Компьютерное зрение с использованием реальных данных

1 Введение

Компьютерное зрение является одной из наиболее динамично развивающихся областей искусственного интеллекта, находящей применение в самых разных сферах — от автономных транспортных средств до медицинской диагностики. Однако успешное применение моделей компьютерного зрения в реальных условиях часто осложняется наличием шума и искажений в данных, что требует разработки нечувствительных к шуму архитектур и методов обучения, способных эффективно работать с неидеальными входными данными. В данной работе мы исследуем возможности различных архитектур нейронных сетей и функций потерь для обработки реальных данных на примере датасета CIFAR-10N, который включает как чистые, так и зашумленные изображения.

CIFAR-10N [25] представляет собой модификацию классического датасета CIFAR-10, где часть данных intentionally искажена для имитации реальных условий, в которых модели могут сталкиваться с шумом, артефактами или ошибками аннотации. В рамках исследования мы рассматриваем три архитектуры: классическую сверточную нейронную сеть (CNN) с настройками, соответствующими экспериментам Xia и соавторов [26], глубокую residual-сеть (ResNet50) и трансформер для изображений (Vision Transformer, ViT). Для обучения моделей используются три вида функций потерь: кросс-энтропия (Cross-Entropy, CE), бинарная кросс-энтропия (Binary Cross-Entropy, B) и функция потерь, учитывающая шум в данных (Noise-Aware Loss, N).

Цель работы — провести сравнительный анализ и эксперименты эффективности выбранной архитектуры и функций потерь в условиях работы с чистыми и зашумлёнными данными, а также выявить наиболее устойчивые подходы для задач классификации в реальных сценариях. Результаты исследования могут быть полезны для разработки моделей, способных сохранять высокую точность даже в условиях значительного уровня шума.

2 Проведённая работа

Обучение с использованием аннотаций реальных людей. Обучение моделей с зашумлёнными метками является важной проблемой в глубинном обучении, особенно при работе с реальными аннотациями, содержащими ошибки. Датасет CIFAR-10N был создан специально для изучения этого аспекта, так как он включает реальные ошибки, допущенные аннотаторами, что делает задачу борьбы с шумом особенно актуальной.

Были рассмотрены две статьи, имеющие непосредственное отношение к данной теме: [26] и [22]. Первая статья предлагает детальное исследование проблемы обучения с зашумлёнными метками, используя датасет CIFAR-10N, и анализирует влияние различных стратегий очистки данных и методов дообучения модели для повышения её устойчивости к шуму. Авторы рассматривают несколько методов, таких как использование функций потерь, устойчивых к шуму, включая Forward Loss Correction, Backward Loss Correction и Symmetric Cross-Entropy, которые корректируют градиенты в зависимости от вероятности ошибок аннотации. Также исследуются методы фильтрации данных. Авторы показывают, что наиболее эффективными стратегиями являются комбинация устойчивых к шуму функций потерь и методов фильтрации, а также использование частично контролируемого обучения, позволяющего повысить точность модели даже при высоком уровне шума.

Вторая статья представляет обзор методов обучения с зашумлёнными метками, классифицируя их на три основные группы: методы обработки данных (data-centric approaches), методы модификации модели (model-centric approaches) и методы, основанные на функциях потерь (loss-centric approaches). Среди рассмотренных методов в категории обработки данных выделяются Bootstrapping и Label Smoothing, позволяющие корректировать метки за счёт усреднения предсказаний модели, а также Meta-Learning Approaches, которые обучают модель находить оптимальные коэффициенты для борьбы с шумом. В категории функций потерь выделяются Generalized Cross-Entropy, которая снижает влияние выбросов в метках за счёт сочетания квадратичной и стандартной кросс-энтропии, а также Self-Adaptive Training, позволяющий модели самостоятельно определять, какие примеры считать шумными, и адаптировать процесс обучения. Авторы данной статьи предлагают систематизированный обзор методов, который помогает выбрать оптимальную стратегию в зависимости от уровня шума и доступных данных. Однако данная работа носит в основном

теоретический характер и не предоставляет конкретных экспериментов с CIFAR-10N, что делает её менее прикладной по сравнению с первой.

