http://dx.doi.org/10.35596/1729-7648-2025-23-3-70-76

УДК 004.93

ОЦЕНКА СХОДСТВА МЕЖДУ НАБОРАМИ ДАННЫХ С ПОМОЩЬЮ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ

А. А. УСАТОВ, А. М. НЕДЗЬВЕДЬ, ГО ЦЗИЖАНЬ

Белорусский государственный университет (Минск, Республика Беларусь)

Аннотация. Рассмотрен подход к определению сходства наборов данных (датасетов) для обучения алгоритмов на примере датасетов с лицами людей. Такой подход позволяет находить похожие наборы данных из разных источников, расширяя детектирование признаков и классов и не нанося серьезного вреда балансировке. Для каждого объекта датасета получено векторное представление (эмбеддинг), затем выполнено сравнение эмбеддингов в обоих датасетах. Эксперименты проводились на примере датасетов с изображениями лиц людей. Для получения эмбеддингов использовалась предобученная сеть ResNet. В процессе исследований один датасет делился на две части, представляющие собой схожие датасеты, затем каждая из частей сравнивалась с отличающимся набором данных. Предлагается новая метрика сходства, которая обладает рядом преимуществ и позволяет находить наиболее похожие датасеты.

Ключевые слова: набор данных, векторное представление, ResNet, сходство датасетов, глубокое обучение.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Благодарность. Авторы выражают благодарность Полине Игоревне Тишковской за помощь в оформлении статьи.

Для цитирования. Усатов, А. А. Оценка сходства между наборами данных с помощью векторных представлений / А. А. Усатов, А. М. Недзьведь, Го Цзижань // Доклады БГУИР. 2025. Т. 23, № 3. С. 70–76. http://dx.doi.org/10.35596/1729-7648-2025-23-3-70-76.

ASSESSING SIMILARITY BETWEEN DATASETS USING VECTOR REPRESENTATIONS

ALEXANDER A. USATOFF, ALEXADER M. NEDZVED, GUO JIRAN

Belarusian State University (Minsk, Republic of Belarus)

Abstract. The article considers an approach to determining the similarity of datasets for training algorithms using datasets with human faces as an example. This approach allows finding similar datasets from different sources, expanding the detection of features and classes and significantly affecting dataset balance. For each dataset object, a vector representation (embedding) was obtained, then the embeddings in both datasets were compared. The experiments were conducted using datasets with images of human faces as an example. To obtain embeddings, a pretrained ResNet network was used. During the research, one dataset was divided into two parts, which were similar datasets, then each of the parts was compared with a different dataset. The new similarity metric is proposed, which has several advantages and allows to find the most similar datasets.

Keywords: dataset, vector representation, ResNet, dataset similarity, deep learning.

Conflict of interests. The authors declare no conflict of interests.

Gratitude. The authors express their gratitude to Polina Igorevna Tishkovskaya for her help in the design of the article.

For citation. Usatoff A. A., Nedzved A. M., Guo Jiran (2025) Assessing Similarity Between Datasets Using Vector Representations. *Doklady BGUIR*. 23 (3), 70–76. http://dx.doi.org/10.35596/1729-7648-2025-23-3-70-76 (in Russian).

Введение

Для решения задач в глубоком обучении [1, 2] необходим большой объем данных. Обучающая выборка должна быть внушительной, чтобы модель определила зависимости и признаки детектируемых объектов и при этом не переобучилась. Как правило, в случае не очень больших наборов данных (датасетов) используется лишь дообучение уже готовой модели, для этого может быть достаточно нескольких тысяч объектов. Тем не менее в случае изображений собрать и разметить даже такой объем данных может быть проблематично. Поэтому сеть часто обучают на публичном датасете, а свой небольшой набор данных используют лишь для валидации модели. Однако, чем меньше публичный набор данных похож на то, что модель будет обрабатывать на практике, тем хуже будет качество модели в реальных условиях.

