УДК 004.85

ИССЛЕДОВАНИЕ ПОВЕДЕНИЯ ПРЕДСТАВЛЕНИЙ ИЗОБРАЖЕНИЙ, НЕ ПРИНАДЛЕЖАЩИХ ИСХОДНОМУ ДОМЕНУ, В СПЕЦИАЛЬНЫХ ПОДПРОСТРАНСТВАХ ФОРМЫ И ТЕКСТУРЫ

И.И. Журавлев, О.А. Милосердов, А.В. Макаренко

На сегодняшний день одной из основных проблем в задачах компьютерного зрения является непредсказуемость ответов сверточных нейронных сетей (СНС) при подаче им на вход изображений, находящихся вне исходного домена. Недавние исследования демонстрируют, что одной из причин некорректной работы СНС в этих задачах является ее смещение в сторону формирования признаков текстуры. В данной работе мы предлагаем новый подход к формированию представлений исходного домена, представляя признаковое пространство изображений в виде двух подпространств формы и текстуры. Затем мы исследуем выходы СНС на данных, не принадлежащих исходному домену.

Ключевые слова: сверточная нейронная сеть, глубокое обучение, форма, текстура, домен, представления изображений.

Введение. Сверточные нейронные сети достигли превосходных результатов в распознавании изображений [1], однако исследования демонстрируют, что при подаче на вход изображения, не принадлежащего *исходному домену* данных, выходы СНС оказываются некорректными [2]. В некоторых случаях семантика новых данных совпадает с семантикой обучающих данных, однако выходы СНС могут говорить об обратном. Аналогично с данными, сущность которых кардинально отличается от обучающих данных: ответы сети могут их отождествлять. Существует множество успешных способов решения задачи адаптации домена для семантически идентичных данных, основанные на дообучении сети [3]. При этом исследования в задачах обнаружения данных, не принадлежащих исходному домену, демонстрируют низкое качество. В данной работе предлагается реализовать метод, формирующий на выходах сети пространства признаков формы и текстуры, затем исследовать его в применении к задаче обнаружения данных вне исходного домена.

Связанные исследования. Одна из причин низкого качества нейросетевых методов при решении задачи обнаружения данных вне исходного домена является качество представления изображений в признаковых пространствах, поэтому многие исследования направленны на обучение качественных энкодеров. Широкую популярность имеют методы, основанные на contrastive learning и metric learning [4-7], которые

уменьшают расстояние между представлениями схожих изображений и отдаляя непохожие. Также ведутся исследования по оценке неопределенности ответов нейросетей [8] с использованием байесовского вывода или использования ансамблей моделей [9] для определения данных вне исходного домена. Однако недавние исследования [10] демонстрируют, что проблема заключается в том, что СНС имеют тенденцию обучаться признакам текстуры, игнорируя информацию о форме. В силу неполноты извлеченных признаков при подаче на вход изображения не из исходного домена возникает неоднозначность.

Форма и текстура. Поскольку в нашей работы мы оперируем такими понятиями, как форма и текстура, необходимо дать им определение, чтобы избежать неоднозначностей при реализации метода и исследовании результатов.

Определение 1. *Форма* объекта – характеристики границы объекта, инвариантные относительно операций поворота, сдвига, масштаба и отражения.

Определение 2. *Текстура* объекта – характеристика объекта, описывающая визуальные паттерны, состоящие из пространственно организованных, повторяющихся объектов или субпаттернов (элементов текстуры), которые имеют характерные для этой текстуры свойства (яркость, цвет, форму, размер, ориентацию).

Метод. Описанная во введении идея формирования подпространств формы S и текстуры T позволит явно штрафовать признаки, формируемые СНС. Схема метода представлена на рис. 1.

