УДК 004.896

АРХИТЕКТУРА СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ДЛЯ КЛАССИФИКАЦИИ ТИПОВ СЦЕН МОБИЛЬНОГО РОБОТА

Каздорф С.Я., Першина Ж.С.

ФГБОУ ВО «Новосибирский государственный технический университет», Новосибирск, e-mail: pershina@tiger.cs.nstu.ru

Актуальной задачей современной наземной мобильной робототехники является повышение степени автономности за счет совершенствования систем управления мобильными роботами, включая навигацию в различных внешних средах (индустриальная среда, городские условия, сеть дорог, пересеченная местность и т.д.) при недетерминированных условиях. Перспективный способ такого управления основан на использовании визуальной информации о внешней среде. Данный подход получил название визуальной навигации. В рамках данного исследования рассматривается задача классификации типа наблюдаемой сцены мобильным роботом с использованием глубоких сверточных нейронных сетей. В работе представлен сравнительный анализ особенностей современных архитектур: VGG-16, Inception v3, ResNet-50, Inception ResNet v2, MobileNet, – и проведены экспериментальные исследования с целью оценки точности и быстродействия, на основе результатов которых выработаны базовые принципы разработки эффективных архитектур, которые позволяют получить наименьшую вычислительную сложность алгоритма при достаточной точности классификации с учетом ограничений по вычислительным ресурсам. Основываясь на этих принципах, разработана собственная архитектура глубокой сверточной нейронной сети, которая обеспечивает точность классификации, сопоставимую с точностью существующих архитектур при большем быстродействии.

Ключевые слова: компьютерное зрение, глубокое обучение, сверточные нейронные сети, классификация, мобильная робототехника

ARCHITECTURE OF CONVOLUTIONAL NEURAL NETWORK FOR SCENES CLASSIFICATION BY MOBILE ROBOT

Kazdorf S.Ya., Pershina Zh.S.

Novosibirsk State Technical University, Novosibirsk, e-mail: pershina@tiger.cs.nstu.ru

The relevant objective of the modern ground mobile robotics is to increase the degree of autonomy by improving the control systems of mobile robots, including navigation in various external environments (industrial environment, urban conditions, roads, rugged terrain, etc.) under nondeterministic conditions. A promising way of such control is based on the usage of visual information about the external environment. This approach is called visual navigation. In the frame of this research, the problem of scene types classification by mobile robot using convolutional neural networks is considered. The paper presents an analysis of state-of-the-art architectures (VGG-16, Inception v3, ResNet-50, Inception ResNet v2, MobileNet) features and experiment results. The basic principles for the development of effective architectures are found based on these results. These principles allow to obtain the least computational complexity of the algorithm with sufficient accuracy of classification, taking into account limitations of computational resources. Using these principles, the own architecture of a deep convolutional neural network is developed, which ensures the accuracy of classification, which is comparable to the accuracy of existing architectures with better performance.

Keywords: computer vision, deep learning, convolutional neural network, classification, mobile robotics

Задача современной наземной мобильной робототехники состоит в повышении степени автономности за счет совершенствования систем управления мобильными роботами, включая навигацию в различных внешних средах (индустриальная среда, городские условия, сеть дорог, пересеченная местность и т.д.) при недетерминированных условиях. Высокая степень автономности может быть достигнута лишь за счет более полного использования всех каналов информации об окружающем мире, включая семантический уровень представления данных. Перспективный способ такого управления основан на использовании визуальной информации о внешней среде. Данный подход получил название визуальной навигации. Такой подход позволит приблизить систему навигации робота к той, которая реализуется интеллектом человека, что, в свою очередь, позволит повысить безопасность и эффективность выполнения сложных операций.

Наибольшего успеха в области распознавания образов на изображении достигли решения, основанные на глубоких сверточных нейронных сетях. Однако стоит отметить, что визуальная навигация в различных типах внешних сред требует распознавания большого количества классов объектов. При этом разработка и реализация алгоритмов распознавания объектов с использованием сверточных нейронных сетей при решении задачи визуальной навигации мобильного робота должна выполняться с учетом ограничений по вычислительным ресурсам. В связи с этим предлагается рассмотреть двухуровневую схему обработки

визуальной информации, при которой на первом уровне выполняется классификация типа наблюдаемой сцены, на втором уровне выполняется распознавание объектов, принадлежащих определенному типу среды, а точнее, их детектирование и локализация. Иными словами, безусловно, можно использовать один уровень, предназначенный для распознавания всех возможных объектов для базовых типов внешних сред, однако это является избыточным, так как нет необходимости искать объекты в тех типах сред, в которых их заведомо быть не должно, из этого следует, что выбор выходных классов объектов зависит от результата классификации типа внешней среды.

