномерно распределяет векторы на гиперсфере. Компоненты $\lambda_{\text{diversity}}$ (стабилизируя дисперсии признаков через ковариационную матрицу) и $\mathcal{L}_{\text{separation}}$ (уве-

личивая минимальное расстояние между классовыми центроидами) формируют необходимые ограничения на ковариацию и структуру распределения.

Динамическая регуляризация $\lambda_{\text{entropy}}(t)$ обеспечивает адаптивный баланс, плавно усиливая влияние энтропийного члена после начальной фазы обучения, что в совокупности реализует осмысленную MaxEnt при заданных ограничениях в полном соответствии с гипотезой H1. С другой стороны, метод HEMS MaxEnt, основанный на функции потерь (10), особенно эффективен для задачи классификации OFDM-сигналов, где данные характеризуются высокой зашумленностью, значительным дисбалансом классов и близостью спектральных характеристик между некоторыми категориями. Компонент $\mathcal{L}_{\text{uniformity}}$ предотвращает излишнюю концентрацию эмбеддингов и способствует лучшему использованию пространства признаков, что критично для меньших классов; $\mathcal{L}_{\text{separation}}$ гарантирует надежную разделимость классов даже при равномерном распределении; а $\lambda_{\text{diversity}}$ обеспечивает активное использование всех измерений эмбеддингов, повышая устойчивость представлений и обобщающую способность модели.

Для оценки равномерности распределения эмбеддингов на единичной гиперсфере использовано среднее косинусное сходство (Mean Cosine Similarity (MeanCosSim)) между различными векторами (после L2-нормализации):

$$\text{MeanCosSim} = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{\mathbf{z}_i^T \mathbf{z}_j}{\|\mathbf{z}_i\| \|\mathbf{z}_j\|}.$$

Для равномерно распределенных на сфере векторов n -мерности MeanCosSim стремится к нулю; следовательно, меньшие значения MeanCosSim соответствуют более равномерному распределению.

Выполним расчет следующих параметров:

- покрытие (coverage) — максимальное евклидово расстояние между любыми двумя эмбеддингами (для нормализованных векторов максимально возможное значение равно 2);
- diversity — эффективная размерность, вычисляемая через собственные значения ковариационной матрицы:

$$\text{diversity} = \frac{\left(\sum_{i=1}^d \lambda_i \right)^2}{\sum_{i=1}^d \lambda_i^2},$$

высокое значение diversity означает, что модель использует все доступное пространство признаков, что соответствует MaxEnt при ограниченной ковариации;

- энтропия распределения сходств вычисляется на основе матрицы косинусных сходств между нормированными векторами:

$$\text{Entropy} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n p_{ij} \ln p_{ij},$$

$$\text{где } p_{ij} = \frac{\exp(\cos_sim(\mathbf{z}_i, \mathbf{z}_j))}{\sum_{k=1}^n \exp(\cos_sim(\mathbf{z}_i, \mathbf{z}_k))},$$

где $\cos_sim(\mathbf{z}_i, \mathbf{z}_j) = (\mathbf{z}_i^T \mathbf{z}_j) / (\|\mathbf{z}_i\| \|\mathbf{z}_j\|)$ — косинусное сходство между эмбеддингами $\mathbf{z}_i$ и $\mathbf{z}_j$; p_{ij} — вероятность того, что вектор $\mathbf{z}_j$ является «соседом» для $\mathbf{z}_i$ после применения функции softmax. Величина p_{ij} является эмпирической характеристикой равномерности распределения эмбеддингов на гиперсфере и не тождественна дифференциальной энтропии, используемой в теоретическом анализе. Энтропия усредняется по всем i и достигает максимума, когда распределение сходств равномерно (все p_{ij} одинаковы). Максимизация метрики *Entropy* является непосредственной целью рассматриваемого метода, так как равномерное распределение эмбеддингов на гиперсфере соответствует MaxEnt при фиксированной норме.

