

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ/ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

DOI: https://doi.org/10.60797/IRJ.2026.169.3 EDN: INTZSF

РЕГУЛЯРИЗАЦИЯ И ИНДУКТИВНОЕ СМЕЩЕНИЕ МЕХАНИЗМОВ САМОВНИМАНИЯ В ГИБРИДНЫХ НЕЙРОННЫХ СЕТЯХ ДЛЯ КЛАССИФИКАЦИИ В КОМПЬЮТЕРНОМ ЗРЕНИИ С ИСПОЛЬЗОВАНИЕМ ШТРАФА НА ОСНОВЕ ЭНТРОПИИ ЦАЛЛИСА

Научная статья

Бабушкин А.С.[1, *]

[1] ORCID : 0009-0006-2182-633X;

[1] Уральский федеральный университет имени первого Президента России Б.Н. Ельцина, Екатеринбург, Российская Федерация

* Корреспондирующий автор (babushkin.alexey[at]urfu.me)

Предложена: 10.02.2026; Принята: 06.07.2026; Опубликована: 17.07.2026

Аннотация

Архитектуры на основе трансформеров занимают лидирующие позиции в задачах глубокого обучения, включая компьютерное зрение, благодаря способности моделировать глобальные зависимости через механизм самовнимания. Однако стандартные методы регуляризации (L1/L2, Dropout) воздействуют преимущественно на веса или нейроны нейросети по отдельности, не оказывая прямого комплексного влияния на распределение весов в матрицах внимания, что может быть недостаточно эффективно для трансформерной архитектуры. В данной работе предлагается экспериментальный подход к регуляризации гибридных сверточно-трансформерных нейросетей с использованием штрафа на основе энтропии Цаллиса. В отличие от энтропии Шеннона, энтропия Цаллиса содержит параметр деформации q, позволяющий гибко управлять чувствительностью к редким и частым событиям и балансировать между разреженным и распределенным вниманием.

Для оценки эффективности метода он был апробирован на двух гибридных моделях (940 тыс. и 5 млн параметров) на наборе данных Imagenette2-160, где оптимизация гиперпараметров проводилась с помощью фреймворка Optuna с одинаковым бюджетом в 100 trials для моделей с добавлением штрафа и без. Результаты экспериментов на 10 различных случайных инициализациях (seeds) показали статистически значимое ($p < 0.001$) преимущество предложенного подхода. Для компактной модели прирост точности составил 3,7% при силе эффекта $d \approx 2,9$, а для модели большего масштаба — 4,1% при $d \approx 4$. Столь высокие значения коэффициента Коэна подтверждают, что внедрение индуктивного смещения через энтропийный штраф радикально улучшает обобщающую способность гибридных архитектур, позволяя получать большую точность, не меняя число параметров или размеры датасетов.

Ключевые слова: трансформеры, сверточно-трансформерные архитектуры, компьютерное зрение, гибридные нейросети, классификация изображений, оптимизация гиперпараметров, Optuna, байесовская оптимизация, контролируемое сравнение.

REGULARISATION AND INDUCTIVE BIAS OF SELF-ATTENTION MECHANISMS IN HYBRID NEURAL NETWORKS FOR COMPUTER VISION CLASSIFICATION USING A TSALLIS ENTROPY-BASED PENALTY

Research article

Babushkin A.S.[1, *]

[1] ORCID : 0009-0006-2182-633X;

[1] Ural Federal University named after the first President of Russia B.N. Yeltsin, Ekaterinburg, Russian Federation

* Corresponding author (babushkin.alexey[at]urfu.me)

Suggested: 10.02.2026; Accepted: 06.07.2026; Published: 17.07.2026

Abstract

Transformer-based architectures are at the forefront of deep learning tasks, including computer vision, due to their ability to model global dependencies through the self-attention mechanism. However, standard regularisation methods (L1/L2, Dropout) primarily affect the weights or neurons of a neural network individually, without exerting a direct, complex influence on the distribution of weights in the attention matrices, which may not be sufficiently effective for transformer architectures. This work proposes an experimental approach to regularising hybrid convolutional-transformer neural networks using a Tsallis entropy-based penalty. Unlike Shannon entropy, Tsallis entropy includes a deformation parameter q, which allows for flexible control of sensitivity to rare and frequent events and strikes a balance between sparse and distributed attention.

