Оптический журнал

Иконика

УДК 51-76, 004.032.26, 004.932.1, 004.8

Представление категорий посредством прототипов согласованной активности нейронов в свёрточных нейронных сетях

© 2021 г. Е. Ю. Малахова

Институт физиологии им. И.П. Павлова РАН, Санкт-Петербург

E-mail: katerina@infran.ru

Поступила в редакцию 21.10.2021

DOI:10.17586/1023-5086-2021-88-12-36-41

Понимание, каким образом отдельное изображение либо категория объектов кодируются в искусственных нейронных сетях, является важным направлением как в области компьютерного зрения, так и для нейронаук. Широко распространен подход, при котором считается, что один скрытый нейрон может детектировать категорию или её доминирующий признак. В данной работе показано, что кодирование осуществляется при помощи коллективной активности нейронов и в обработке большинства категорий участвует до 93% нейронов слоя. Предложен подход к представлению категории посредством построения прототипа, сформированного как матрица ковариации активации нейронов слоя. Подход позволяет учитывать согласованный ответ популяции искусственных нейронов, а также сложность распределённого паттерна активации на различных этапах обработки, начиная с низкоуровневой статистики изображения и до уровня абстрактных и семантических признаков.

Ключевые слова: *свёрточные нейронные сети, представление информации, интерпретируемое глубокое обучение, модель зрительной системы.*

Коды OCIS: *200.4260, 330.4060.*

1. ВВЕДЕНИЕ

Свёрточные нейронные сети (СНС) учатся классифицировать изображения на основе признаков, присутствующих в сцене. Отдельные функциональные единицы (фильтры или нейроны) СНС часто называют детекторами признаков, исходя из их способности обнаруживать наличие определённого пространственного паттерна на изображении. Внутреннее представление изображения для СНС, таким образом, формируется за счёт активности существующих детекторов признаков.

Различные техники фокусируются на исследовании того, как изображение или понятие категории (класса) представлено в сети [1–5], что необходимо для понимания процесса классификации, осуществляемого сетью, повышения эффективности передачи знаний между моделями, контроля и интерпретируемости знаний, кодируемых моделью.

В качестве объекта исследования в большинстве работ выступает активность одного нейрона. Например, работа [6] посвящена поиску и применению фильтров, кодирующих определённый семантический концепт. Однако без дополнительных ограничений, которые принуждают классическую СНС заключать представление концепции в пределах одного нейрона [7, 8], такое предположение упускает из виду сложность кодирования образов на уровне слоя нейронной сети. В некоторых случаях применяется оценка коллективной активности нейронов, например, для расчёта схожести изображений [9, 10], либо

для переноса стиля с одного изображения на другое [11], однако в задачи данных методик не входит изучение кодирования зрительных категорий.

Цель данной работы заключается в изучении представления категорий на уровне коллективной и согласованной активности фильтров, когда ответ нейронного слоя рассматривается как единое целое. В работе анализируются свёрточные слои, так как интерес представляет их сходство с процессами, протекающими в зрительной системе человека [12, 13]. Формирование функционального профиля более поздних слоёв обусловлено поставленной задачей, в то время как свёрточные слои отражают статистику натуральных сцен [14], взаимосвязь которой с процессом распознавания образов широко исследована в нейрофизиологии зрительной системы [15–19].

2. МЕТОДЫ

2.1. Модель и набор данных

Расчёты проводились на СНС классической архитектуры VGG16 [20], обученной на полном наборе данных ImageNet [21]. Применённый подход может быть использован и для других сетей, в том числе биологических. Архитектура VGG16 включает 13 свёрточных слоёв, 5 слоёв обобщения и 3 полностью связанных слоя. В ходе анализа, модели были представлены 10 случайно выбранных категорий ImageNet тренировочного множества (около 1300 изображений на категорию) и полный набор данных из валидационного множества (1000 категорий, по 50 изображений в каждой). Изображения подавались в сеть и активации всех свёрточных слоёв регистрировались в виде многомерных векторов ($k\times dim\times dim\times n$), где k — количество изображений, dim обозначает длину или ширину карты признаков, а n равно количеству фильтров в слое.

2.2. Значимые направления

Каждый элемент слоя нейронной сети определяет направление в многомерном пространстве. Когда входной сигнал, поступающий в модель, достигает слоя, он проецируется на пространство, образованное набором его детекторов признаков.