Учитывание неопределённостей. Одним из подходов к оценке неопределенности регрессионных моделей является гетероскедастическая (heteroscedastic) регрессия, которая учитывает как среднее значение переменной, так и дисперсию [18, 20]. Таким образом, модель обучается прогнозированию средних значений и дисперсий, и неопределенность прогнозов модели может быть оценена с использованием значений дисперсии. К счастью, модели классификации также могут использовать squared error (SE) loss. Хо и Белкин [10] продемонстрировали, что визуальные модели, основанные на SE и CE функциях потерь, близки по точности. Тем не менее, SE loss требует чуть больше эпох обучения. Кендалл и Гал [11] рассмотрели два типа неопределенности: алеаторическая (неопределенность данных) и эпистемологическая (неопределенность модели), и предложили два подхода к оценке неопределенности. Кендалл и Гал [11] заявили, что примеры, основанные на отсутствии данных, нельзя отождествлять с алеаторической неопределенностью. Авторы также предложили подход, сочетающий алеаторическую и эпистемологическую неопределенности. Дальнейшая работа ван Амерсфорта и других [23] посвящена методу количественной оценки детерминированной неопределенности. Предлагаемая модель изучает положения центроидов классов и обучает ядра оценивать расстояние между входной выборкой и центроидами, что позволяет модели логического вывода распознавать выборку, в которой отсутствуют данные, как неопределенную. Сенсой и другие [21] разработали теорию доказательной базы и представили предсказания модели в виде распределения плотности Дирихле по выходным данным (softmax outputs), а также предложили новую функцию потерь. Коллиер и другие [3] предложили метод обучения глубинных классификаторов в условиях гетероскедастического шума меток (label noise). Метод основан на softmax temperature tuning, которая позволяет контролировать соотношение смещения и дисперсии.

Ансамблирование, test-time augmentation и label smoothing. Ашуха и другие [1] продемонстрировали, что многие методы создания ансамблей моделей эквивалентны ансамблю из нескольких независимо обученных сетей с точки зрения производительности тестирования. Test-time augmentation - это метод, который улучшает производительность модели с помощью усреднения прогнозов [14]. Вероятно, самыми простыми способами повысить устойчивость моделей к шуму в метках является label smoothing [24] и аугментирование данных [19].

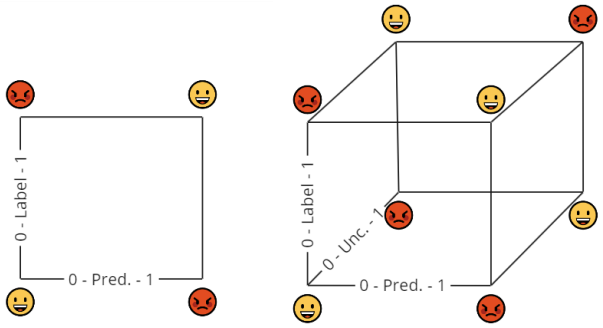
Оценка неопределённости данных на практике Искажённые входные данные [13] и искажённые метки [27], распределения в области определения и вне её [15, 11, 3] являются одними из полюсов исследований в области оценки неопределённости данных. Типичными тестами моделей на практике является использование общедоступных наборов данных с повреждёнными (зашумлёнными) метками на этапах обучения и валидации, но с чистыми метками на этапе тестирования [26, 27]. Существует ряд методов, направленных на выявление входных сэмплов с неправильными метками и удаление [3, 26, 27] или занижение веса этих сэмплов [11, 4]. Хан и другие [8] заявили, что модели сначала изучают данные с чистыми метками, а затем с зашумлёнными, и предложили новую парадигму под названием co-teaching с обучением двух сетей.

