DOI: 10.25586/RNU.V9187.21.01.P.111 УДК 004.032.26

К.О. Гнидко, М.А. Еремеев, С.В. Пилькевич, Т.Р. Сабиров*

АНАЛИЗ СТЕПЕНИ ПРИГОДНОСТИ ОТКРЫТЫХ БАЗ ДАННЫХ АННОТИРОВАННЫХ ИЗОБРАЖЕНИЙ ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ НЕЙРОСЕТЕЙ В ЗАДАЧАХ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННО-ПСИХОЛОГИЧЕСКОЙ БЕЗОПАСНОСТИ

Представлена методика проведения анализа степени пригодности открытых баз данных аннотированных изображений для глубокого обучения нейросетей с целью обеспечения информационно-психологической безопасности пользователей сети Интернет. Проведено исследование соответствия модели представления и распознавания визуальных образов сверточными нейронными сетями и особенностей зрительного восприятия этих изображений человеком. Осуществлен сравнительный анализ результатов экспериментальных исследований по визуализации карт признаков обученной сверточной сети и точек фиксации взгляда испытуемых для одних и тех же визуальных стимулов. Предложены подходы к повышению качества результатов подобных исследований в дальнейшем.

Ключевые слова: сверточные нейросети, машинное обучение, информационно-психологическая безопасность, сеть Интернет, аннотированные изображения.

K.O. Gnidko, M.A. Eremeev, S.V. Pilkevich, T.R. Sabirov

ANALYSIS OF THE SUITABILITY OF OPEN DATABASES OF ANNOTATED IMAGES FOR DEEP LEARNING OF NEURAL NETWORKS IN THE TASKS OF INFORMATION-PSYCHOLOGICAL SECURITY

The article presents a methodology for analyzing the degree of suitability of open databases of annotated images for deep learning of neural networks in order to ensure information and psychological security of Internet users. The study of the correspondence between the model of representation and recognition of visual images by convolutional neural networks and the peculiarities of visual perception of these images by humans has been carried out. A comparative analysis of the results of experimental studies on the visualization of maps of signs of the trained convolutional network and points of fixation of the gaze of the subjects for the same visual stimuli is carried out. Approaches are proposed to improve the quality of the results of such studies in the future.

Keywords: convolutional neural networks, machine learning, information and psychological security, the Internet, annotated images.

Вводные замечания

Настоящая работа продолжает исследования авторов в области обеспечения информационно-психологической безопасности потребителей мультимедийного контента [1, 5].

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-29-22064 «Модели и методы выявления и интеллектуальной обработки деструктивного мультимедийного интернет-контента».

Информационная безопасность

Основной целью настоящего этапа исследования являлся анализ пригодности открытых баз аннотированных изображений, доступных в сети Интернет, для построения интеллектуальной системы обеспечения информационно-психологической безопасности пользователей с применением методов глубокого обучения.

В связи с тем, что наиболее эффективной технологией распознавания изображений к настоящему времени являются сверточные нейронные сети (СНС) [4], важной задачей является исследование соответствия модели представления и распознавания визуальных образов СНС и особенностей зрительного восприятия этих изображений человеком. Для решения данной задачи в работе проводится сравнительный анализ результатов экспериментальных исследований по визуализации карт признаков обученной сверточной сети и точек фиксации взгляда испытуемых для одних и тех же визуальных стимулов.

Исходные данные

В качестве исходных данных был использован набор эмоциогенных изображений из открытой международной базы GAPED [2]. Все изображения набора GAPED в процессе подготовки экспериментальных данных были разделены на два класса. В первый класс (520 шт.) вошли файлы изображений, содержащие эмоционально- и когнитивно-негативные визуальные стимулы (змеи, пауки, страдания людей и животных), во второй (210 шт.) – все оставшиеся изображения, обозначенные как нейтральные или позитивные.

Из исходного набора изображений (730 шт.) для каждого из двух классов путем случайной безвозвратной выборки в соотношении 0,75/0,15/0,10 сформированы обучающий, валидационный и тестовый наборы соответственно.