Значимым считалось направление, ответ которого на специфическую категорию (обучающие и валидационные данные) существенно отличался от среднего ответа фильтра, т.е. имел непересекающиеся доверительные интервалы уровня 0,95. Средний ответ фильтра рассчитывался на основе предъявления 50000 изображений из валидационного набора данных ImageNet. Для получения среднего ответа фильтра первоначально проводилась операция подвыборки, выделяющая максимальную активацию детектора на всем пространстве изображения, а затем было подсчитано среднее по всем изображениям, принадлежащим к классу.

2.3. Ковариация активности фильтров

В этой работе показано, что ковариационная матрица активаций фильтров является удобным способом исследования скрытых представлений изображений в нейронной сети. Она указывает на наличие и направление линейной связи между двумя переменными. Ковариационная матрица для нейронного слоя является симметричной матрицей, где каждое значение отражает взаимосвязь двух фильтров, а сама матрица представляет все зависимости внутри слоя.

Чтобы оценить ковариацию активности фильтров для одного изображения, ответ свёрточного слоя ($k\times dim\times dim\times n$) был преобразован в матрицу X' размера ($m\times n$), где n равно количеству фильтров слоя, а m — длина уплощенных карт активации одного фильтра в ответ на одно изображение. Каждая строка матрицы была центрирована с использованием среднего ответ фильтра, полученного ранее:

$$X_i = X'_i - \mu(Z_i),$$

где i — индекс строки матрицы (фильтра), а $\mu(Z_i)$ обозначает средний ответ i-го фильтра слоя, полученный в результате предъявления всего валидационного множества. Затем была рассчитана ковариационная матрица C:

$$C = \beta X X^T,$$

где β — это коэффициент $1/(m-1)$, m — длина уплощенной карты признаков. В полученной матрице размером $m\times m$ элемент C_{ij} является ковариацией активности i-го и j-го фильтров, а C_{ii} — дисперсия активности i-го фильтра.

2.4. Представление категорий на основе прототипов

На основе ответов модели на обучающие данные были сформированы прототипы категорий путём усреднения матриц ковариаций всех изображений, принадлежащих к классу. Затем ранее не видимые сетью изображения десяти классов пропускались через модель и им присваивались метки на основе их сходства с эталонными прототипами (рассчитанными на всем объёме обучающих данных), аналогично широко распространенному в статистике методу эталонов или ближайших соседей [22]. Сходство с прототипом оценивалось двумя способами: Евклидовой метрикой и через коэффициент корреляции. Изображению присваивалась метка класса ближайшего прототипа, и результаты присвоения сравнивались с реальными метками.

В качестве оценки эффективности было проведено сравнение двух подходов к созданию прототипов: (1) при помощи среднего вектора активации, как, например, в метрике схожести [9] изображений и (2) в виде ковариационной матрицы активаций фильтров.

2.5. Сложность представления

Ответы, вызванные набором изображений одной категории, образуют облако точек данных, спроецированных на оси пространства слоя сети. Анализ главных компонент этих точек данных помогает понять структуру репрезентаций категорий. Метод главных компонент уменьшает сложность данных, преобразуя их в меньшее количество измерений с максимальной дисперсией. Чем больше измерений требуется для объяснения дисперсии данных, тем сложнее представление.

Для матрицы данных X главные направления являются её собственными векторами, а проекции данных на эти оси, называемые главными компонентами, могут быть выражены собственными значениями матрицы. Было проведено сингулярное разложение прототипов, построенных на основе матрицы ковариаций.

3. РЕЗУЛЬТАТЫ

3.1. Значимые направления

Фильтр, чьи ответы на изображения категории значительно отличались от его средней активности, считался значимым направлением, связанным с категорией. Количество значимых направлений (как и их относительное значение) постепенно увеличивается от начальных слоёв и глубже в сеть. При расчёте на тренировочных данных до 93% фильтров демонстрируют активность, отличную от базовой, в ответ на предъявление изображения. Для валидационных данных это количество, по крайней мере, на 30% меньше (57–70%), что говорит о подстройке сети к обучающим изображениям, в то время как новые данные демонстрируют меньше изученных признаков и, таким образом, менее различимы в сформированном пространстве признаков.