Детали эксперимента

Архитектура модели. Использована оригинальная гибридная нейросетевая архитектура FlexibleCNN, специально разработанная для обработки одномерных спектральных данных (OFDM-сигналов). Модель состоит из следующих блоков.

Сверточная часть. Три одномерных сверточных слоя с числом каналов 8, 16 и 32, размером ядра 5, каждый сопровождается пакетной нормализацией (BatchNorm1d), активацией GELU и подвыборкой (MaxPool1d) с окном 3. Входной сигнал имеет длину $N = 200$ и подается как одноканальный тензор.

Механизм самовнимания. После сверток тензор перепорядочивается в формат (batch, seq_len, channels) и поступает в блок самовнимания, который вычисляет взвешенную сумму по временным позициям; используется остаточная связь и слоевая нормализация.

Внимание по каналам. Адаптивно перемасштабирует каналные признаки с помощью сжимающих полностью связанных слоев и сигмоидной активации.

Усреднение по времени. Глобальное среднее пулинг преобразует тензор размерности (batch, seq_len, channels) в вектор признаков размерности 32.

Полносвязная часть. Вектор размерности 32 проходит через два полносвязных слоя с размерами 256 и 128; каждый слой сопровождается пакетной нормализацией, активацией GELU и Dropout с вероятностью 0,3. Выход последнего скрытого слоя (размерность 128) используется как эмбеддинг объекта и перед вычислением компонентов функции потерь нормируется к единичной длине (L2нормализация).

Классификационная головка. Финальный линейный слой преобразует нормированный эмбеддинг размерности 128 в логиты для четырех классов.

Инициализация весов. Для сверточных слоев — метод He (kaiming_normal), для линейных слоев — нормальное распределение со стандартным отклонением 0,01.

Функция потерь. Модель обучается путем минимизации совокупной функции уравнения (10) HEMS.

Гиперпараметры HEMS. $\lambda_{\text{max}} = 0,1$ (максимальное значение коэффициента энтропийной регуляризации); $T_{\text{warmup}} = 10$ эпох (длительность разогрева); температура $t = 0,1$ в $\mathcal{L}_{\text{uniformity}}$; $\gamma = 0,1$ — вес штрафа за корре-

ляции в $\mathcal{L}_{\text{diversity}}$; $\alpha = 0,99$ — коэффициент сглаживания при обновлении скользящих центроидов в $\mathcal{L}_{\text{separation}}$; порог разделения $m = 1,0$ в $\mathcal{L}_{\text{separation}}$; $\lambda_{\text{diversity}} = 0,01$, $\lambda_{\text{sep}} = 0,05$ (подобраны на валидационной выборке).

Оптимизация. AdamW, learning rate = 0,001, снижение в два раза при плато на валидации (терпение 5 эпох).

Размер батча: 256.

Разбиение данных. Случайное стратифицированное разделение на обучающую (70 %) и тестовую (30 %) выборки. Для подбора гиперпараметров из обучающей выборки дополнительно выделялась валидационная подвыборка (20 % от обучающей, т. е. 14 % от общего объема данных). Все метрики сообщаются по тестовой выборке. Нормализация входных признаков (центрирование и масштабирование) выполнялась только по обучающим данным.

Повторяемость. Все эксперименты проведены с 5 различными случайными seed'ами; в таблицах приведены средние значения $\pm$ стандартные отклонения.

Результаты эксперимента

Результаты численных исследований, проведенных на корпусе данных большой емкости (не менее 17 000 образцов сигналов каждого типа) при общем числе тестовых образцов 34 667 и значении параметра $N = 200$, приведены в табл. 1. В графах значений достигнутых метрик показаны полученные величины для двух случаев применений представленной модели: для принципа MaxEnt при построении оценок и для традиционной функции потерь (CE, без HEMS (MaxEnt) регуляри-

зации). Числа поддержек (столбец support (табл. 1)) соответствуют тестовой выборке объемом 34 667 образцов, что составляет 29,2 % от общего объема корпуса (118 784). Небольшое отклонение от 30 % обусловлено особенностями стратифицированного разбиения при целочисленных размерах классов и не влияет на достоверность результатов.