Совмещение различных датасетов в машинном обучении представляет собой сложную задачу, так как возникает множество факторов, влияющих на качество модели и точность прогнозов. Основная проблема заключается в том, что данные из разных источников часто имеют различия в структуре, формате или распределении признаков. Это может привести к тому, что модель будет обучаться на непоследовательных или даже противоречивых примерах. Различия в масштабах и единицах измерения между данными из разных источников также создают проблемы. Если одни признаки представлены в больших численных значениях, а другие – в меньших, это может исказить важность таких признаков для модели. Даже после нормализации или стандартизации различия в характере данных могут сохраняться, особенно, если они были собраны с использованием разных методологий или инструментов. Это может привести к появлению шума в данных или усилению некоторых биасов, которые снижают обобщающую способность модели.

Еще одна трудность связана с тем, что различные датасеты могут содержать разные уровни полноты или иметь разные стратегии обработки пропущенных значений. В одном наборе данных могут быть удалены записи с пропусками, тогда как в другом такие записи могли быть заполнены средними значениями или другими методами. Это может создать искусственные различия между данными, которые не отражают реальной картины. Таким образом, совмещение данных требует тщательной предварительной обработки, чтобы минимизировать подобные искажения и обеспечить согласованность между источниками информации. Одно из решений перечисленных проблем – нахождение похожих датасетов. Но для этого необходимо определять некую меру сходства наборов данных, чтобы использовать для обучения тот набор, который ближе всего к реальным данным, с которыми модель будет работать на практике.

Цель исследований авторов – повышение эффективности определения сходства датасетов с использованием векторных представлений (эмбеддингов) – объектов фиксированной и, как правило, относительно невысокой размерности, которые содержат наиболее важную информацию об исходных данных. Эмбеддинги несут в себе определенный «смысл» объекта, а близкие эмбеддинги означают, что объекты похожи между собой [3]. Эксперименты проводились на примере датасетов с изображениями лиц людей.

Определение сходства изображений

Математическое определение сходства изображений основывается на вычислении расстояния или сходства между их представлениями в численном пространстве. Один из классических подходов – использование метрик, таких как среднеквадратичная разность (MSE) или коэффициент корреляции Пирсона, которые сравнивают пиксельные значения двух изображений напрямую. Например, MSE вычисляется как среднее значение квадратов разностей между соответствующими пикселями двух изображений, что позволяет количественно оценить различия [4]. Однако такие методы имеют существенные ограничения: они чувствительны к небольшим изменениям, не учитывают более сложные семантические характеристики изображений и игнорируют структурные особенности объектов [5]. Другим подходом является использование гистограмм цветов или текстур для сравнения изображений. В этом случае каждое изображение представляется в виде распределения значений, характеризующих его цветовые или текстурные свойства. Затем применяются метрики, такие как расстояние Хэмминга или Кульбака – Лейблера между двумя распределениями [6]. В данном случае оценка более устойчива к некоторым преобразованиям, но ограничена в восприятии высокоуровневых признаков, включая форму объектов. Кроме того, она чувствительна к шуму или изменению освещенности [7].

Использование эмбеддингов для сравнения изображений – современный и эффективный подход. Векторные представления отражают высокие уровни абстракции, связанные с содержанием изображений, и позволяют сравнивать их с использованием таких метрик, как евклидово расстояние или косинусная близость [8].

Особенности применения векторных представлений

При нейросетевом методе получения эмбеддингов для изображений обычно используются сверточные нейронные сети, принцип устройства которых продемонстрирован на рис. 1.

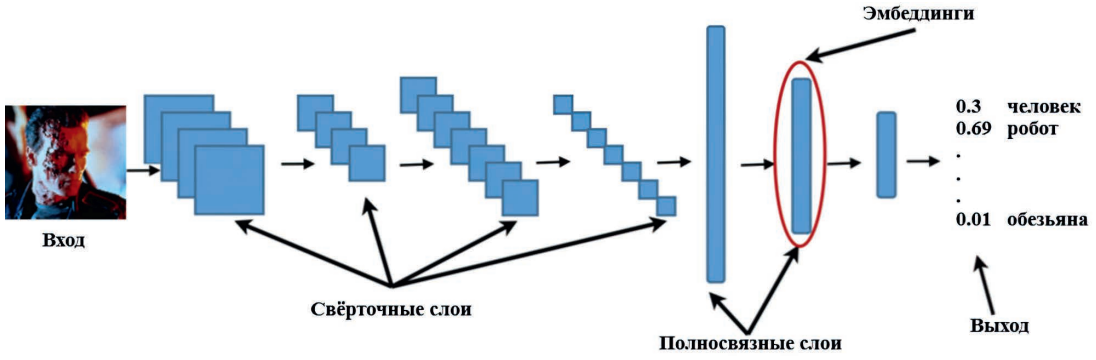


Рис. 1. Получение векторных представлений с помощью нейронной сети
Fig. 1. Obtaining vector representations using a neural network