3.2. Прототип категории

Учитывая большое количество фильтров, участвующих в обработке категории, имеет смысл рассмотреть представление категории на уровне целого слоя. Результаты показывают (рис. 1), что классификация, использующая ковариационные матрицы в качестве прототипа категории, превосходит классификацию, основанную на векторе средних активаций.

3.3. Сложность категории

Результаты метода главных компонент (рис. 2) представлений категорий показывают, что для первого свёрточного слоя первые 10 глав-

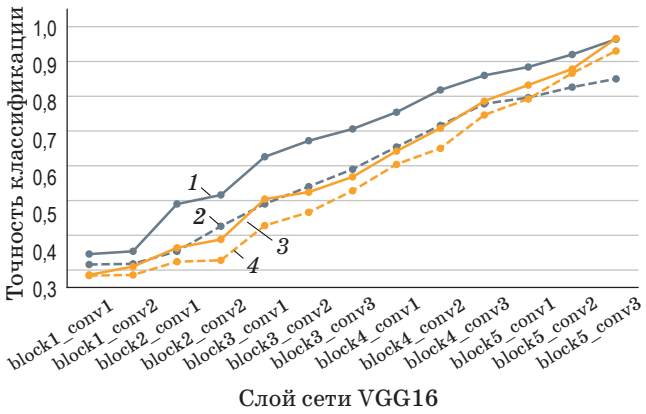

Рис. 1. Точность классификации на основе близости к прототипу для 10 классов, отобранных из валидационного набора данных. Подход с использованием ковариационной матрицы для формирования прототипов отражают кривые *1, 2*, подход на основе средних активаций фильтров – кривые *3, 4*. Сплошной линией обозначено применение Евклидова расстояния для классификации, пунктирной — меры корреляции.

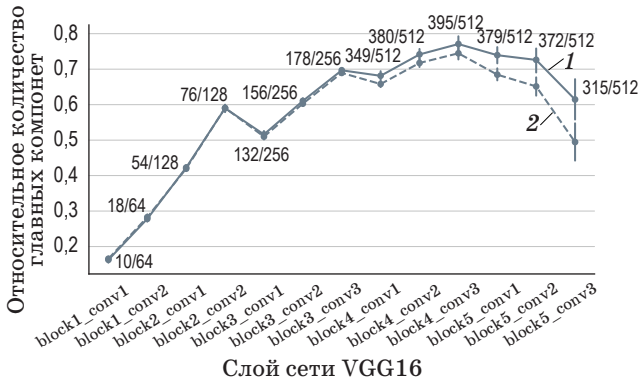

Рис. 2. Сложность представления категории: доля главных компонент, необходимая для покрытия 95% дисперсии ответа слоя. Кривая *1* — для обучающих данных, кривая *2* — для тех же категорий из валидационного набора данных.

ных компонент покрывают около 95% дисперсии. На более поздних этапах обработки количество осей значительно увеличивается, что означает, что репрезентация категорий требует задействования большинства фильтров слоя, формируя сложные распределённые представления. Полученные результаты свидетельствуют об ограниченной применимости методов, когда один фильтр рассматривается в качестве детектора категории [6].

3.4. Оценка прототипа

Доступ к обучающим данным не всегда возможен. Более того, новые поступающие данные могут не всегда наилучшим образом отражать то, что сеть видела и чему она научилась ранее. Поэтому важно понимать, насколько хорошо прототип, построенный на частичных или ранее не виденных данных, схож с эталонным прототипом, рассчитанном на всем объёме обучающих данных.

Рисунок 3а иллюстрирует, как количество изображений, отобранных из обучающих данных, влияет на качество прототипа. Для большинства слоёв достаточно 100 изображений для достижения относительно хорошего приближения. Чем более высокоуровневый слой, тем меньшее количество данных требуется для формирования прототипа.

Зафиксировав количество изображений 50 объектами, было рассмотрено, каким образом влияет использование ранее не виденных сетью данных на формирование прототипа.