В табл. 2 приведены сравнительные метрики классификации (HEMS vs без HEMS), табл. 3 содержит сводную оценку достигнутого качества («вариант HEMS» против варианта «без HEMS»).

В табл. 2 MeanCosSim вычислен по нормированным эмбедингам на тестовой выборке; меньшее значение означает более равномерное распределение. MaxEnt по сходствам для тестовой выборки (34 667 образцов) равна 10,453 нат, при этом достигнутое значение 10,34 нат близко к теоретическому пределу. В табл. 3 отражен вклад компонентов HEMS в качество решения классификационной задачи.

В табл. 4 приведены результаты сравнения, предложенного метода HEMS с альтернативными регуляризаторами.

Экспериментальные результаты (табл. 2–4) подтверждают, что предложенный метод HEMS систематически превосходит стандартные приемы регуляризации, а каждый его компонент вносит положительный вклад. Применение HEMS обеспечило формирование высоко-равномерного пространства эмбедингов: MeanCosSim снизилось до 0,064, coverage достигло 1,71 (0,85 от максимума равного 1,0), а entropy — 10,34 нат (0,99 от теоретического предела равного 1,0). Достигнутый баланс между равномерностью и разделимостью

Таблица 1. Результаты численных исследований (данные усреднены по 5 запускам)

Table 1. Results of numerical studies (data averaged over 5 runs)

Класс	Precision*	Recall*	F1-score*	Support	Интерпретация
Class_ATL	0,966/0,958	0,973/0,965	0,970/0,961	5,268	Отличный баланс
BackGround	0,988/0,985	0,992/0,990	0,990/0,988	18,590	Почти идеально
Class_DJ	0,982/0,976	0,960/0,954	0,971/0,965	4,955	Хороший баланс
Class_F_ANLG	1,000/1,000	1,000/1,000	1,000/1,000	5,854	Идеальное разделение
Интегральные показатели					
Accuracy*				0,9862/0,9825	
Macro F1-score*				0,9827/0,9783	

Примечание: * — метрики, для которых показаны результаты: значение 1 соответствует принципу максимальной энтропии MaxEnt; значение 2 — традиционной функции потерь.

Таблица 2. Сравнительные метрики классификации

Table 2. Comparative Classification Metrics

Метрика	с HEMS	без HEMS
Accuracy	0,986 $\pm$ 0,121	0,982 $\pm$ 0,001
Macro F1-score	0,982 $\pm$ 0,002	0,978 $\pm$ 0,003
MeanCosSim	0,064 $\pm$ 0,003	0,098 $\pm$ 0,004
Coverage	1,710 $\pm$ 0,021	1,652 $\pm$ 0,021
Diversity	50,350 $\pm$ 1,201	35,092 $\pm$ 1,501
Entropy, нат	10,340 $\pm$ 0,052	7,912 $\pm$ 0,101

Таблица 3. Вклад компонентов HEMS
Table 3. Contribution of HEMS Components

Конфигурация	Accuracy	MeanCosSim ↓	Diversity ↑
CE only	0,9825 ± 0,0015	0,098 ± 0,004	35,1 ± 1,5
CE + $\mathcal{L}_{\text{uniformity}}$	0,9842 ± 0,0010	0,072 ± 0,003	41,8 ± 1,0
CE + $\mathcal{L}_{\text{uniformity}}$ + $\mathcal{L}_{\text{diversity}}$	0,9851 ± 0,0008	0,068 ± 0,002	48,2 ± 0,9
CE + $\mathcal{L}_{\text{uniformity}}$ + $\mathcal{L}_{\text{diversity}}$ + $\mathcal{L}_{\text{separation}}$ (полный HEMS)	0,9860 ± 0,0012	0,064 ± 0,003	50,3 ± 1,2

Примечание: ↓ — (стрелка вниз) — чем меньше значение метрики, тем лучше; ↑ (стрелка вверх) — чем больше значение метрики, тем лучше.