Обучение сверточной нейронной сети

Исходная архитектура сверточной сети имеет простую структуру, включающую четыре сверточных слоя и бинарный выход с сигмоидальной функцией активации (рис. 1).

Модель №1

Тип слоя	Размерность	Кол-во параметров
conv2d_13 (Conv2D)	(None, 148, 148, 32)	896
max_pooling2d_13 (MaxPooling	(None, 74, 74, 32)	0
conv2d_14 (Conv2D)	(None, 72, 72, 64)	18496
max_pooling2d_14 (MaxPooling	(None, 36, 36, 64)	0
conv2d_15 (Conv2D)	(None, 34, 34, 128)	73856
max_pooling2d_15 (MaxPooling	(None, 17, 17, 128)	0
conv2d_16 (Conv2D)	(None, 15, 15, 128)	147584
max_pooling2d_16 (MaxPooling	(None, 7, 7, 128)	0
flatten_5 (Flatten)	(None, 6272)	0
dropout_4 (Dropout)	(None, 6272)	0
dense_9 (Dense)	(None, 512)	3211776
dense_10 (Dense)	(None, 1)	513

Общее количество обучаемых параметров: 3 453 121

Рис. 1. Архитектура исходной сверточной сети

Ввиду малого количества изображений в исходном наборе GAPED для снижения эффектов переобучения сети использовались методы расширения данных (генерация новых изображений на основе уже имеющихся за счет растяжения, поворота, зеркального отображения и др.). Кроме того, в структуру сети был включен слой прореживания (dropout) с коэффициентом «забывания» 0,5.

По итогам обучения представленной сверточной сети в ходе 45 эпох получены результаты, свидетельствующие о переобучении сети, несмотря на принятые меры (рис. 2).

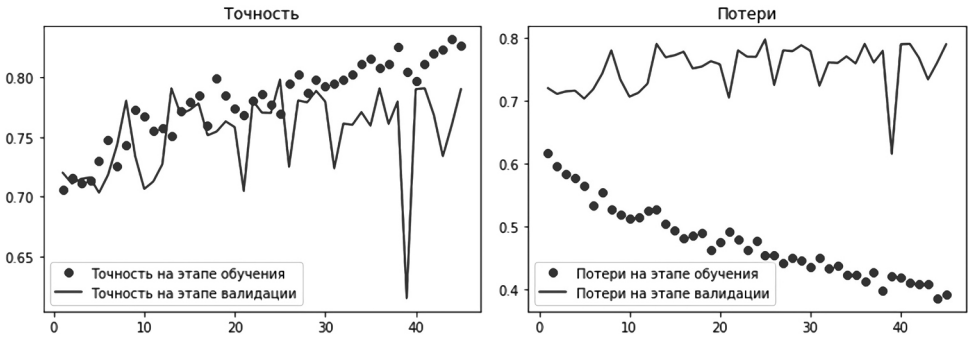


Рис. 2. Результаты обучения сверточной сети (модель № 1) на наборе изображений GAPED

Эффект переобучения особенно заметен на графике потерь: на обучающем наборе потери распознавания постоянно падают, а на валидационном наборе потери не убывают, колеблясь в диапазоне 0,7 … 0,8.

Таким образом, может быть сделан промежуточный вывод, что объем существующих баз данных аннотированных эмоциогенных изображений непригоден для решения задач обеспечения информационно-психологической безопасности пользователей сети Интернет с применением методов и алгоритмов на основе глубокого обучения.

Визуализация знаний, обобщенных сетью

Визуализация промежуточных активаций (карт признаков). Целью данного шага является отображение карт признаков, которые выводятся различными сверточными и объединяющими слоями СНС в ответ на определенные входные данные.

В качестве примера рассмотрим изображение из исходного набора данных GAPED (рис. 3, *слева*).

Отметим, что в ходе предыдущих исследований с применением программно-аппаратного комплекса Gaze Point H3 для каждого конкретного изображения из БД GAPED авторами была получена визуализация точек фиксации взгляда испытуемых (рис. 3, *справа*).