3 Методология

В общем, рассмотрим модель $\mathbf{f}[\mathbf{x}, \mathbf{w}]$ параметризованную по весам $\mathbf{w}$, которая сначала сопоставляет входные значения $\mathbf{x}$ к логитам $\mathbf{z}$, а затем гипотезу $\mathbf{h}$, которая аппроксимирует ground truth $\mathbf{y}$. The negative log-likelihood minimization [18, 2, 6] позволяет формализовать следующие функции потерь с учетом неопределенности для задач подбора и классификации с использованием различных типов распределений для выходных данных моделей.

3.1 B-loss

В этом параграфе представлена интерпретация модели бинарной классификации, основанной на минимизации *uncertainty-aware negative log-likelihood с распределением Бернулли* (B-model, B-loss). Предлагаемая модель обучена таким образом, что истинные прогнозы становились определёнными, а ложные прогнозы, если они имеют место, - неопределёнными (см. fig. 1). Бинарный классификатор оценивает значение определённости $c \in (0, 1)$, что является основной задачей в предлагаемой формализации. Кроме того, классификатор оценивает и усиливает сходство δ между гипотезой $\mathbf{h}$ и ground truth $\mathbf{y}$, что составляет вторую задачу в предлагаемой формализации.



(a) (b)

Рис. 1: Основа бинарной классификации с BCE loss (a) и предложенный бинарный B-loss eq. (4) (b) по отношению к значениям выходных данных модели (неопределённости $u = 1 - c$, предсказания h), и метки y.

Бинарная классификация. Рассмотрим i^{th} сэмпл и модель с логитами $z^{(i)} = \left[z^{(i)}_{pred}, z^{(i)}_{cert}\right]$ которые соответствуют предсказанию $h^{(i)} = \sigma(z^{(i)}_{pred})$, и достоверность $c_i = \sigma(z^{(i)}_{cert})$, связанную с предсказанием, соответственно. Затем сравним предсказание $h^{(i)}$ и данную метку $y^{(i)}$, используя метрику скалярного произведения $\delta_i = y^{(i)} h^{(i)}$, и сопоставим эту метрику как псевдомаркировку бинарной оценки неопределенности с параметрами массовой функции вероятности Бернулли [18]:

$$p_i = p(\delta_i | c_i) = \begin{cases} 1 - c_i & \text{if } \delta_i \to 0, \\ c_i & \text{if } \delta_i \to 1, \end{cases} \tag{1}$$

где $\delta_i \in (0, 1)$ - сглаженная псевдомаркировка, которая характеризует сходство между меткой и предсказанием.

eq. (1) представляет собой дискретное распределение вероятности для случайной величины, которая принимает значение 0 с вероятностью $1 - c_i$, которая является неверным прогнозом, соответствующим неопределенности прогноза, и значением 1 с вероятностью c_i, которая является правильным предсказанием, соответствующим достоверности предсказания. Распределение Бернулли имеет эквивалентную форму степенного закона [18]:

$$p_i = c_i^{\delta_i} (1 - c_i)^{1 - \delta_i}. \tag{2}$$

Для roll-out датасета из m и пар $\{x^{(i)}, y^{(i)}\}$, связанных с выходными данными модели $\{h^{(i)}, c_i\}$, совместная вероятность [2] для заданной функции массы вероятности eq. (2) принимает следующую форму:

$$P(\delta_1, \ldots, \delta_m \mid c_1, \ldots, c_m) = \prod_{i=1}^{m} c_i^{\delta_i} (1 - c_i)^{1 - \delta_i}. \tag{3}$$

Отрицательный логарифм совместной вероятности eq. (3) представляет предложенный uncertainty-aware B-loss для бинарной классификации:

$$\mathcal{L}_{\text{B}} = -\frac{1}{m} \sum_{i=1}^{m} \left[\delta_i \log c_i + (1 - \delta_i) \log(1 - c_i)\right]. \tag{4}$$

Интуиция eq. (4) продемонстрирована на fig. 1. B-loss может быть обобщен для случая многоклассовой классификации.