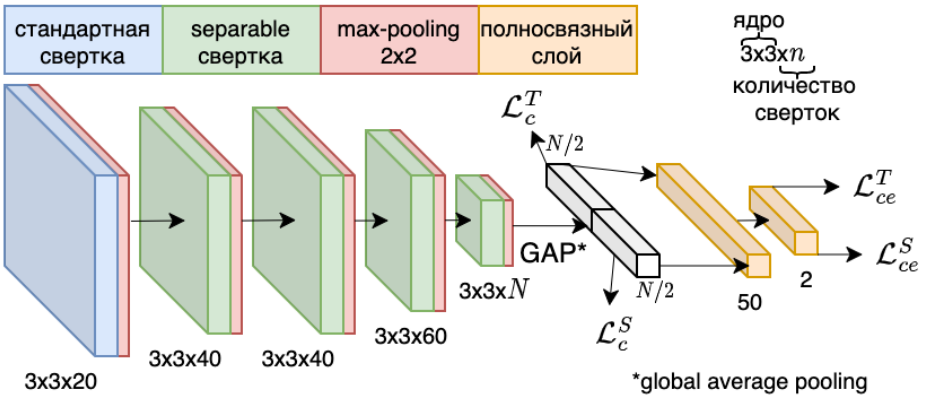

Рис. 1. Схема метода

Формирование подпространств достигается за счет использования функции потерь, описываемой уравнением (1):

$$\mathcal{L}_c = \sum_{i=1}^{K} \left|\left|x_i - c\right|\right|_2^2, \quad (1)$$

где x_i – проекция вектора признаков исходного пространства на подпространство формы или текстуры, c – центр кластера отдельного

класса в S или T, K – размер мини - пакета. Размерность данных подпространств $N/2$. Классификатор обучается посредством кросс-энтропии $\mathcal{L}_{\text{ce}}$ и служит дополнительным штрафом за близость кластеров. Функция потерь (1) позволяет штрафовать векторы признаков в подпространствах S и T за отдаление от общего центра класса. Таким образом, объекты одной формы должны находиться в одном кластере. Аналогично для объектов одной текстуры. Конечный вид функционала эмпирического риска имеет вид:

$$Q = \alpha\mathcal{L}_c^S + \beta\mathcal{L}_c^T + \gamma\mathcal{L}_{ce}^S + \theta\mathcal{L}_{ce}^T, \quad (2)$$

где верхние индексы означают подпространства, которые формирует сеть, коэффициенты при функциях потерь регулируют их вклад в обучение.

Данные. Представляют собой изображения объектов 5-и различных форм в комбинации с 5-ю различными текстурами. Таким образом, обучающее множество содержит 25 различных классов. Классы формы: треугольник, эллипс, месяц, квадрат, крест. Классы текстуры: зебра, гладкость, точки, градиент, шум. К каждому сгенерированному изображению применяются случайным образом преобразования сдвига, поворота и масштабирования. Мы определяем целевой класс формы и текстуры и ожидаем, что предложенный метод позволит сформировать кластеры признаков целевых объектов, исключающие проникновение в них объектов, находящихся вне исходного домена.

Обучение. Для обучения использовался метод оптимизации Adam. Скорость обучения задавалась следующим образом: начальное значение 0.001, уменьшение в 10 раз после 15 и 25 эпох. Общее число эпох 30. Размер мини-пакета 512. Размер обучающего и валидационного множества: 5000 и 1000 изображений на класс соответственно. Коэффициенты α и β равны 1, γ и θ 0.01 до 20-й эпохи и 0.1 с 20-й по 30-ю эпоху. Для оценки качества обучения основной метрикой кластеризации выступала метрика силуэт Sil [11].

Результаты обучения приведены в таблице 1. Sil характеризует разделимость 4-х классов в исходном пространстве признаков, Sil_{ST}^* характеризует внутриклассовую разделимость объектов одной формы и произвольной текстуры или одной текстуры, но произвольной формы. Sil^* характеризует разделимость кластеров целевой и нецелевой формы или текстуры.

Таблица 1

Результаты обучения

N	Sil	Sil_{ST}^S	Sil^S	Sil_{ST}^T	Sil^T
10	0.84±0.01	0.06±0.02	0.84±0.02	0.02±0.03	0.91±0.01

Исходя из значений метрик кластеризации можно предполагать, что метод обучился формировать признаки формы и текстуры объекта. Теперь перейдем к анализу признаков изображений, не принадлежащих исходному домену.