Таблица 4. Сравнение предложенного метода HEMS с альтернативными регуляризаторами
Table 4. Comparison with Alternative Regularizers

Метод	Accuracy	Macro F1-score
CE (базовый)	0,9825 ± 0,0015	0,978 ± 0,003
CE + Weight Decay	0,9830 ± 0,0014	0,979 ± 0,003
CE + Dropout (0,5)	0,9833 ± 0,0013	0,979 ± 0,003
CE + Label Smoothing (0,1)	0,9837 ± 0,0011	0,980 ± 0,002
CE + Mixup ($\alpha = 0,2$)	0,9841 ± 0,0012	0,981 ± 0,002
HEMS (предложенный)	0,9860 ± 0,0012	0,982 ± 0,002

(diversity = 50,35, межклассовые расстояния сохранены благодаря $\mathcal{L}_{\text{separation}}$) позволил повысить Accuracy до 0,986 и Macro F1-score до 0,982. Таким образом, данные полностью подтверждают гипотезу H1: максимизация энтропии распределения признаков при заданных ограничениях улучшает обобщающую способность модели.

Заключение

В работе представлено обоснование и экспериментальная проверка гипотезы H1, устанавливающей прямую связь между максимизацией энтропии распределения эмбедингов при заданных ограничениях и улучшением обобщающей способности модели. Теоретический вклад заключается в обосновании того, что распределение с максимальной энтропией при фиксированных моментах минимизирует априорные предположения, обеспечивает наилучшее покрытие пространства признаков и снижает радиемахеровскую сложность класса моделей. Это, в свою очередь, сужает границы обобщения и уменьшает риск больших ошибок. Экспериментальная часть, выполненная на задаче классификации зашумленных OFDM-сигналов с выраженным дисбалансом классов, подтвердила практическую эффективность предложенного подхода — регуляризации методом максимизации энтропии на гиперсфере с разделением классов. Важно отметить, что даже при достижении близкой к максимальной энтропии распределения (10,34 нат, 0,99 от максимума равного 1,0) и почти ортогонального углового расстояния между эмбедингами (86° , $\arccos(0,064) = 86^\circ 18'$), модель сохранила отличную дискриминативность, что подтверждается точностью 0,986 и сбалансированным Macro F1-score = 0,982. Ключевым достижением рабо-

ты является демонстрация оптимального компромисса: компонент $\mathcal{L}_{\text{uniformity}}$ обеспечил равномерное распределение на гиперсфере, максимизируя энтропию, в то время как компоненты $\mathcal{L}_{\text{separation}}$ и $\mathcal{L}_{\text{diversity}}$ выступили в роли необходимых ограничений, сохранивших межклассовую разделимость и предотвративших ухудшение признаков.

Основные ограничения работы связаны со следующими предположениями теоретической модели: необходимость симметрии распределения эмбедингов, что на практике обеспечивается центрированием; требование фиксированных моментов (ковариации), предполагающее предварительную нормализацию данных; асимптотический характер полученных оценок; стандартное для теории статистического обучения предположение о липшицевости моделей и функций потерь.

Перспективные направления дальнейших исследований, направленные на развитие представленных результатов, включают: ослабление предположения о симметрии; вывод неасимптотических оценок для малых выборок; учет специфики современных архитектур глубоких нейросетей (например, механизмов внимания) и масштабную экспериментальную верификацию на более широком классе задач.