Таким образом, основываясь на двухуровневой схеме обработки визуальной информации с целью распознавания объектов, участвующих в навигации мобильного робота, первоочередной задачей является задача классификации типов сцен.

Цель исследования: решение задачи классификации с использованием сверточных нейронных сетей предполагает выполнение следующих этапов: выбор или разработка архитектуры сверточной нейронной сети; формирование обучающего, валидационного и тестового наборов данных для обучения; выбор параметров обучения; обучение; оценка точности.

В рамках данного исследования предполагается выполнить анализ особенностей современных архитектур, предназначенных для классификации изображений, осуществить их обучение на наборе данных Places365 [1], который состоит из изображений классов различных типов сцен, с целью оценки точности классификации и быстродействия, после чего выполнить разработку собственной архитектуры сверточной нейронной сети для решения задачи классификации типов внешних сред мобильного робота с целью достижения лучших показателей точности и быстродействия по сравнению с существующими архитектурами.

Анализ современных архитектур сверточных нейронных сетей (классификация)

Наибольшего успеха в области классификации изображений, как уже отмечалось выше, достигли решения, основанные на глубоких сверточных нейронных сетях, в частности на таких архитектурах, как VGG-16 [2], Inception v3 [3], ResNet-50 [4], Inception ResNet v2 [5], MobileNet [6].

Особенностью архитектуры VGG-16 является использование фильтров одинаковой размерности 3x3. Предшествующие архитектуры, в частности архитектура AlexNet [7], состоят из фильтров различной размерности, что позволяет извлекать признаки различного масштаба, но при этом большая размерность фильтров увеличивает вычислительную сложность свертки:

$$O(N^2) = N_k^2 \cdot C_{in}, \quad (1)$$

где N_k – размер фильтра, а C_{in} – количество каналов на входе. Так, например, при использовании фильтра размерностью 5x5 вычислительная сложность свертки соответствует значению $25 \cdot C_{in}$. При этом сохранения масштаба признака, а именно, размерности рецептивного поля, можно добиться за счет использования фильтров размерностью 3x3 на трех последовательных слоях, такой подход позволяет сократить вычислительную сложность до $18 \cdot C_{in}$.

Архитектура Inception v3 является одной из версии семейства архитектур, основанных на совокупности базовых комбинаций фильтров (рис. 1, а). В данной версии снижение вычислительной сложности операции свертки достигается за счет замены фильтра размерностью $n \times n$, последовательностью одномерных сверток размером $n \times 1$ и $1 \times n$. Также стоит отметить, что одним из принципов построения архитектур сверточных нейронных сетей может являться исключение резкого снижения размерности представления данных, и зачастую для этого используются слои подвыборки (англ. pooling layers). Однако при использовании слоев подвыборки необходимо увеличивать глубину выходных данных в два раза за счет использования дополнительной свертки в глубину размерностью 1×1.

Появление семейства архитектур ResNet явилось результатом исследований возможности создания сверточных нейронных сетей большей глубины. Одной из фундаментальных проблем глубокого обучения является затухающий градиент. Для решения этой проблемы разработчики архитектуры ResNet предложили применение остаточной функции в виде остаточного блока (англ. residual block), обосновав его эффективность как эмпирически, так и математически (рис. 2).

Дальнейшим развитием глубоких сверточных нейронных сетей является попытка объединения ключевых идей семейств Inception и ResNet, в результате чего появляется новая более глубокая и эффективная архитектура Inception ResNet v2.

До недавнего времени, несмотря на достаточно высокие показатели точности всех вышеописанных архитектур, возможность их эффективного применения во встраиваемых системах не являлась возможной. Решением проблемы стала идея совместного использования операции свертки в глубину слоя (англ. depthwise convolution), с последующим применением фильтра размерностью 1x1 (англ.

pointwise convolution), данный подход получил название разделяемой свертки в глубину слоя (англ. depthwise separable convolution). При таком подходе происходит уменьшение вычислительной сложности слоя. В формуле (2) отражена вычислительная сложность сверточного слоя без разделения, в формуле (3) – вычислительная сложность разделяемой свертки в глубину слоя. Данный подход явился основополагающим при разработке семейств архитектур MobileNet.