Мультиклассовая (N-classes) классификация. Рассмотрим i^{th} сэмпл и модель с логитами $\mathbf{z}^{(i)} = [\mathbf{z}^{(i)}_{pred}, z^{(i)}_{cert}]$, которые соответствуют вектору предсказаний $\mathbf{h}^{(i)} = \text{softmax}(\mathbf{z}^{(i)}_{pred})$, $\mathbf{h}^{(i)} \in \mathcal{R}^N$, и достоверность $c_i = \sigma(z^{(i)}_{cert})$ связанную с предсказаниями, соответственно. Затем сравним вектор предсказаний $\mathbf{h}^{(i)}$ и данный one-hot encoded label vector $\mathbf{y}^{(i)}$, используя термины скалярного произведения $\delta^{(i)}_k = y^{(i)}_k h^{(i)}_k$, и сопоставим эти показатели в виде псевдомаркировок с функцией массы вероятности:

$$p_i = \prod_{k=1}^{N} \left(\frac{c_i}{N}\right)^{\delta^{(i)}_k} \left(\frac{1 - c_i}{N}\right)^{1 - \delta^{(i)}_k}, \tag{5}$$

где $\delta_k^{(i)} \in (0,1)$ - сглаженная one-hot encoded псевдомаркировка, которая характеризует сходство между k^{th} компонентами метки и вектором предсказаний, N - количество классов.

Следуя логической последовательности, приведенной в section 3.1 и в [18], совместная вероятность для eq. (5) может быть получена, а затем преобразована в negative log-likelihood (NNL):

$$NLL = -\frac{1}{m}\sum_{i=1}^{m}\left(\cos(\mathbf{h}^i,\mathbf{y}^i)\log\left(\frac{c^{(i)}}{N}\right)+(N-1)\left(1-\cos(\mathbf{h}^i,\mathbf{y}^i)\right)\log\left(\frac{1-c^{(i)}}{N}\right)\right), \tag{6}$$

где $\cos(\mathbf{h}^i,\mathbf{y}^i)$ - сглаженная псевдометка, характеризующая косинусное сходство между двумя N-мерными векторами: вектор предсказания и вектор метки с one-hot encoding.

Наконец, предложенные uncertainty-aware B-loss для N-классовой классификации является расхождением Кульбака-Леберга между двумя распределениями: распределение сглаженных псевдометок one-hot encoding и распределение NNL eq. (6):

$$\mathcal{L}_{\text{B}} = \frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{N}\delta_k^{(i)}\log\delta_k^{(i)} + NLL. \tag{7}$$

где m - количество сэмплов (в батче), N - количество классов, $\delta_i = y^{(i)}h^{(i)}$ являются членами скалярного произведения вектора меток с one-hot encoding и вектора предсказания модели, $c^{(i)}$ - достоверность предсказания (fig. 4b).

3.2 N-loss

Поскольку бинарную классификацию можно рассматривать как частный случай многоклассовой классификации, в этом разделе бинарная классификация пропущена.

Мультиклассовая классификация. Рассмотрим i-th сэмпл и модель с логитами $\mathbf{z}^{(i)} = [\mathbf{z}_{mean}^{(i)}, z_{var}^{(i)}]$, которая сопоставляет параметры многомерного нормального распределения: гипотеза или среднее $\mathbf{h}^{(i)} = \mathbf{z}_{mean}^{(i)}$, которая аппроксимирует ground truth $\mathbf{y}^{(i)}$, и дисперсия $\sigma_{(i)}^2 = \exp(z_{var}^{(i)})$, характеризующая неопределенность гипотезы, $f[\mathbf{x}^{(i)},\mathbf{w}] = [\mathbf{h}^{(i)},\sigma_{(i)}^2]$. Другими словами, предполагается, что условное распределение вероятностей $p = p(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) = p(\mathbf{y}^{(i)}|\mathbf{f}[\mathbf{x}^{(i)},\mathbf{w}])$ имеет вид многомерного нормального распределения, характеризующегося равными дисперсиями (сферическими ковариациями) в N-мерном пространстве [17]:

$$p^{(i)} = \frac{\exp\left(-\frac{\sum_{k=1}^{N}\left(y_k^{(i)}-h_k^{(i)}\right)^2}{2\sigma_{(i)}^2}\right)}{\left(2\pi\sigma_{(i)}^2\right)^{\frac{N}{2}}}, \tag{8}$$