To evaluate the method's effectiveness, it was tested on two hybrid models (940,000 and 5 million parameters) using the ImageNet-160 dataset, where hyperparameter optimisation was carried out with the Optuna framework, using an identical budget of 100 trials for both models with and without a penalty term. The results of experiments using 10 different random initialisations (seeds) showed a statistically significant ($p < 0.001$) advantage of the suggested approach. For the compact model, the accuracy gain was 3.7% with an effect size of $d \approx 2.9$, while for the larger-scale model it was 4.1% with $d \approx 4$. Such high values of the Cohen coefficient confirm that the introduction of inductive bias via an entropy penalty radically improves

the generalisation ability of hybrid architectures, enabling higher accuracy to be achieved without changing the number of parameters or the size of the datasets.

Keywords: transformers, convolutional-transformer architectures, computer vision, hybrid neural networks, image classification, hyperparameter optimisation, Optuna, Bayesian optimisation, controlled comparison.

Введение

Несмотря на то, что концепция механизмов внимания была известна ранее как вспомогательный инструмент для рекуррентных моделей, работа А. Васвани и соавторов [1] стала переломной и показала, что можно исключить рекуррентные слои и использовать только самовнимание и FFN блоки. За прошедшее десятилетие этот подход доказал свою эффективность, однако стандартные методы регуляризации до сих пор слабо адаптированы к специфике распределения весов в матрицах внимания и не могут контролировать их. L1/L2 регуляризации контролируют лишь размеры весов, в то время как dropout не позволяет модели опираться на одни и те же нейроны при обучении. При этом из-за своей сложности, сам процесс применения самовнимания эти методы регуляризации практически не затрагивают напрямую, т.к. самовнимание является комплексным процессом взаимодействия всех токенов друг с другом и не может контролироваться грубым влиянием на отдельные веса или нейроны в полносвязных слоях. Самостоятельный контроль самовнимания моделью требует модифицирования обучения модели с использованием дополнительных штрафов на распределения самовнимания с применением метода градиентного спуска для подстраивания весов полносвязных слоев определенным образом, приводящим к нужным распределениям матриц весов после применений softmax.

Используемый в исследовании подход реализует механизм осознанного (структурного) контроля информационных потоков. Предлагается использование штрафа на основе энтропии Цаллиса для промежуточных слоёв нейросети, который, как предполагается, должен внести в трансформерную модель большую степень регуляризации, а также добавить дополнительное индуктивное смещение на основе понимания требуемой энтропии распределений весов самовнимания, в конкретной форме, выведенной из обобщённой энтропии Цаллиса [2] байесовской оптимизацией (Optuna).

Данное направление в методах регуляризации нейросетей привлекает внимание в последнее время. Различные исследования, используют контроль энтропии в своих целях [3], [4], [5]. При этом исследований конкретно энтропии Цаллиса при использовании с матрицами самовнимания для стандартных архитектур с трансформерными блоками в значимом объеме не проводилось.

Помимо сравнения изменения точности моделей с добавлением штрафа, в ходе эксперимента также будет наблюдаться оптимальное по итогам оптимизации значение параметра деформации q в формуле энтропии Цаллиса, определяющего степень деформации (неэкстенсивности) системы, который и отличает её от энтропии Шеннона.

Методы и принципы исследования

2.1. Архитектура исследуемых моделей

В эксперименте используются две гибридные архитектуры, сочетающие сверточные слои и механизмы самовнимания (Self-Attention). Модели различаются масштабом, то есть количеством обучаемых параметров, что позволяет оценить универсальность предлагаемого метода: модель 1 содержит около 940 тыс. параметров, модель 2 — около 5 млн параметров. Обе архитектуры имеют последовательную иерархическую структуру, состоящую из классических основных компонентов.

2.1.1. Сверточный экстрактор признаков

Сверточный экстрактор признаков включает три последовательных блока. Каждый блок увеличивает глубину каналов и уменьшает пространственное разрешение в два раза. Это позволяет сформировать компактные и информативные локальные признаки перед подачей в трансформерную часть. Сверточный экстрактор обладает явным индуктивным смещением, характерным для сверточных слоев.

2.1.2. CLS-токен

CLS-токен представляет собой обучаемый токен, который конкатенируется с последовательностью признаков, приведённой к виду [batch, seq_len, emb_dim], для сбора общей информации об изображении.

2.1.3. Позиционные эмбеддинги

Позиционные эмбеддинги представляют собой обучаемые векторы, суммируемые с последовательностью признаков перед трансформерной частью для добавления позиционной информации.