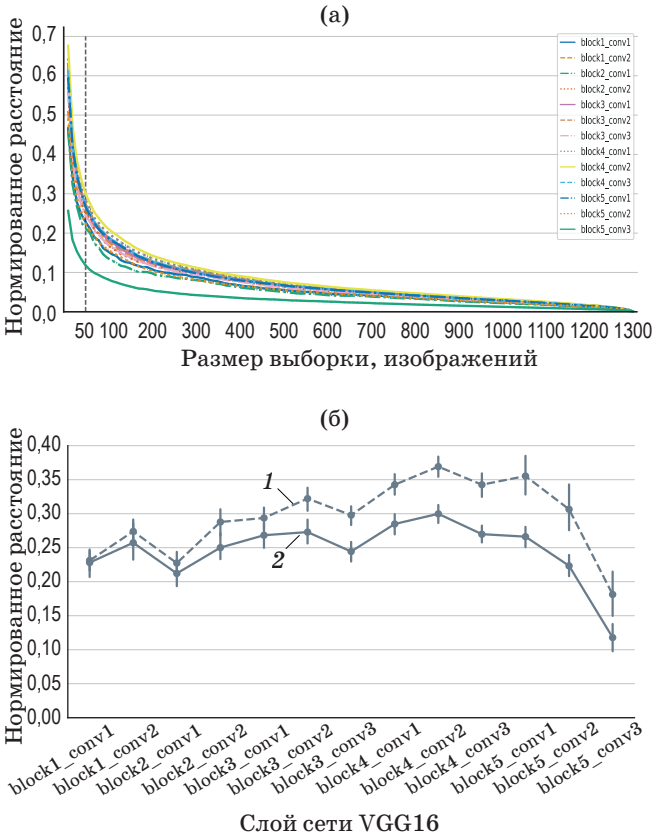

Рис. 3. Аппроксимация прототипа категории на основании частичных или новых данных. Влияние количества изображений, используемых для составления прототипа на его отдалённость (нормированное Евклидово расстояние) от эталонного прототипа (а), сопоставление качества аппроксимации прототипа при использовании обучающих и новых данных (б). Использована выборка размером 50 изображений. Для обучающих данных – кривая *1*, кривая *2* — для тех же категорий из валидационного набора данных.

На рис. 3б показаны расстояния между прототипами, сформированными с использованием 50 изображений, отобранных либо из обучающих, либо из тестовых данных. Можно заметить, что чем выше сложность представления (п. 3.3), тем сложнее его приближение посредством малого количества новых данных (см. средние слои на рис. 3б).

4. ВЫВОДЫ

В работе проведено исследование представления категорий на уровне комплексной активности популяции фильтров в слоях свёрточной сети. Предложен новый подход к рас-

смотрению категории посредством прототипа, описывающего согласованную активность элементов слоя. Впервые проведён многосторонний анализ применения матрицы ковариаций для формирования прототипа и описания категории зрительных образов.

Показано, что количество значимых направлений (а также их относительное значение) увеличивается от начальных слоёв и глубже в сеть. При этом, расчёты, проведённые на обучающих данных, демонстрируют на 30% больше значимых направлений по сравнению с валидационным набором данных. Такие различия могут возникать из-за того, что при настройке представлений сети на учебные изображения веса (и, следовательно, представления категорий) были оптимизированы для обучения, а новые данные не проявляют всех изученных признаков и, таким образом, менее дифференцируемы в заданном пространстве признаков. Значимые направления полезны для понимания того, как отдельные фильтры участвуют в представлении категории, тем не менее они не дают информации о характере совместной активности единиц слоя. Чтобы лучше понять, как происходит кодирование категории внутри слоя, необходимо рассматривать коллективную реакцию всех фильтров слоя.

Показано, что предложенный подход — представление категории посредством прототипа, сформированного ковариационной матрицей активаций всех функциональных единиц слоя, даёт более точные результаты, как на тренировочных, так и на новых данных.

Анализ сложности представлений категорий показывает, что число главных компонент значительно увеличивается на более поздних стадиях обработки, формируя распределённые представления категории, задействующие большинство фильтров слоя.

Проведены расчёты о качестве аппроксимации прототипа при малом количестве данных и при отсутствии доступа к обучающим данным. Для большинства слоёв достаточно 50–100 изображений для достижения относительно хорошего приближения. Чем более высокоуровневый слой, тем меньшее количество данных требуется для формирования прототипа, но тем более сложным является повторение оригинального прототипа, полученного на тренировочных данных при помощи новых изображений.

Таким образом, при изучении кодирования зрительных образов и категорий нейронными сетями, как искусственными, так и биологическими, предлагается применять комплексный анализ репрезентаций, сформированный ответом всех фильтров выбранного слоя.