Данные вне исходного домена. Как говорилось во введении, СНС имеют тенденцию обучаться признакам текстуры, поэтому мы разработали алгоритм, позволяющий явно формировать как признаки формы, так и текстуры. В данном разделе мы описываем данные, позволяющие проверить, на сколько СНС способно различать форму и текстуру, принадлежащую исходному домену. При этом мы исследуем кластеры целевой формы и текстуры. Если на вход сети подается объект с целевой формой, принадлежащей исходному домену, то признаки объекта в подпространстве формы должны находиться в одном кластере целевой формы объекта. Аналогично для признаков текстуры.

Для начала мы сгенерируем объекты, сильно отличающиеся как по форме, так и по текстуре от целевого класса (рис. 2).

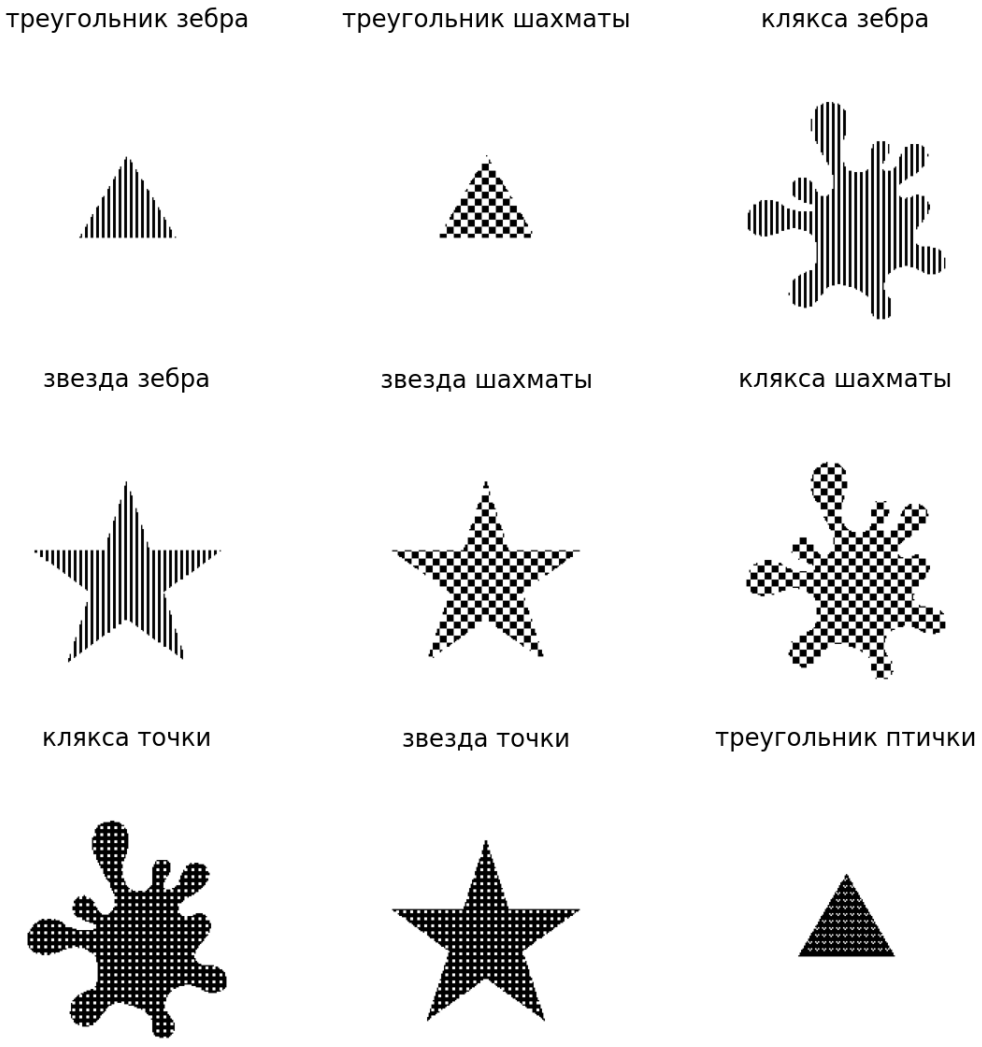

Рис. 2. Изображения вне исходного домена

Данные изображения подаются на вход сети, а затем для полученных векторов признаков формы и текстуры производится оценивание расстояния до кластера целевой формы и текстуры в соответствующих подпространствах. Поскольку кластеры могут быть произвольной формы, оценка производится на основе медианного евклидового расстояния до 10 - и ближайших соседей. На рис. 3 показано распределение медианных расстояний для объектов внутри целевых кластеров, а также для объектов, находящихся вне исходного домена. Темно-синим цветом обозначены распределения медианных расстояний для объектов целевых кластеров. В легенде через после двоеточия указаны значения расстояний.