Таким образом, наибольший интерес представляет степень соответствия признаков (объектов, областей), выделяемых на изображении сверточными нейросетями, обученными на разных открытых наборах данных-изображений, и областей на изображениях, на которых чаще всего возникает фиксация взгляда испытуемого.

Информационная безопасность

Рис. 3. Тестовое изображение для визуализации карт признаков обученной сверточной нейросети

Рассматриваемая СНС имеет четыре сверточных слоя. В 1-м сверточном слое (см. рис. 1) формируется 32 карты признаков, во 2-м – 64, в 3-м – 128 и в 4-м – 128 (рис. 4).

Первый слой детектирует цветовые сочетания и простые контуры – линии, углы, точки. По мере подъема по слоям кодируются все более абстрактные понятия («членистоногие» и др.), при этом пустые фильтры говорят о том, что шаблон, соответствующий данному фильтру нейросети, в анализируемом изображении не найден. Разреженность активаций, таким образом, растет по мере увеличения глубины слоя.

Из представленной визуализации сверточных слоев видно, что СНС, обученная на наборе изображений GAPED, детектирует как признаки, относящиеся к человеческой фигуре, так и к объекту класса «паук».

Однако, как было отмечено выше, объема специализированной обучающей выборки в БД GAPED (520 → несколько сотен тысяч программно сгенерированных примеров путем модификаций исходного изображения) оказалось недостаточно для построения эффективного нейросетевого распознавателя.

Визуализация тепловых карт активации классов на предобученной сети VGG-16. Для анализа конфигурации так называемых тепловых карт, «подсвечивающих» области пикселей, которые имеют наибольшие весовые коэффициенты для заданного класса, была применена более масштабная сверточная сеть – VGG-16 [6], которая была предварительно обучена на наборе данных изображений ImageNet [3] (более 14 млн изображений, разбитых на 21 841 класс) и включающая более 138 млн обучаемых параметров.

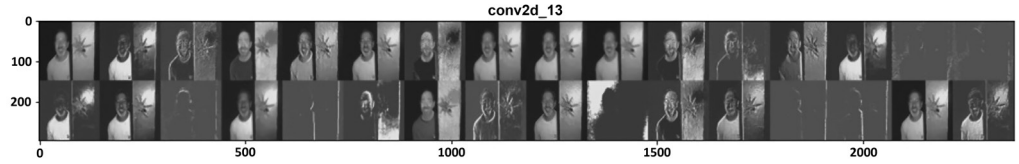
В качестве примера было вновь использовано изображение, представленное на рисунке 3.

Полученный результат распознавания оказался неудовлетворительным.

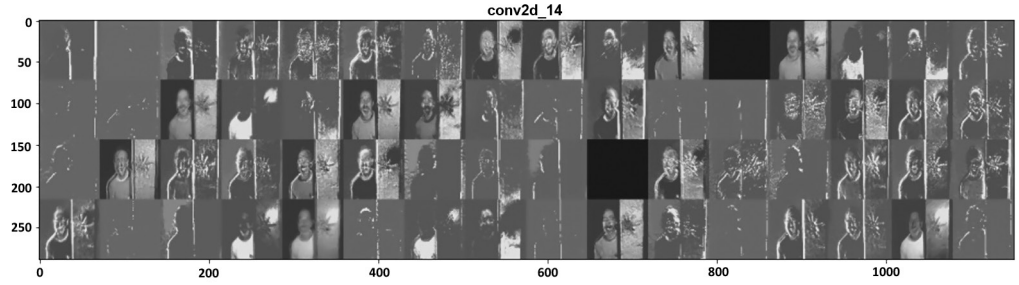
Вывод СНС получился следующий: первая тройка классов, имеющих наибольшую вероятность отнесения к классам изображений: «обелиск» (0,099), «пьедестал» (0,058), «свеча» (0,039). Соответствующая тепловая карта представлена на рисунке 5.

Очевидно, в силу особенностей обучающей выборки СНС реагирует на самый очевидный, а не на самый значимый в контексте сложной сцены набор признаков (столб).