Многомерное нормальное распределение (8) может быть применено к критерию negative log-likelihood с использованием uncertainty-aware negative log-likelihood loss (N-loss) для регрессии [18]:

$$\mathcal{L}_N = \frac{1}{2m}\sum_{i=1}^{m}\left(\sum_{k=1}^{N}\frac{\left(y_k^{(i)}-h_k^{(i)}\right)^2}{\sigma_{(i)}^2} + N\left(s^{(i)}+r\right)\right), \tag{9}$$

где m - число сэмплов (в батче), $\mathbf{y}^{(\text{i})}$, $s^{(i)} = \log\sigma_{(i)}^2$ - логарифмическая дисперсия, $r = \log 2\pi$ - константа.

Последнее слагаемое в (9) представляет собой константу, которой можно пренебречь. Кендалл и другие в [11] рекомендуют обучать модели прогнозированию логарифмических отклонений $s^{(i)} = \log\sigma_{(i)}^2$, потому что она более стабильна численно, чем дисперсия $\sigma_{(i)}^2$ и позволяет избежать потенциального деления на ноль в loss-функции:

$$\mathcal{L}_N = \frac{1}{m}\sum_{i=1}^{m}\left(e^{-s^{(i)}}\sum_{k=1}^{N}\left(y_k^{(i)}-h_k^{(i)}\right)^2 + Ns^{(i)}\right). \tag{10}$$

Таким образом, (10) представляет собой гетероскедастическую регрессионную потерю [18, 11], обобщенный для случая пространства из N измерений. Наше предложение состоит в том, чтобы использовать эту loss-функцию для задач классификации.

Стандартной loss-функцией в задаче мультиклассовой классификации является cross-entropy loss [6, 18, 2]:

$$L_{CE} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{N}y_k^{(i)}\log h_k^{(i)}. \tag{11}$$

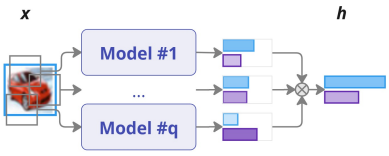


Рис. 2: Ансамбль моделей q делает предсказание $\mathbf{h}$ для q аугментированной копии входного сэмпла $\mathbf{x}$. Продемонстрирована классификация на два класса.

3.3 Ансамблирование

Множество q моделей $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_q$ при разной случайной инициализации веса обучаются с одним и тем же набором данных. Такое агрегирование сокращает переобучение и обеспечивает более надежные оценки за счет усреднения отдельных ошибок модели. [1]. Каждая j^{th} модель предсказывает индекс класса для данного i^{th} входа:

$$\hat{y}^{(i,j)} = \arg\max_k h_k^{(i,j)}, \tag{12}$$

где i, j, k - индексы, которые ссылаются на j^{th} аугментированную версию i^{th} сэмпла, и k^{th} - компонент вектора предсказания или индекса класса, $i \in (1, m), j \in (1, q), k \in (0, N-1)$.

Окончательный предсказанный класс ансамбля определяется *голосованием большинства*, основанном на предсказании каждой индивидуальной модели j^{th}.

$$\hat{y}^{(i)} = \text{mode}\left(\hat{y}^{(i,1)}, \hat{y}^{(i,2)}, \ldots, \hat{y}^{(i,q)}\right). \tag{13}$$

Окончательный предсказанный класс ансамбля также может быть определён с помощью *взвешенных прогнозы, основанных на достоверности* (см. fig. 2). Каждая модель предсказывает класс $\hat{y}^{(i,j)}$ и обеспечивает значение доверия $co^{(i,j)}$ для этого предсказания:

$$co^{(i,j)} = \max_k(h_k^{(i,j)}). \tag{14}$$

Совокупная достоверность для класса k:

$$co_k^{(i)} = \sum_{j=1}^{q} co^{(i,j)} \cdot \mathbb{I}\left(\hat{y}^{(i,j)} = k\right), \tag{15}$$

где $\mathbb{I}(\cdot)$ - это функция индикатора, которая возвращает 1 если условие истинно и 0 в остальных случаях, $co_k^{(i)}$ это суммарная уверенность в классе k среди всех моделей.