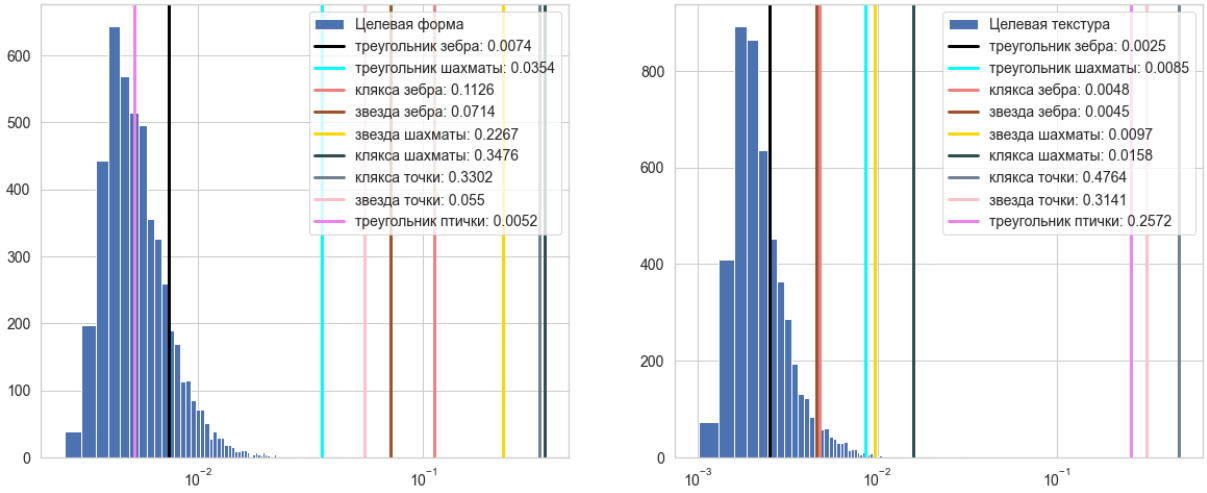

Рис. 3. Медианное расстояние до 10-и ближайших соседей относительно кластеров целевой формы и текстуры

Мы видим, что некоторые нецелевые объекты, например треугольник с текстурой «птички», действительно не попадают в целевой кластер текстуры. Однако объект с текстурой «шахматы» демонстрирует нежелательный результат в обоих подпространствах, поэтому мы проводим еще одну серию опытов, чтобы выявить причину, по которой нейросеть некорректно формирует векторы признаков.

Сформируем объекты целевой формы, но с текстурой, состоящей из паттернов 5-и форм, из которых формировались объекты обучающего множества, будем называть их патчами. Патч имеет 4 размера, поскольку чем больше патч, тем ближе он к объекту данной формы из обучающего множества, что может оказывать влияние на работу СНС. При этом к данным объектам так же, как в обучающем множестве, применяются преобразования сдвига, поворота и масштабирования. Для каждого патча генерируется 20 изображений. В данной работе мы представляем часть результатов, которых достаточно для формулировки выводов по данному исследованию. Изображения представлены на рис. 4.