2.1.4. Трансформерный стек

Трансформерный стек состоит из четырех блоков трансформера для модели 1 и из шести блоков трансформера для модели 2. Реализация включает стандартные механизмы многоголового самовнимания и полносвязные блоки. Архитектура модифицирована для извлечения матриц внимания после операции Softmax, что необходимо для вычисления штрафа на основе энтропии Цаллиса с сохранением графа вычислений для обратного распространения ошибки. Введение штрафа на основе энтропии Цаллиса выступает в роли регуляризующего индуктивного смещения, накладывающего структурные ограничения на распределение весов внимания. Длина векторов признаков, определяемая выходом сверточных экстракторов, составляет 128 и 256 для малой и большей модели соответственно.

2.1.5. Полносвязный классификатор

Полносвязный классификатор принимает CLS-токен с агрегированной информацией об изображении и выдает 10 логитов для последующего вычисления кросс-энтропии.

2.1.6. Нормализация и функции активации

В сверточной части используется групповая нормализация (GroupNorm) вместо BatchNorm для обеспечения стабильности при малых размерах батча. В трансформерной части и классификаторе применяется нормализация по

слою (LayerNorm). В качестве функций активации во всех слоях используется GELU, что соответствует стандартному выбору для многих трансформерных архитектур.

2.2. Датасет

В ходе исследования использовался набор данных Imagenette2-160 (версия 2020 года). Данный датасет представляет собой сбалансированную выборку из 10 классов ImageNet (tench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, parachute). Выбор данного набора обусловлен необходимостью проведения большого количества итераций обучения (в рамках оптимизации гиперпараметров) при сохранении сложности реальных фотографических данных.

Изображения имеют фиксированный размер 160 px по меньшей стороне. Перед подачей в модель все изображения приводились к строгому разрешению 160x160 пикселей с использованием метода Resize библиотеки torchvision.transforms.

Структура выборки: В работе использовано современное разбиение (split), обеспечивающее высокую статистическую значимость оценки работы модели:

— Обучающая выборка: 9 469 изображений.

— Валидационная выборка: 3 925 изображений.

2.3. Регуляризация на основе энтропии Цаллиса

Ключевой особенностью предлагаемого метода является модификация функции потерь за счет добавления регуляризационного члена $R_{Tsallis}$, воздействующего на распределение весов в матрицах самовнимания. В отличие от стандартных подходов, данный метод минимизирует отклонение текущей энтропии распределений внимания от заданного целевого значения H_{target}.

Энтропия Цаллиса вычисляется для каждого распределения внимания, соответствующего отдельной голове и каждому токену-запросу, и затем агрегируется по всем головам и слоям модели. Тем самым вводится индуктивное смещение, направленное на контроль степени концентрации информационных потоков в механизме самовнимания

Вид штрафа (формулы 1 и 2):

$$S_q(P) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right) \quad (1)$$

Для обеспечения численной стабильности при q ≈ 1 (в диапазоне 0.95–1.05) используется энтропия Шеннона, являющаяся частным случаем энтропии Цаллиса:

$$S_{Shannon}(P) = -\sum_{i=1}^{n} p_i \ln(p_i) \quad (2)$$

где:

p_i — веса самовнимания для каждого токена;

q — параметр деформации, определяющий чувствительность меры к форме распределения.

При близости q к 1, подобное изменение формулы позволяет избегать операций с очень маленькими или очень большими числами из-за близости знаменателя к 0, что важно при использовании смешанной точности для ускорения на современных видеокартах, так как это уменьшает численные нестабильности/погрешности во время обучения.

Регуляризационный член функции потерь определяется как квадратичное отклонение нормированной энтропии от целевого уровня:

$$R_{Tsallis} = \lambda \cdot \left(\frac{S_q(P)}{S_q^{max}} - H_{\text{target}}\right)^2 \quad (3)$$

p — веса внимания для каждого токена;

q — параметр деформации, определяющий степень концентрации или разреженности внимания;

λ — вес штрафа;

$S_q(P)$ — используемая энтропия;

S^{max} — максимально возможная энтропия для данной длины последовательности n.

Формула S^{max} для нормализации энтропии выглядит как $S_{q,max} = \frac{1-n^{1-q}}{q-1}$ (n - количество элементов), при использовании версии Шеннона как $S_{max} = \ln n$ и выводится через уравнивание всех членов последовательности.