Окончательный прогнозируемый класс:

$$\hat{y}^{(i)} = \arg\max_k co_k^{(i)}. \tag{16}$$

Оценка неопределенности в глубоком ансамблировании выводится из дисперсии прогнозов отдельных моделей. Более высокая дисперсия между выходными данными моделей указывает на большую неопределенность, обеспечивая меру эпистемологической неопределенности.

3.4 Метрики

Стандартными показателями классификации для сбалансированных наборов данных являются точность (accuracy), receiver operating characteristic - area under curve (ROC-AUC) [2, 6]. Набор более специфичных показателей, используемых при количественной оценке неопределенности, включает оценку достоверности (см. eq. (14)) [15]: Bries score, энтропия, ожидаемая ошибка калибровки (ECE), negative log-likelihood (NLL), prediction interval coverage probability (PICP), резкость, и другие [16, 7, 5, 9].

Поскольку предлагаемая B-модель (см. eq. (7)) и N-модель (см. eq. (10)) имеют дополнительный выход, могут быть соблюдены следующие дополнительные показатели достоверности:

- $c^{(i)} \in (0, 1)$ для B-модели ;
- $1 - sigm(s^{(i)}) \in (0, 1)$ для N-модели.

Обе вышеперечисленные метрики могут быть использованы в качестве весов в eq. (16), таким образом, взвешенные прогнозы, основанные на достоверности, должны быть выполнены.

Таблица 1: Accuracy (%) моделей, обученных на датасете CIFAR-10N с чистыми и зашумленными метками, а также со сглаживанием меток (LS). Каждая модель была обучена с использованием семи различных начальных значений сида в течение 20 эпох, затем объединена в ансамбль. Средняя точность тестирования отдельных моделей сравнивается с точностью ансамбля с использованием голосования большинства (EMV). Наилучшие результаты выделены жирным шрифтом.

Метод	**Арх-ра**	**#Пар-ры.** (трен. пар-ры)	**LS**	**Accuracy (single model / EMV)%** Чистые	Шумные
Baseline CE loss	9-l.CNN	4.4 M (all)	0.0	$83.87 \pm 0.006/87.41$	$71.55 \pm 0.008/74.57$
Proposed B-loss				$84.24 \pm 0.002/87.78$	$\mathbf{72.62 \pm 0.01/77.01}$
Proposed N-loss				$\mathbf{86.09 \pm 0.005/89.15}$	$72.48 \pm 0.005/75.72$