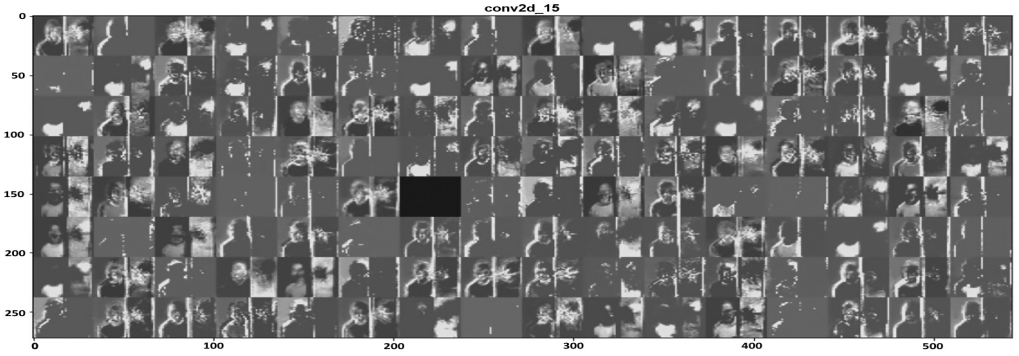
В обучающем наборе ImageNet доля изображений с архитектурными элементами значительно превышает долю специфических стимулов, интересующих авторов в рамках проводимого исследования.



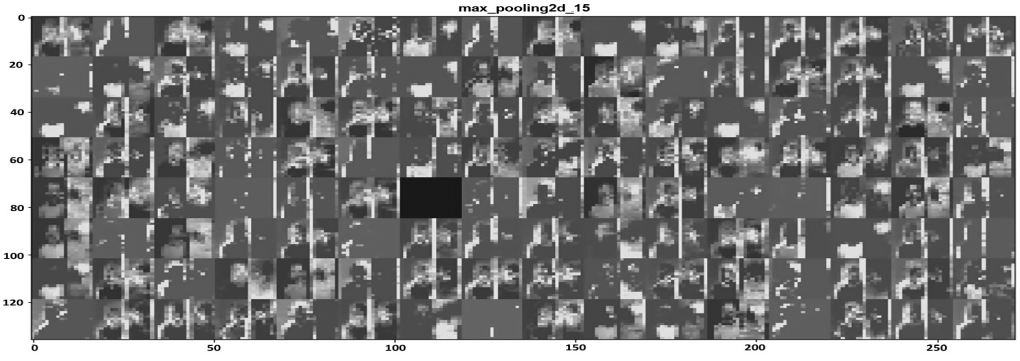
1-й сверточный слой



2-й сверточный слой



3-й сверточный слой



4-й сверточный слой

Рис. 4. Тестовое изображение для визуализации карт признаков обученной сверточной нейросети

Информационная безопасность

Рис. 5. Тепловая карта активации сверточной сети VGG-16 для изображения из набора БД GAPED

Вместе с тем человеческий глаз безошибочно выделяет области интереса, и зритель определяет класс картинки как «негативный» с учетом контекста (опасная близость человека и паука и др.).

При этом на изображениях, не содержащих сложного контекста, метод визуализации тепловых карт сверточных сетей показывает более правильные и ожидаемые результаты, в целом совпадающие с результатами исследований испытуемых с применением видеоокулографа (ай-гейз-трекера).

Таким образом, СНС, обученные на больших неспециализированных выборках изображений, являясь, безусловно, полезным инструментом автоматического выделения сложной иерархической структуры признаков объектов различных классов на изображении, оказываются непригодными в их существующем виде для решения задач защиты пользователей от негативного информационно-психологического контента, поскольку они не способны достоверно идентифицировать объекты, относимые к классу деструктивных.

Направления дальнейших исследований

Для повышения качества полученных результатов целесообразно рассмотреть два основных подхода:

1) наращивание объема специализированных выборок эмоциогенных стимулов;

2) использование СНС, предварительно обученных на больших размеченных выборках изображений (ImageNet, MS-COCO, OpenImagesDataset и др.) путем «замораживания» отдельных слоев и последующего дообучения на выборке эмоциогенных изображений.