Сверточные слои являются ключевыми для сверточной нейронной сети. Они выполняют операцию свертки, применяя фильтры (ядра) к исходному изображению, чтобы выделить важные признаки. Эти признаки могут включать края, углы, текстуры и другие визуальные элементы. После применения фильтров результаты передаются через функции активации (например, ReLU) и слои субдискретизации, которые уменьшают размерность и обобщают признаки. Обычно применяется несколько сверточных слоев, где каждый последующий извлекает признаки более высокого уровня. Например, на первом сверточном слое могут извлекаться такие признаки, как границы, углы, текстуры, на втором – более сложные формы – изгибы бровей или колеса автомобилей, на третьем сверточном слое могут выявляться уже конкретные объекты или их крупные части. За сверточными слоями следуют полносвязные слои, которые выполняют классификацию на основе извлеченных сверточными слоями признаков. Последний полносвязный слой обычно содержит столько нейронов, сколько классов в задаче классификации, а функция активации softmax преобразует выходные данные в вероятности каждого класса, что и является выходом нейронной сети. Эмбеддингами обычно являются выходы одного из полносвязных слоев сети. Эти признаки затем сравниваются с помощью таких метрик, как косинусное сходство или манхэттенское расстояние, чтобы определить степень их схожести [9].

Воспользуемся вышеупомянутым свойством, что похожие изображения имеют близкие эмбеддинги, чтобы считать сходство изображений. Будем считать сходство изображений как косинусное сходство между их эмбеддингами согласно формуле

$$cosine(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}, \quad (1)$$

где A, B – векторы, между которыми считается косинусное сходство; n – размер векторов A и B.

Формула (1) – один из самых распространенных способов вычисления близости эмбеддингов, который используется, например, в модели word2vec [10]. Для определения сходства датасетов проводилось сравнение содержащихся в них элементов и их распределения:

– для каждого объекта из первого датасета находили самый близкий объект во втором датасете, затем наоборот;

– для каждого объекта первого датасета считали близость со всеми объектами второго датасета, затем усредняли.

Первый вариант не учитывает плотность распределения объектов в признаковом пространстве, поэтому для оценки сходства использовалось второе определение, где оценка выполнялась по формуле

$$datasets\_similarity(D1, D2) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} cosine(D1_i, D2_j)}{nm}, \quad (2)$$

где $D1$, $D2$ – датасеты, сходство которых вычислялось; *cosine* – косинусное сходство, описанное в (1); n, m – размер датасетов $D1$ и $D2$ соответственно.

Такой подход имеет важное преимущество: эмбеддинги учитывают не только низкоуровневые признаки, но и высокоуровневые семантические характеристики, что делает их гораздо более релевантными для оценки человеческой перцепции сходства. Исследования показывают, что эмбеддинги, полученные с помощью предобученных CNN, демонстрируют высокую корреляцию с результатами тестов, основанных на субъективной оценке людей [11].

Использование объектов при сравнении наборов данных

Для проверки работоспособности предлагаемого подхода оценки сходства между датасетами проводились эксперименты на датасетах с фотографиями лиц людей. Чтобы получить эмбеддинги таких изображений, использовалась предобученная сеть ResNet18 [12] без слоя классификации. В экспериментах применялось три датасета: два – очень близкие, а третий несколько от них отличался. Такая организация данных позволила проверить адекватность оценки на основе эмбеддингов, поскольку сходство между первыми двумя датасетами было больше, чем их сходство с третьим. Для получения двух максимально схожих между собой наборов данных один датасет разбивался на два тестовых набора данных.