Таким образом, работа подтверждает, что явная максимизация энтропии представлений в рамках заданных ограничений является обоснованным принципом для улучшения обобщающей способности. Этот принцип не только объясняет успех многих современных методов контрастного и метрического обучений, но и задает методологическую основу для разработки новых регуляризационных техник, направленных на построение более устойчивых и эффективных моделей машинного обучения.

Литература

1. Bengio Y., Courville A., Vincent P. Representation learning: A review and new perspectives // *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013. V. 35. N 8. P. 1798–1828. doi: 10.1109/tpami.2013.50
2. Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations // *Proc. of the 37th International Conference on Machine Learning*. 2020. P. 1597–1607.
3. He K., Fan H., Wu Y., Xie S., Girshick R. Momentum contrast for unsupervised visual representation learning // *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020. P. 9726–9735. doi: 10.1109/cvpr42600.2020.00975
4. Wang T., Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere // *Proc. of the 37th International Conference on Machine Learning*. 2020. P. 9929–9939.
5. Chakraborty D., LeCun Y., Rudner T.G.J., Learned-Miller E. Improving pre-trained self-supervised embeddings through effective entropy maximization // *arXiv*. 2024. arXiv:2411.15931. doi: 10.48550/arXiv.2411.15931
6. Берестова С.А., Копытов Н.П., Митюшов Е.А. Дискретные ориентации космического аппарата // *Инженерный журнал: наука и инновации*. 2017. № 7 (67). С. 7. doi: 10.18698/2308-6033-2017-7-1661
7. Pang T., Yang X., Dong Y., Xu K., Zhu J., Su H. Boosting adversarial training with hypersphere embedding // *Proc. of the 34th Conference on Neural Information Processing Systems (NeurIPS)*. 2020.
8. Jakovác A., Telcs A. A note on representational understanding // *Entropy*. 2022. V. 24. N 9. P. 1313. doi: 10.3390/e24091313
9. Garg V.K., Jegelka S., Jaakkola T. Generalization and representational limits of graph neural networks // *Proc. of the 37th International Conference on Machine Learning*. 2020. P. 3419–3430.
10. Ren B., Teh K.C., An H., Gunawan E. OFDM modulation classification using Cross-SKNet with blind IQ imbalance and carrier frequency offset compensation // *IEEE Transactions on Vehicular Technology*. 2024. V. 73. N 6. P. 8389–8403. doi: 10.1109/tvt.2024.3356606
11. Данфорд Н., Шварц Дж. Линейные операторы. Т. 1. М.: Иностранная литература, 1966. С. 326–330.
12. Рудин У. Основы математического анализа. М.: Мир, 1976. С. 99.
13. Anderson T.W. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities // *Proceedings of the American Mathematical Society*. 1955. V. 6. N 2. P. 170–176. doi: 10.1090/S0002-9939-1955-0069229-1
14. Prekopa A. On logarithmic concave measures and functions // *Acta Scientiarum Mathematicarum*. 1973. V. 34. P. 335–343.
15. Mohri M., Rostamizadeh A., Talwalkar A. *Foundations of Machine Learning*. MIT Press, 2018. 504 p.
16. Bartlett P.L., Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results // *Journal of Machine Learning Research*. 2002. V. 3. P. 463–482.
17. Cover T.M., Thomas J.A. *Elements of Information Theory*. John Wiley & Sons, 2006. 784 p.
18. El-Khamy S.E., Elsayed H.A., Rizk M.M. C45. Classification of OFDM signals using higher order statistics and clustering techniques // *Proc. of the 29th National Radio Science Conference (NRSC)*. 2012. P. 541–549. doi: 10.1109/nrsc.2012.6208563

Автор

Тимофеев Андрей Владимирович — доктор технических наук, научный директор, ООО «Центр ИИ для Стран ШОС+», Санкт-Петербург, 197341, Российская Федерация, [sc 56689367600](https://orcid.org/0000-0001-7212-5230), <https://orcid.org/0000-0001-7212-5230>, timofeev.andrey@gmail.com, tav@centeraisco.com