Итоговая функция эта сумма кросс-энтропии (10 классов) и введённого регуляризационного члена:

$$L = L_{CE} + R_{Tsallis} \quad (4)$$

L — общая функция потерь;

L_{CE} — функция потерь кросс энтропии;

$R_{Tsallis}$ — энтропийный штраф.

2.4. Протокол эксперимента

Процесс оценки эффективности предложенного метода разделен на два последовательных этапа.

1. Оптимизация гиперпараметров.

Для каждой из четырех конфигураций (Модель 1 и Модель 2 в вариантах со штрафом и без него) проводился поиск оптимальных гиперпараметров с помощью фреймворка Optuna. Использовалась байесовская оптимизация (TPE sampler multivariate) для подбора гиперпараметров.

Таблица 1 - Полное пространство поиска гиперпараметров

DOI: https://doi.org/10.60797/IRJ.2026.169.3.1

Гиперпараметр	Тип	Диапазон / Значения	Описание
lr	float (логарифмический)	5e-5 – 1e-3	Скорость обучения
batch_size	категориальный выбор	32 / 64 / 128	Число примеров в 1 шаге оптимизации
weight_decay	категориальный выбор	0 / 1e-4 / 1e-3 / 1e-2	Корректная L2 регуляризация встроенная в оптимизатор AdamW
dropout	категориальный выбор	0 / 0,1 / 0,2 / 0,3 / 0,4 / 0,5	Вероятность зануления нейронов в блоках трансформеров во время обучения
tsallis_q	float	0,2 – 4,0	Параметр деформации. Инструмент, позволяющий учитывать неэкстенсивность энтропии, определяющий её чувствительность к малым или большим вероятностям.
entropy_target	float	0,0 – 1,0	Целевой уровень энтропии. Нормализован в диапазон [0; 1] делением на максимально возможный уровень энтропии для заданной длины последовательности
penalty_weight	float	1,0 – 500,0	Вес штрафа. Определяет относительную силу штрафа в сравнении с основной функцией потерь. Размеры градиентов зависят не только от веса, но и от производных, поэтому значение равное 1 не обязательно вносит сопоставимый вклад с основной функцией потерь и его нужно подбирать в широком диапазоне.

Для базовых моделей (без штрафа) пространство поиска включало 4 гиперпараметра (lr, batch_size, weight_decay, dropout). Вычислительный бюджет составил 100 проб (trials) с 20 случайными пробами в начале, чего должно быть достаточно для нахождения субоптимальных значений в пространстве низкой размерности.

Для моделей с регуляризацией Цаллиса количество оптимизируемых параметров увеличилось до 7. Дополнительно подбирались tsallis_q, entropy_target и penalty_weight (см. таблицу «Полное пространство поиска гиперпараметров»). Вычислительный бюджет остался в рамках 100 проб (trials)

Стоит отметить, что несмотря на равенство в бюджете, сложность поиска растёт комбинаторно от числа параметров т.е. для исследования пространства поиска с регуляризацией на основе энтропии Цаллиса, сложность подбора гиперпараметров с помощью Optuna значительно выше, чем для стандартного обучения. Существует априорное смещение в пользу базовых моделей из-за большей сложности подбора гиперпараметров в пространстве с дополнительными размерностями, что делает любые положительные результаты регуляризации Цаллиса

статистически более значимыми. В том числе, это связано с увеличением числа возможных двойных или тройных взаимодействий в 21/6 = 3,5 раза и 35/4 = 8,75 раз соответственно, учёт которых или выделение алгоритмом отсутствия связи между которыми занимает большее число trials.

Прочие условия обучения также были равны для всех моделей и основаны на общепринятых правилах или достаточны для данной задачи: косинусное расписание скорости обучения (learning rate), максимум 100 эпох для обучения, а также ранняя остановка при отсутствии какого-либо увеличения валидационной точности в течение 6 эпох подряд.

С целью анализа данных, собранных во время работы с Optuna и для защиты от их потери в случае прерывания обучения, Optuna сохраняла все результаты запусков в базе данных Architectures_optimization.db с 4 study - исследованиями 4 вариантов обучения (2 модели с/без штрафа).

2. Сбор статистических данных.