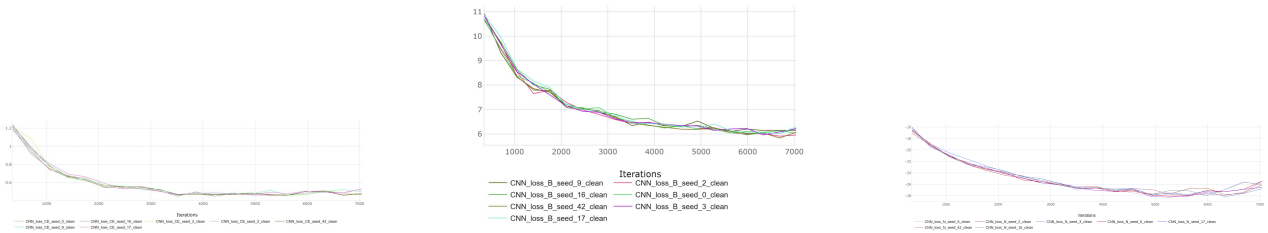
4 Результаты и обсуждение

Датасет CIFAR-10N [25] был разделен на обучающий, валидационный и тестовый наборы в количестве [45000, 5000, 10000] сэмплов, соответственно. Модели были обучены с помощью 9-слойной сверточной нейронной сети [8, 26]. Сеть имеет 4,4 миллиона параметров, которые были случайным образом инициализированы во время обучения. Архитектура CNN и большинство настроек соответствуют экспериментам Xia et al. [26] с небольшими изменениями: модели обучались в течение 20 эпох (200 в оригинальной статье) с использованием оптимизатора Adam с импульсом 0.9, размером батча 128 и с постоянной скоростью обучения 0.001 (в оригинальной работе начальная скорость обучения линейно снижалась до нуля, начиная с 80^{th} эпохи); образцы изображений были преобразованы в тензоры и нормализованы с помощью средних значений [0.491, 0.482, 0.447] и стандартных отклонений [0.247, 0.243, 0.261].

Некоторые экспериментальные настройки могут отличаться от настроек в статье [26]: была применена техника построения модельного ансамбля; обрезка произвольного размера в диапазоне масштаба [0.8, 0.1] и в диапазоне соотношений сторон [0.9, 1.1] была применена ко всем сэмплам во всех наборах в качестве преобразования, таким образом, выборка тестового набора была реализована с использованием увеличения времени тестирования с обрезкой произвольного размера; для вывода были использованы модели с наименьшим валидационным лоссом.

Все эксперименты проводились семь раз со случайными сидами [42, 0, 17, 9, 3, 16, 2]. Затем были представлены среднее значение и стандартное отклонение результатов эксперимента. Множественные прогнозы, полученные с помощью объединения моделей, позволили рассчитать окончательные прогнозы с использованием голосования большинства.

Accuracy была использована [12] как метрика. Полученные результаты не сравнивались с результатами, полученными по последнему слову техники. Последние объединяют множество методов и сложные сетевые архитектуры. Таким образом, сравнение было бы несправедливым.



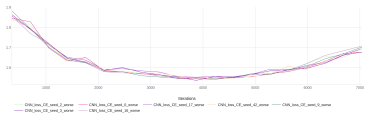
(a) CE-model (b) B-model (c) N-model

Рис. 3: Значения валидационного лосса во время обучения ансамбля моделей за 20 эпох на чистых данных: CE-model (a), B-model (b), and N-model (c).

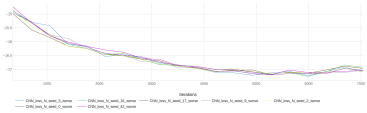
Как видно из графиков, лучшие результаты показывают модели на N-loss с чистыми данными и с B-loss на зашумлённых данных.

5 Этика

LLM-модели использовались в основном для перевода текста как задания, так и сопутствующих статей, а также для небольших правок latex файла.



(a) CE-model



(c) N-model

(b) B-model (отсутствует из-за ошибки ClearML)

Рис. 4: Значения валидационного лосса во время обучения ансамбля моделей за 20 эпох на зашумлённых данных: CE-model (a), B-model (b) (отсутствует из-за ошибки ClearML), and N-model (c).

6 Заключение

В данной работе мы исследовали проблему обучения моделей компьютерного зрения на данных с шумными метками, используя датасет CIFAR-10N, который содержит как чистые, так и зашумленные аннотации, основанные на реальных человеческих ошибках. Были рассмотрены архитектура CNN, а также затронуты теоретически ResNet50 и ViT, и три функции потерь: кросс-энтропия (CE-loss), бинарная кросс-энтропия (B-loss) и N-loss, что позволило провести всесторонний анализ устойчивости моделей к шуму. Результаты показали, что зашумленные данные значительно снижают точность моделей.