Поскольку в датасетах для глубокого обучения обычно минимум десятки тысяч изображений, а для статистической значимости результатов необходимо провести эксперименты много раз, применялись подвыборки размером 100 изображений. Кроме того, поиск для каждого изображения наиболее близкого выполнялся квадратичное время, что затрудняло использование больших подвыборок. Для эксперимента наборы данных с идентичным содержанием, включающим объекты одного типа в одинаковой проекции и при одинаковых условиях съемки, определялись как очень похожие датасеты, а наборы данных объектов одного типа, но в разных условиях съемки и в разном положении, – как похожие датасеты. Эксперименты проводились следующим образом:

1) использовалось два датасета – $D1$ и $D2$. Датасет $D1$ делился на две части $D1\_1$ и $D1\_2$, которые представляли собой очень похожие наборы данных. Датасет $D2$ имел небольшие отличия от $D1$ (ниже приведены примеры и описание отличий для конкретных датасетов, но в общем случае это не принципиально);

2) далее 100 раз повторялись следующие действия:

a) из датасетов $D1\_1$, $D1\_2$ и $D2$ брались выборки размером 100 элементов – $d1\_1$, $d1\_2$ и $d2$ соответственно;

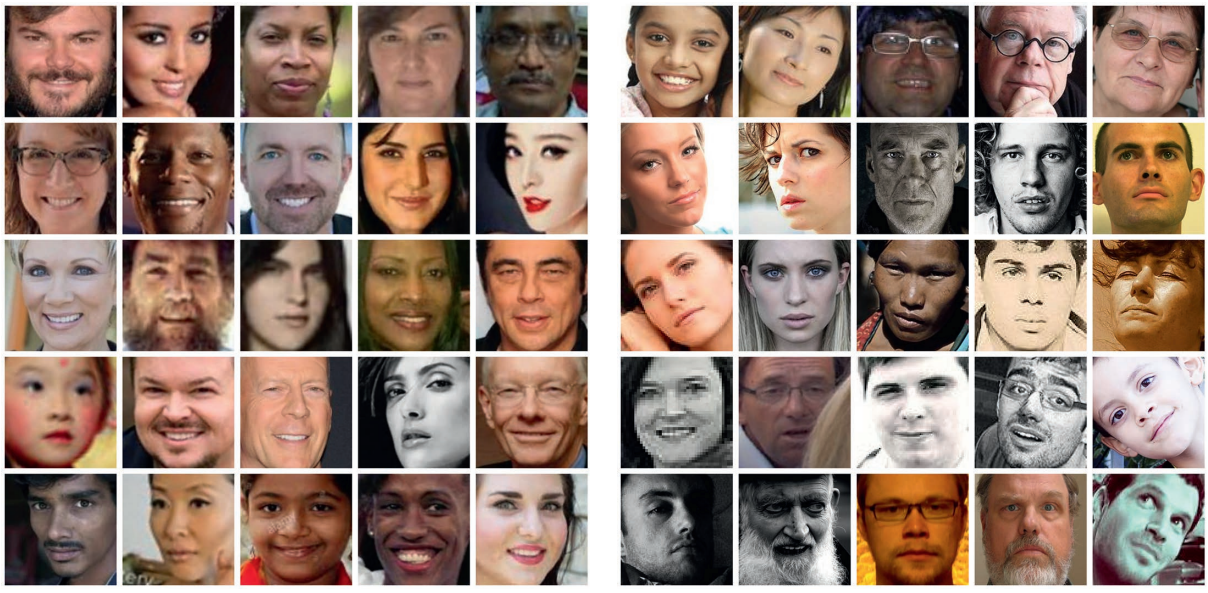
b) по формуле (2) считалось сходство между выборками $\{d1\_1, d1\_2\}$, $\{d1\_1, d2\}$ и $\{d1\_2, d2\}$;

3) по результатам вычислений по пункту 2b строились гистограммы сходства между всеми датасетами (точнее, их оценки по выборкам) и вычислялась точность определения более похожего датасета.

На рис. 2, *a* показан фрагмент набора изображений для классификации лиц. Этот набор делился на два, представлявшие собой очень похожие датасеты. В качестве не очень похожего датасета были использованы картинки из набора данных для определения ключевых точек лица (рис. 2, *b*), где присутствуют различные искажения – контрастные тени, засветки, повернутые изображения и т. д. На рис. 3 приведено распределение косинусной меры сходства для различных датасетов.