На рис. 5 приведены диаграммы разброса медианных значений расстояний как для не инвертированных, так и для инвертированных данных. Диаграммы «форма» и «текстура» демонстрируют распределение расстояний между объектами соответствующих целевых кластеров. Исследования расположения признаков относительно кластеров целевой формы и текстуры для двух групп изображений демонстрируют неоднозначность. Желаемый результат достигается в большинстве случаев для инвертированных изображений, в то время как для не инвертированных сеть формирует некорректные признаковые представления.

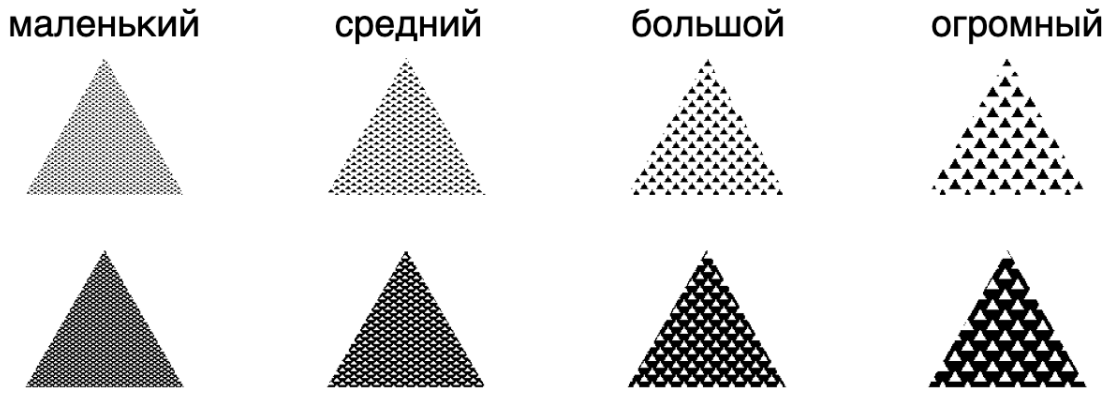

Рис. 4. Верхний ряд – не инвертированные изображения, нижний ряд – инвертированные изображения

Также важно отметить, что доля некорректных признаковых представлений увеличивается при увеличении размера патча. Размер «усов» диаграмм, отображающие 5 и 95 квантили, объясняется степенью искаженности текстуры при масштабировании и повороте.