Статья поступила в редакцию 18.01.2026
Одобрена после рецензирования 16.04.2026
Принята к печати 20.05.2026

References

1. Bengio Y., Courville A., Vincent P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, vol. 35, no. 8, pp. 1798–1828. doi: 10.1109/tpami.2013.50
2. Chen T., Kornblith S., Norouzi M., Hinton G. A simple framework for contrastive learning of visual representations. *Proc. of the 37th International Conference on Machine Learning*, 2020, pp. 1597–1607.
3. He K., Fan H., Wu Y., Xie S., Girshick R. Momentum contrast for unsupervised visual representation learning. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735. doi: 10.1109/cvpr42600.2020.00975
4. Wang T., Isola P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *Proc. of the 37th International Conference on Machine Learning*, 2020, pp. 9929–9939.
5. Chakraborty D., LeCun Y., Rudner T.G.J., Learned-Miller E. Improving pre-trained self-supervised embeddings through effective entropy maximization. *arXiv*, 2024. arXiv:2411.15931. doi: 10.48550/arXiv.2411.15931
6. Berestova S.A., Kopytov N.P., Mityushov E.A. Spacecraft discrete orientations. *Engineering Journal: Science and Innovation*, 2017, no. 7 (67), pp. 7. (in Russian). doi: 10.18698/2308-6033-2017-7-1661
7. Pang T., Yang X., Dong Y., Xu K., Zhu J., Su H. Boosting adversarial training with hypersphere embedding. *Proc. of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.
8. Jakovác A., Telcs A. A note on representational understanding. *Entropy*, 2022, vol. 24, no. 9, pp. 1313. doi: 10.3390/e24091313
9. Garg V.K., Jegelka S., Jaakkola T. Generalization and representational limits of graph neural networks. *Proc. of the 37th International Conference on Machine Learning*, 2020, pp. 3419–3430.
10. Ren B., Teh K.C., An H., Gunawan E. OFDM modulation classification using Cross-SKNet with blind IQ imbalance and carrier frequency offset compensation. *IEEE Transactions on Vehicular Technology*, 2024, vol. 73, no. 6, pp. 8389–8403. doi: 10.1109/tvt.2024.3356606
11. Dunford N., Schwartz J. *Linear Operators*. Part 1. Interscience Publishers, 1958, pp. 326–330
12. Rudin W. *Principles of Mathematical Analysis*. Moscow, Mir Publ., 1976, pp. 99. (in Russian)
13. Anderson T.W. The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society*, 1955, vol. 6, no. 2, pp. 170–176. doi: 10.1090/S0002-9939-1955-0069229-1
14. Prekopa A. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 1973, vol. 34, pp. 335–343.
15. Mohri M., Rostamizadeh A., Talwalkar A. *Foundations of Machine Learning*. MIT Press, 2018, 504 p.
16. Bartlett P.L., Mendelson S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 2002, vol. 3, pp. 463–482.
17. Cover T.M., Thomas J.A. *Elements of Information Theory*. John Wiley & Sons, 2006, 784 p.
18. El-Khamy S.E., Elsayed H.A., Rizk M.M. C45. Classification of OFDM signals using higher order statistics and clustering techniques. *Proc. of the 29th National Radio Science Conference (NRSC)*, 2012, pp. 541–549. doi: 10.1109/nrsc.2012.6208563

Author

Andrey V. Timofeev — D.Sc., Chief Scientific Officer, LLC “AI Center for SCO+ Countries”, Saint Petersburg, 197341, Russian Federation, [sc 56689367600](https://orcid.org/0000-0001-7212-5230), <https://orcid.org/0000-0001-7212-5230>, timofeev.andrey@gmail.com, tav@centeraisco.com

Received 18.01.2026
Approved after reviewing 16.04.2026
Accepted 20.05.2026



Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»