Для исключения влияния фактора случайности при инициализации весов, лучшие конфигурации, найденные на этапе оптимизации, обучались повторно. Для каждого из 4 случаев было проведено по 10 запусков с использованием различных значений случайного начального состояния (seed) для всех используемых компонентов. На основе полученных результатов рассчитывались:

— средняя точность;

— стандартное отклонение.

Эти данные собирались в соответствующие .csv файлы и на их основе были рассчитаны статистическая значимость и сила эффекта.

Основные результаты

Optuna: Лучшие подобранные данные за 100 проб с байесовской оптимизацией, а также их оцененная сила влияния.

1. Архитектура 1.

Без штрафа:

— lr: 0.0007609;

— batch_size: 64;

— weight_decay: 0.001;

— dropout: 0.2;

— Точность на валидации: 0.7259.

Со штрафом:

— lr: 0.0006563;

— batch_size: 64;

— weight_decay: 0.01;

— dropout: 0.5;

— tsallis_q: 0.934;

— entropy_target: 0.743;

— penalty_weight: 16.56;

— Точность на валидации: 0.7455.

2. Архитектура 2.

Без штрафа:

— lr: 0.0001536;

— batch_size: 32;

— weight_decay: 0.01;

— dropout: 0.3;

— Точность на валидации: 0.7154.

Со штрафом:

— lr: 0.0008381;

— batch_size: 128;

— weight_decay: 0.001;

— dropout: 0.3;

— tsallis_q: 1.466;

— entropy_target: 0.974;

— penalty_weight: 71.26;

— Точность на валидации: 0.7434.

В dashboard Optuna функция fANOVA показывает, что вклад гиперпараметров, связанных со штрафом при его добавлении, превышает 50%. Следует отметить, что все оценки fANOVA являются стохастическими и могут немного меняться между запусками, поскольку процесс сэмплирования случаен и на этапе поиска гиперпараметров использовались ускоряющие оптимизации, которые дают значимый прирост, но нарушают детерминированность.

Статистический анализ при разных seed представлен в таблице ниже.

Таблица 2 - Данные о результатах проверки 2 архитектур

DOI: https://doi.org/10.60797/IRJ.2026.169.3.2

Архитектура	Без штрафа (среднее ± СКО)	Со штрафом (среднее ± СКО)	Средняя разница	Статистическая значимость (p)	Сила эффекта (d Коуэна)
1	0,7031 ± 0,0121	0,7410 ± 0,0088	0,03789	$6{,}40\times10^{-6}$ ***	2,95
2	0,6845 ± 0,0089	0,7261 ± 0,0102	0,04163	$5{,}39\times10^{-7}$ ***	3,96

Обсуждение

Модель с большим числом параметров оказалась слишком сложной для имеющегося датасета и задачи и показала худшие результаты, однако также показала большее улучшение точности с добавлением штрафа, что говорит о возможно большем уровне регуляризации при его добавлении. Разница же между моделями с и без штрафа была статистически значимой ($p < 0{,}001$) и обладала очень большим размером эффекта (Cohen's $d \approx 2{,}9$–4). Стоит также отметить, что улучшение в 3-4% при небольшом размере датасета является крайне значимым и в машинном обучении в целом, а не только относительно разброса при разном seed.

Параметр деформации q, в свою очередь, оказался вариативен и принял разные значения, меньше для малой модели и больше для большей, что привело к разнице в том числе по целевому уровню энтропии и весу штрафа. Причины подобной неоднородности требуют дальнейших исследований.

Также стоит ещё раз отметить, что разница в размерах пространств при одинаковом числе trials создавала ограничения для подбора и с большими вычислительными бюджетами для больших пространств результаты от добавления штрафа могли быть ещё более выраженными.

Заключение

В ходе исследования было установлено, что добавление энтропийного штрафа к трансформерной части приводит к увеличению top-1 accuracy (точности на валидационной выборке) гибридной нейросети в среднем на 3–4% по сравнению с базовой моделью без энтропийной регуляризации. Полученный эффект находится в диапазоне улучшений, характерных для ряда стандартных методов регуляризации, таких как dropout и L2-регуляризация, однако в рамках проведённого эксперимента действует аддитивно поверх них.

Анализ влияния параметра деформации q показал, что как значения ниже 1, так и существенно выше 1 могут приводить к достижению лучшего качества по итогам байесовской оптимизации, что может указывать на чувствительность предложенного подхода к форме используемой энтропии и необходимость дальнейшего исследования роли энтропии Цаллиса в сравнении с предельным случаем Шеннона при q стремящимся к 1.