Результаты работы также имеют практическую значимость в области компьютерного зрения. Подход к формированию прототипов на основе матриц ковариации может быть применён для оценки качества новых данных и детекции аномалий, как метрика переобучения или регуляризации сети, а также для проведения диагностического анализа по соответствию обученной модели текущим задачам.

ЛИТЕРАТУРА

1. *Erhan D., Bengio Y., Courville A., Vincent P.* Visualizing higher-layer features of a deep network // University of Montreal Press. 2009. V. 1341. P. 3.
2. *Simonyan K., Vedaldi A., Zisserman A.* Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034. 2013.
3. *Zeiler M.D., Fergus R.* Visualizing and understanding convolutional networks // Proceedings of the European conference on computer vision. Zurich, Switzerland. September 6–12. 2014. P. 818–833.
4. *Mahendran A., Vedaldi A.* Understanding deep image representations by inverting them // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. P. 5188–5196.
5. *Nguyen A., Yosinski J., Clune J.* Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. arXiv preprint arXiv:1602.03616. 2016.
6. *Bau D., Zhou B., Khosla A., Oliva A., Torralba A.* Network dissection: Quantifying interpretability of deep visual representations // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI. USA. July 21–26. 2017. P. 3319–3327.
7. *Kågebäck M., Mogren O.* Disentanglement by penalizing correlation // Proceedings of the NIPS Workshop on Learning Disentangled Features. Long Beach, CA, USA. Dec 9. 2017. P. E1–8.

8. *Higgins I., Amos D., Pfau D., Racaniere S., Matthey L., Rezende D., Lerchner A.* Towards a definition of disentangled representations. arXiv preprint arXiv:1812.02230. 2018.
9. *Dosovitskiy A., Brox T.* Generating images with perceptual similarity metrics based on deep networks // In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16). Barcelona, Spain. Dec 5–10. 2016. P. 658–666.
10. *Gao F., Wang Y., Li P., Tan M., Yu J., Zhu Y.* Deepsim: Deep similarity for image quality assessment // Neurocomputing. 2017. V. 257. P. 104–114.
11. *Gatys L.A., Ecker A.S., Bethge M.* Image style transfer using convolutional neural networks // Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, NV, USA. June 26 — July 1. 2016. P. 2414–2423.
12. *Kruger N., Janssen P., Kalkan S., Lappe M., Leonardis A., Piater J., Rodriguez-Sanchez A.J., Wiskott L.* Deep hierarchies in the primate visual cortex: What can we learn for computer vision? // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. V. 35. P. 1847–1871.
13. *Cadieu C.F., Hong H., Yamins D.L.K., Pinto N., Ardila D., Solomon E.A., Majaj N.J., DiCarlo J.J.* Deep neural networks rival the representation of primate IT cortex for core visual object recognition // PLoS Computational Biology. 2014. V. 10. P. 12.
14. *Малахова Е.Ю.* Пространство описания зрительной сцены в искусственных и биологических нейронных сетях // Оптический журнал. 2020. № 10. С. 50–58.
15. *Глезер В.Д., Цукерман И.И.* Информация и зрение. АН СССР: М-Л. 1961. С. 183.
16. *Шелепин Ю.Е.* Ориентационная избирательность и пространственно-частотные характеристики рецептивных полей нейронов затылочной коры кошки // Нейрофизиология. 1981. Т. 13 (3). С. 227–232.
17. *Campbell F.W., Robson J.G.* Application of fourier analysis to the visibility of gratings // The Journal of Physiology. 1968. V. 197(3). P. 551–566.
18. *Field D.J.* What the statistics of natural images tell us about visual coding // Human Vision, Visual Processing, and Digital Display. 1989. V. 1077. P. 269–276.
19. *Vinje W.E., Gallant J.L.* Sparse coding and decorrelation in primary visual cortex during natural vision // Science. 2000. V. 287(5456). P. 1273–1276.
20. *Simonyan K., Zisserman A.* Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
21. *Deng J., Dong W., Socher R., Li L.J., Li K., Fei-Fei L.* Imagenet: A large-scale hierarchical image database // Proceedings IEEE conference on computer vision and pattern recognition. Miami Beach, FL, USA. June 20–21. 2009. P. 248–255.
22. *Fix E., Hodges J.L.* Discriminatory analysis. Nonparametric discrimination: Consistency properties // International Statistical Review. 1989. V. 57(3). P. 238–247.