В ходе экспериментов рассчитанное в соответствии с предложенным подходом сходство между датасетами $D1\_1$ и $D1\_2$ всегда было больше, чем сходство между $D1\_1$ и $D2$ и датасетами $D1\_2$ и $D2$, то есть:

$$\forall d1\_1 \in D1\_1, d1\_2 \in D1\_2, d2 \in D2$$
$$datasets\_similarity(d1\_1, d1\_2) > datasets\_similarity(d1\_1, d2)$$
$$datasets\_similarity(d1\_1, d1\_2) > datasets\_similarity(d1\_2, d2).$$



a *b*

Рис. 2. Примеры изображений: *a* – датасетов, которые делились на два; *b* – несколько отличающегося датасета
Fig. 2. Examples of images: *a* – datasets that were divided into two; *b* – several different datasets

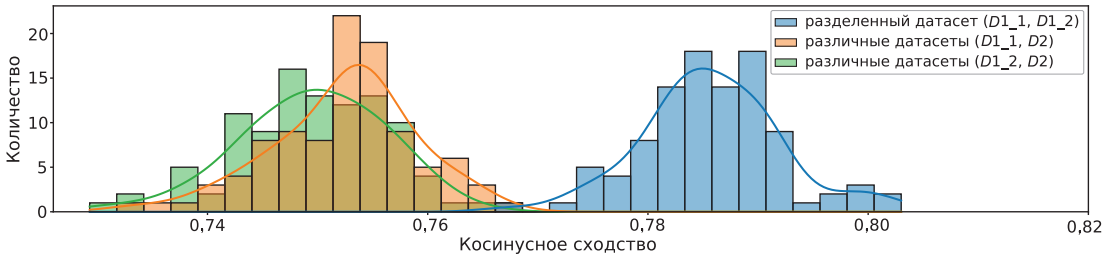


Рис. 3. Распределение косинусной меры сходства для различных датасетов
Fig. 3. Distribution of cosine similarity measure for different datasets

Одной из популярных метрик определения сходства изображений является MSE. Предложенный в статье подход превосходит метод оценки расстояния между датасетами на основе MSE. Для уменьшения влияния случайностей при сравнении методов эксперименты с применением MSE выполнялись на тех же подвыборках изображений, что и с использованием эмбеддингов. На рис. 4 приведено распределение среднеквадратичной разности для различных датасетов.

Из рис. 4 видно, что в части случаев оценка с использованием MSE дает неверный результат. Кроме того, оценка получается неустойчивой, поскольку датасеты *D*1_1, *D*1_2 являются частями одного и того же датасета, но их расстояния от *D*2 сильно отличаются, чего не наблюдается при оценке сходства с использованием эмбеддингов. При использовании MSE точность определения более схожего датасета составила 0,975, в то время как применение формулы (2) позволило полностью исключить ошибки, обеспечив точность, равную единице. Это подтверждает эффективность предложенной метрики в сравнении наборов данных

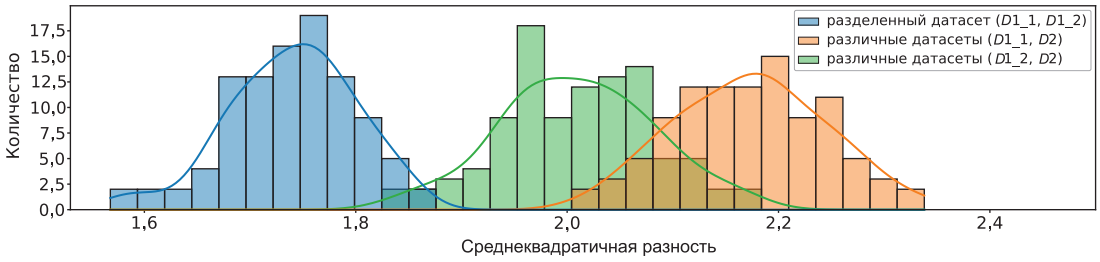


Рис. 4. Распределение MSE для различных датасетов
Fig. 4. Distribution of MSE for different datasets

Заключение

1. Рассмотрен подход к определению сходства набора данных (датасетов) на примере датасетов с изображениями лиц людей. Для получения векторных представлений использовалась предобученная сеть ResNet18 без слоя классификации. Предложена эффективная метрика сходства датасетов, которая позволяет определять меру сходства наборов данных и выбирать датасет, наиболее близкий к образцу.