Полученные результаты позволяют обоснованно предположить, что энтропийная регуляризация может быть эффективным дополнением к стандартным методам регуляризации, не требующим существенного увеличения вычислительных затрат на подбор гиперпараметров. Однако для подтверждения универсальности подхода необходимы дополнительные исследования на других архитектурах и задачах компьютерного зрения.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Vaswani A. Attention Is All You Need / A. Vaswani, N. Shazeer, N. Parmar et al. // Advances in Neural Information Processing Systems; edited by Guyon I. Luxburg U. V. Bengio S. Wallach H. M. Fergus R. Vishwanathan S. V. N. Garnett R. — 30. — Long Beach, Ca: Curran Associates, Inc., 2017. — P. 6000–6010. — URL: https://dl.acm.org/doi/10.5555/3295222.3295349. (accessed: 15.04.26). doi: 10.5555/3295222.3295349

2. Tsallis C. Possible generalization of Boltzmann-Gibbs statistics / C. Tsallis // Journal of Statistical Physics. — 1988. — Vol. 52. — P. 479–487. — URL: https://link.springer.com/article/10.1007/BF01016429 (accessed: 05.02.26). — DOI: 10.1007/BF01016429

3. Zhai S. Stabilizing Transformer Training by Preventing Attention Entropy Collapse / S. Zhai, T. Likhomanenko, E. Littwin et al. // Proceedings of the 40th International Conference on Machine Learning (PMLR); edited by Krause A. Brunskill E. Cho K. Engelhardt B. Sabato S. Scarlett J — 202. — New York: Proceedings of Machine Learning Research, 2023. — P. 40770–40803. — URL: https://proceedings.mlr.press/v202/zhai23a.html. (accessed: 06.02.26).

4. Lam W.-L Energy-Entropy Regularization: The True Power of Minimal Looped Transformers / W.-L Lam // arXiv. — 2026. — №2601. — URL: https://arxiv.org/abs/2601.09588. (accessed: 04.02.26) doi: 10.48550/arXiv.2601.09588

5. Peruzzo E. Spatial Entropy as an Inductive Bias for Vision Transformers / E. Peruzzo, E. Sangineto, Y. Liu et al. // Machine Learning. — 2024. — 113(9). — P. 6945–6975. — URL: https://link.springer.com/article/10.1007/s10994-024-06570-7 (accessed: 06.02.26). — DOI: 10.1007/s10994-024-06570-7

6. Akiba T. Optuna: A Next-generation Hyperparameter Optimization Framework / T. Akiba, S. Sano, T. Yanase et al. // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; — New York: Association for Computing Machinery, 2019. — P. 2623–2631. — URL: https://dl.acm.org/doi/10.1145/3292500.3330701. (accessed: 05.02.26). doi: 10.1145/3292500.3330701

7. Franceschi L. Hyperparameter Optimization in Machine Learning / L. Franceschi, M. Donini, V. Perrone et al. // Foundations and Trends in Machine Learning. — 2025. — 18 (6). — P. 975–1109. — URL: https://doi.org/10.1561/2200000088 (accessed: 06.02.26). — DOI: 10.1561/2200000088

8. Howard J. Imagenette: a smaller subset of 10 easily classified classes from ImageNet. Version 2 / J. Howard // GitHub repository fastai/imagenette. — 2020. — URL: https://github.com/fastai/imagenette. (accessed: 05.02.26)

9. Ba J.L. Layer Normalization / J.L. Ba, J.R. Kiros, G.E. Hinton // arXiv. — 2016. — URL: https://arxiv.org/abs/1607.06450. (accessed: 12.02.26) doi: 10.48550/arXiv.1607.06450

10. Dai Z. CoAtNet: Marrying Convolution and Attention for All Data Sizes / Z. Dai, H. Liu, Q.V. Le et al. // Advances in Neural Information Processing Systems; — 34. — Red Hook, Ny: Curran Associates, Inc., 2021. — P. 3965–3977. — URL: https://proceedings.neurips.cc/paper/2021/hash/20568692db622456cc42a2e853ca21f8-Abstract.html. (accessed: 07.02.26). doi: 10.48550/arXiv.2106.04803

11. Babushkin A.S TsallisPenaltyResearch / A.S Babushkin // TsallisPenaltyResearch. — 2026. — URL: https://github.com/nelex12/TsallisPenaltyResearch. (accessed: 20.03.26)