В силу отсутствия строгой формальной теории моделирования и применения СНС, а также исключительно высокой размерности пространства настраиваемых параметров, оказывающих существенное влияние на результат (конкретный тип архитектуры сети, вид функций активации слоев, алгоритм модификации весов, количество эпох обучения и др.) оба варианта являются сложными нетривиальными задачами.

Таким образом, оптимизация достигнутых результатов может быть осуществлена только итеративным путем на основе экспериментальных исследований с применением аппаратного вычислителя на основе графического процессора, поддерживающего технологию CUDA.

Заключение

Таким образом, размещенные в открытом доступе выборки аннотированных изображений в результате применения технологий глубокого обучения не обеспечивают правильного распознавания визуального контента, способного нанести ущерб психике пользователя, в силу недостаточного объема либо существенно отличной специфики представленных в них изображений.

Вместе с тем такие выборки являются ценным источником пополнения признакового пространства для рассматриваемой предметной области и представляют интерес для дальнейшего исследования.

Литература

1. *Гнидко К.О., Ломако А.Г.* Экспериментальное исследование бессознательных реакций потребителей мультимедийного контента на эмоционально значимые визуальные стимулы // Материалы 29-й Науч.-практ. конф. 29–30 сентября 2020 г. СПб.: Изд-во Политехнического университета, 2020. С. 21–23.
2. *Dan-Glauser E.S., Scherer K.R.* The Geneva Affective Picture Database (GAPED): a New 730-picture Database Focusing on Valence and Normative Significance // Behavior Research Methods. Springer. 2011. Vol. 43, no. 2. Pp. 468–477.
3. *Deng J., Dong W., Socher R., Li L.-J. et al.* Imagenet: A Large-Scale Hierarchical Image Database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. Pp. 248–255.
4. *Kindel W.F., Christensen E.D., Zylberberg J.* Using Deep Learning to Reveal the Neural Code for Images in Primary Visual Cortex / ArXiv.org – Cornell University [Digital Resource]. – URL: https://arxiv.org/abs/1706.06208 (Date of the Application: 10.11.2020).
5. *Pilkevich S.V., Gnidko K.O.* Formation of the System of Signs of Potentially Harmful Multimedia Objects // Intelligent Distributed Computing XIII / ed. Kotenko I.V. et al. Cham: Springer International Publishing, 2020. Pp. 266–271.
6. *Qassim H., Verma A., Feinzimer D.* Compressed Residual-VGG-16 CNN Model for Big Data Places Image Recognition // 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2018. Pp. 169–175.

Literatura

1. *Gnidko K.O., Lomako A.G.* Eksperimental'noe issledovanie bessoznatel'nykh reaktsij potrebitelej mul'timedijnogo kontenta na emotsional'no znachimye vizual'nye stimuly // Materialy 29-j Nauch.-prakt. konf. 29–30 sentyabrya 2020 g. SPb.: Izd-vo Politekhnicheskogo universiteta, 2020. S. 21–23.
2. *Dan-Glauser E.S., Scherer K.R.* The Geneva Affective Picture Database (GAPED): a New 730-Picture Database Focusing on Valence and Normative Significance // Behavior Research Methods. Springer. 2011. Vol. 43, no. 2. Pp. 468–477.
3. *Deng J., Dong W., Socher R., Li L.-J. et al.* Imagenet: A Large-Scale Hierarchical Image Database // 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009. Pp. 248–255.
4. *Kindel W.F., Christensen E.D., Zylberberg J.* Using Deep Learning to Reveal the Neural Code for Images in Primary Visual Cortex / ArXiv.org – Cornell University [Digital Resource]. – URL: https://arxiv.org/abs/1706.06208 (Date of the Application: 10.11.2020).
5. *Pilkevich S.V., Gnidko K.O.* Formation of the System of Signs of Potentially Harmful Multimedia Objects // Intelligent Distributed Computing XIII / ed. Kotenko I.V. et al. Cham: Springer International Publishing, 2020. Pp. 266–271.
6. *Qassim H., Verma A., Feinzimer D.* Compressed Residual-VGG-16 CNN Model for Big Data Places Image Recognition // 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). IEEE, 2018. Pp. 169–175.