2. Качество оценки, основанной на векторных представлениях, зависит от архитектуры модели и данных, на которых она была обучена, особенно если модель обучалась на ограниченном наборе категорий [13]. Однако использование векторных представлений остается одним из самых перспективных направлений для решения задачи определения сходства изображений благодаря способности выявлять сложные паттерны и отношения между объектами.

3. Для специфических данных, таких как снимки с дрона или медицинские изображения, имеет смысл использовать дообученную сеть из той же области, поскольку датасет ImageNet [14], на котором была обучена сеть ResNet, во многом отличается от узкоспециализированных наборов данных [15]. Предложенную метрику можно использовать не только для определения сходства датасетов, но и, например, для расширения набора данных путем добавления в него похожих объектов из других датасетов.

Список литературы

1. Ивахненко, А. Г. Кибернетические предсказывающие устройства / А. Г. Ивахненко, В. Г. Лапа. Киев: Акад. наук Укр. ССР, 1965.
2. Gradient-Based Learning Applied to Document Recognition / Y. Lecun [et al.] // Proceedings of the IEEE. 1998. Vol. 86, Iss. 11. P. 2278–2324.
3. Label-Embedding for Image Classification / Z. Akata [et al.] // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2015. Vol. 38, No 7. P. 1425–1438. DOI: 10.1109/TPAMI.2015.2487986.
4. Image Quality Assessment: From Error Visibility to Structural Similarity / Z. Wang [et al.] // IEEE Transactions on Image Processing. 2024. Vol. 13, No 4. P. 600–612. DOI: 10.1109/TIP.2003.819861.
5. Rubner, Y. The Earth Mover's Distance as a Metric for Image Retrieval / Y. Rubner, C. Tomasi, L. J. Guibas // International Journal of Computer Vision. 2000. Vol. 40, No 2. P. 99–121. DOI: 10.1023/A:1026543900054.
6. Lin, J. Divergence Measures Based on the Shannon Entropy / J. Lin // IEEE Transactions on Information Theory. 1991. Vol. 37, Iss. 1. P. 145–151. DOI: 10.1109/18.61115.
7. Swain, M. J. Color Indexing / M. J. Swain, D. H. Ballard // International Journal of Computer Vision. 1991. Vol. 7, No 1. P. 11–32.
8. Simonyan, K. Very Deep Convolutional Networks for Large-Scale Image Recognition / K. Simonyan, A. Zisserman // arXiv.1409.1556. 2014. Vol. 1.
9. Self-Similarity Guided Probabilistic Embedding Matching Based on Transformer for Occluded Person Re-Identification / Y. Pang [et al.] // Expert Systems with Applications. 2024. Vol. 237. https://doi.org/10.1016/j.eswa.2023.121504.
10. Efficient Estimation of Word Representations in Vector Space / T. Mikolov [et al.] // arXiv:1301.3781. 2013. http://arxiv.org/abs/1301.3781.
11. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric / R. Zhang [et al.] // arXiv:1801.03924. 2023. https://doi.org/10.48550/arXiv.1801.03924.
12. Deep Residual Learning Forimage Recognition / K. He [et al.] // arXiv:1512.03385. 2015. https://doi.org/10.48550/arXiv.1512.03385.
13. Learning Transferable Visual Models from Natural Language Supervision / A. Radford [et al.] // arXiv:2103.00020. 2021. https://doi.org/10.48550/arXiv.2103.00020.
14. Imagenet: A Large-Scale Hierarchicalimage Database / Jia Deng [et al.] // 2009 IEEE Conference on Computer Vision and Pattern Recognition. P. 248–255.
15. Недзьведь, А. М. Анализ изображений для решения задач медицинской диагностики / А. М. Недзьведь, С. В. Абламейко. Минск: Объедин. ин-т проблем информ. Нац. акад. наук Беларуси, 2012.