Проблема шумных меток является одной из ключевых в машинном обучении, особенно в условиях, когда данные размечаются людьми или автоматическими системами, подверженными ошибкам. Наше исследование подчеркивает, что методы, разработанные для искусственного шума, не всегда эффективны на реальных данных, таких как CIFAR-10N. Это указывает на необходимость разработки более гибких и адаптивных подходов, которые учитывают сложную природу шума в реальных условиях.

Одним из ключевых выводов работы является то, что шум в данных не является однородным. Например, классы с визуально схожими признаками (такие как "кошка"и "собака") имеют более высокий уровень шума, что требует разработки методов, способных учитывать контекст и семантику данных. Это открывает новые возможности для исследований в области semi-supervised обучения и методов, основанных на самообучении, где модель может самостоятельно корректировать ошибки в метках.

Хотя наша работа не решает глобальных проблем, она вносит вклад в развитие нечувствительных методов машинного обучения, которые могут быть применены в реальных задачах, таких как медицинская диагностика, автономные транспортные средства и анализ спутниковых изображений. Устойчивые к шуму модели способны повысить надежность и безопасность таких систем, что в конечном итоге может улучшить качество жизни людей.

Список литературы

[1] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*, 2020.

[2] C. M. Bishop and N. M. Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] M. Collier, B. Mustafa, E. Kokiopoulou, R. Jenatton, and J. Berent. A simple probabilistic method for deep classification under input-dependent label noise. *arXiv preprint arXiv:2003.06778*, 2020.

[4] E. Englesson, A. Mehrpanah, and H. Azizpour. Logistic-normal likelihoods for heteroscedastic label noise, 2023.

[5] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[6] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[7] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. *International Conference on Machine Learning*, pages 1321–1330, 2017.

[8] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels, 2018.

[9] F. Hernandez, L. Bertino, G. Brassington, E. Chassignet, J. Cummings, F. Davidson, M. Drevillon, G. Garric, M. Kamachi, J. M. Lellouche, et al. Probabilistic forecasting in meteorology: A review. *Quarterly Journal of the Royal Meteorological Society*, 141(688):318–350, 2015.

[10] L. Hui and M. Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks, 2021.

[11] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

[12] A. Kumar, P. Liang, and T. Ma. Verified uncertainty calibration, 2020.

[13] E. Mintun, A. Kirillov, and S. Xie. On interaction between augmentations and corruptions in natural corruption robustness, 2021.

[14] D. Molchanov, A. Lyzhov, Y. Molchanova, A. Ashukha, and D. Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation, 2020.

[15] T. Pearce, A. Brintrup, and J. Zhu. Understanding softmax confidence and uncertainty. *CoRR*, abs/2106.04972, 2021. URL `https://arxiv.org/abs/2106.04972`.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] S. Prince. *Computer Vision: Models Learning and Inference*. Cambridge University Press, 2012.

[18] S. J. Prince. *Understanding Deep Learning*. MIT Press, 2023. URL `http://udlbook.com`.

[19] S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[20] M. Seitzer, A. Tavakoli, D. Antic, and G. Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

[21] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty, 2018.

[22] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee. Learning from noisy labels with deep neural networks: A survey, 2022.

[23] J. van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *CoRR*, abs/2003.02037, 2020. URL `https://arxiv.org/abs/2003.02037`.

[24] J. Wei, H. Liu, T. Liu, G. Niu, M. Sugiyama, and Y. Liu. To smooth or not? when label smoothing meets noisy labels. *arXiv preprint arXiv:2106.04149*, 2021.

[25] J. Wei, Z. Zhu, H. Cheng, T. Liu, G. Niu, and Y. Liu. Learning with noisy labels revisited: A study using real-world human annotations, 2022. URL `https://arxiv.org/abs/2110.12088`.

[26] X. Xia, T. Liu, B. Han, M. Gong, J. Yu, G. Niu, and M. Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels, 2021.

[27] Q. Yao, H. Yang, B. Han, G. Niu, and J. Kwok. Searching to exploit memorization effect in learning from corrupted labels, 2020.