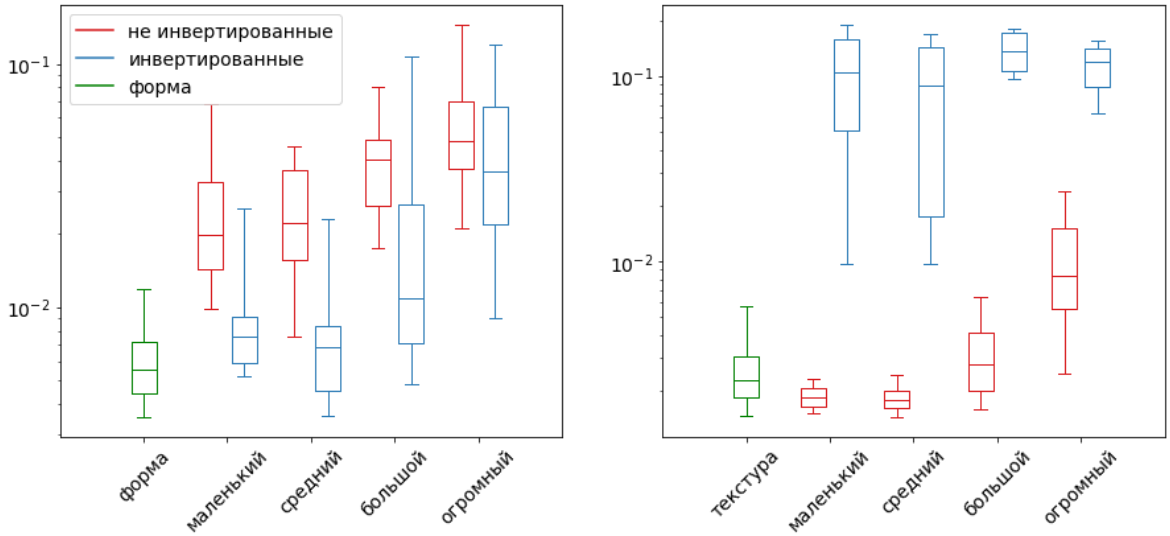

Рис. 5. Диаграммы расстояний для изображений вне исходного домена

Обсуждение результатов. Исходя из проведенного анализа мы видим, что часть объектов с формой, не принадлежащей исходному домену, сеть не относит в кластер целевой формы и текстуры. Это свидетельствует не только о том, что сеть обучилась извлекать отдельно как признаки формы, так и признаки текстуры, но и формировать уникальные представления для целевых классов. Однако наблюдается, что многие объекты с текстурой вне исходного домена попадают в целевые кластеры. Существует неоднозначность работы сети при анализе ее выходов для ряда изображений (см. рис. 4). Мы предполагаем, что данные результаты являются следствием возникших неоднозначностей в обучающем множестве. Граница объекта целевой формы не представляла собой ограниченный контур и геометрическая фигура обрела очертания треугольника за счет расположения элементов текстуры «зебра». В не инвертированных изображениях также отсутствовала четкая граница объекта, что могло привести к некорректной работе СНС. Также негативным фактором мог быть размер элементов текстуры и расстояние между ними. В данной работе использовалась сеть с 5-ю слоями max - pooling, который мог сильно повлиять на извлечение локальных свойств объекта, а именно, на текстуру.

Заключение. В данной работе был предложен и разработан метод, позволяющий решать задачи определения объектов вне исходного домена. Ключевая особенность метода – явное обучение сети формированию двух признаковых пространств формы и текстуры. Результаты обучения свидетельствуют о том, что сеть действительно обучилась извлекать данные признаки. Исследования выходов сети для данных вне обучающего множества также продемонстрировали, что данный метод может использоваться в задаче определения данных вне исходного домена, однако наблюдаются неоднозначности для некоторых типов изображений, связанные с неоднозначностью в исходных данных. Данные неоднозначности будут учтены в последующих исследованиях предложенного метода. Также стоит отметить, что метод продемонстрировал желаемые результаты на данных, которые не допускали в себе неоднозначности

Библиографический список

1. Yu J. et al. CoCa: Contrastive Captioners are Image-Text Foundation Models // arX iv preprint arXiv:2205.01917. 2022.
2. Pan S. J., Yang Q. A survey on transfer learning // IEEE Transactions on knowledge and data engineering. 2009. Vol. 22(10). P. 1345-1359.
3. Redko I. et al. A survey on domain adaptation theory: learning bounds and theoretical guarantees //arXiv preprint arXiv:2004.11829. 2020.
4. Sohn K. et al. Learning and evaluating representations for deep one-class classification // arXiv preprint arXiv:2011.02578. 2020.

5. Schroff F., Kalenichenko D., Philbin J. Facenet: A unified embedding for face recognition and clustering // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 815-823.
6. Wen Y. et al. A discriminative feature learning approach for deep face recognition // European conference on computer vision. Springer: Cham, 2016. P. 499-515.
7. Liu W. et al. Sphereface: Deep hypersphere embedding for face recognition // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 212-220.
8. Gal Y., Ghahramani Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning // international conference on machine learning. PMLR, 2016. P. 1050-1059.
9. Lakshminarayanan B., Pritzel A., Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles //Advances in neural information processing systems. 2017. Vol. 30.
10. Geirhos R. et al. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness //arXiv preprint arXiv:1811.12231. 2018.
11. Rousseeuw P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis //Journal of computational and applied mathematics. 1987. Vol. 20. P. 53-65.