Поступила 10.12.2024 Принята в печать 25.04.2025

References

1. Ivakhnenko A. G., Lapa V. G. (1965) *Cybernetic Predictive Devices*. Kyiv, Academy of Sciences of the Ukrainian SSR (in Russian).
2. Lecun Y., Bottou L., Bengio Y., Haffner P. (1998) Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*. 86 (11), 2278–2324.
3. Akata Z., Perronnin F., Harchaoui Z., Schmid C. (2015) Label-Embedding for Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 38 (7), 1425–1438. DOI: 10.1109/TPAMI.2015.2487986.

4. Wang Z., Bovik A. C., Sheikh H. R., Simoncelli E. P. (2024) Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*. 13 (4), 600–612. DOI: 10.1109/TIP.2003.819861.
5. Rubner Y., Tomasi C., Guibas L. J. (2000) The Earth Mover's Distance as a Metric for Image Retrieval. *International Journal of Computer Vision.* 40 (2), 99–121. DOI: 10.1023/A:1026543900054.
6. Lin J. (1991) Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*. 37 (1), 145–151. DOI: 10.1109/18.61115.
7. Swain M. J., Ballard D. H. (1991) Color Indexing. *International Journal of Computer Vision*. 7 (1), 11–32.
8. Simonyan K., Zisserman A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv.1409.1556*. 1.
9. Pang Y., Zhang H., Zhu L., Liu D., Liu L. (2024) Self-Similarity Guided Probabilistic Embedding Matching Based on Transformer for Occluded Person Re-Identification. *Expert Systems with Applications*. 237. https://doi.org/10.1016/j.eswa.2023.121504.
10. Mikolov T., Chen K., Corrado G., Dean J. (2013) Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*. http://arxiv.org/abs/1301.3781.
11. Zhang R., Isola P., Efros A. A., Shechtman E., Wang O. (2023) The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv:1801.03924*. https://doi.org/10.48550/arXiv.1801.03924.
12. He K., Zhang X., Ren S., Sun J. (2015) Deep Residual Learning for Image Recognition. *arXiv:1512.03385*. https://doi.org/10.48550/arXiv.1512.03385.
13. Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., et al. (2021) Learning Transferable Visual Models from Natural Language Supervision. *arXiv:2103.00020*. https://doi.org/10.48550/arXiv.2103.00020.
14. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, et al. (2009) ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255.
15. Nedzved A., Ablameyko S. (2012) *Image Analysis for Tasks of Medical Diagnostic*. Minsk, United Institute of Informatics Problems of the National Academy of Sciences of Belarus (in Russian).

Received: 10 December 2024 Accepted: 25 April 2025

Вклад авторов

Усатов А. А. провел эксперименты и подготовил рукопись статьи.

Недзьведь А. М. осуществил постановку задачи для проведения исследования, дал рекомендации по выполнению экспериментов и написанию статьи.

Го Цзижань предложил возможные применения предложенного подхода, принимал участие в подготовке рукописи статьи.

Authors' contribution

Usatoff A. A. conducted experiments and prepared the manuscript of the article.

Nedzved A. M. set the task for the research, gave recommendations on conducting the experiments, and writing the article.

Guo Jiran suggested possible applications of the proposed approach, participated in the preparation of the manuscript of the article.

Сведения об авторах

Усатов А. А., магистр, асп. каф. информационных систем управления, Белорусский государственный университет (БГУ)

Недзьведь А. М., д-р техн. наук, доц., зав. каф. информационных систем управления, БГУ

Го Цзижань, асп. каф. информационных систем управления, БГУ

Адрес для корреспонденции

220030, Республика Беларусь,
Минск, просп. Независимости, 4, к. 501
Белорусский государственный университет
Тел.: +375 17 209-52-45
E-mail: alexander.usatoff@gmail.com
Усатов Александр Андреевич

Information about the authors

Usatoff A. A., Master's, Postgraduate at the, Department of Information Management Systems, Belarusian State University (BSU)

Nedzved A. M., Dr. Sci. (Tech.), Associate Professor, Head of the Department of Information Management Systems, BSU

Guo Jiran, Postgraduate at the Department of Information Management Systems, BSU

Address for correspondence

220030, Republic of Belarus,
Minsk, Nezavisimosti Ave., 4, Off. 501
Belarusian State University
Tel.: +375 17 209-52-45
E-mail: alexander.usatoff@gmail.com
Usatoff Alexander Andreevich