$$O(N^2) = N_k^2 C_{in} N_f^2 C_{out}, \qquad (2)$$

$$O(N^2) = (N_k^2 + C_{out}) N_f^2 C_{in}, \qquad (3)$$

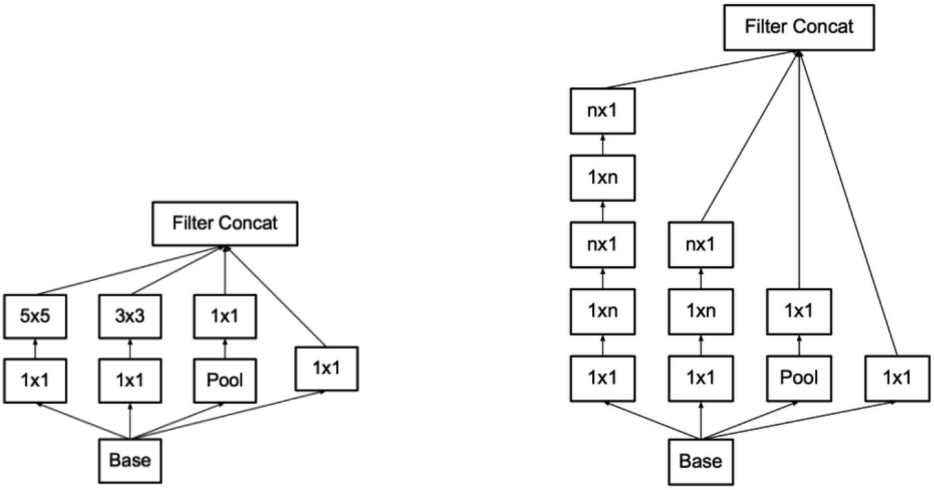
где N_k – размер фильтра, N_f – высота и ширина слоя, C_{in} – количество каналов на входе, C_{out} – количество каналов на выходе.

С целью сравнительного анализа вышеописанных архитектур применительно к решению задачи классификации внешних сред, каждая архитектура обучена на наборе данных Places365 [1]. Данный набор содержит около 1,8 млн изображений, характеризующих 365 типов сцен. В рамках данного исследования выбраны 6 типов сцен, 3 из которых представляют различные урбанистические окружения (завод, моторный отсек, вокзал), а оставшиеся 3 относятся к сценам природы (пляж, поле, лес). При этом тестовая выборка составляет 20% от всего объема изображений, оставшиеся данные разделены в соотношении 30% и 70% на валидационную и обучающую выборку соответственно. Параметры обучения приведены в табл. 1.

Точность классификации (англ. accuracy) рассчитана по формуле (4). Результаты оценки точности приведены в табл. 2, время обработки оценивалось на графическом процессоре Nvidia GTX 1080ti.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (4)$$

где *TP* (англ. True Positives) – верно положительный результат, *TN* (англ. True Negatives) – верно отрицательный результат, *FP* (англ. False Positive) – ложноотрицательный результат (ошибка второго рода), *FN* (англ. False Negative) – ложноположительный результат (ошибка первого рода).



а) *б)*

Рис. 1. Особенности архитектур семейства Inception: а) базовая комбинация фильтров в Inception v1; б) базовая комбинация фильтров в Inception v3 [3]

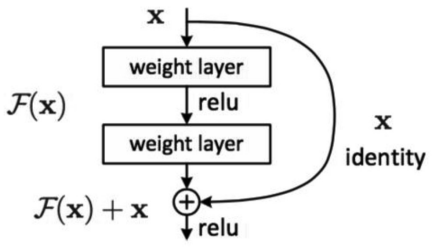


Рис. 2. Структура остаточного блока (residual block) семейства архитектур ResNet [4]

Таблица 1

Параметры обучения

Число шагов	Размер минибатча	Алгоритм оптимизации	Начальный темп обучения	Момент	Коэффициент затухания
8,7 тыс.	32	SGD	0,01	0,9	10^{-3}

Таблица 2

Показатели точности и времени обработки алгоритмов классификации внешних сред

Название модели	Количество параметров сети, млн	Top1 accuracy,%	Время обработки, мс
VGG-16	138	69,0	8,2
ResNet-50	25	71,1	11,1
Inception v3	23	77,2	16,9
Inception ResNet v2	55	**79,9**	34,6
MobileNet v2	4	75,9	**6,4**

Таким образом, видно, что лучшим показателем точности обладает алгоритм классификации на основе наиболее глубокой архитектуры Inception ResNet v2, однако ее размер негативно отражается на времени исполнения данного алгоритма. При этом наименьшее время исполнения показал алгоритм на основе архитектуры MobileNet v2 с точностью на 4% меньшей Inception ResNet v2. В связи с этим возникает необходимость проведения исследований в этом направлении с целью достижения лучших показателей производительности и точности классификации по сравнению с рассматриваемыми архитектурами.

Разработка архитектуры сверточной нейронной сети

Одной из основных характеристик системы визуальной навигации мобильного робота является быстродействие, соответственно, при разработке архитектуры сверточной нейронной сети, предназначенную для классификации наблюдаемых типов сцен мобильным роботом, необходимо руководствоваться принципами, позволяющими получить наименьшую вычислительную сложность при достаточной точности распознавания.

Анализ принципов построения современных архитектур показал следующее:

1) разделяемые свертки в глубину слоя позволяют уменьшить вычислительную сложность сети, при незначительных потерях в точности;

2) остаточные блоки позволяют избежать проблемы затухающего градиента;

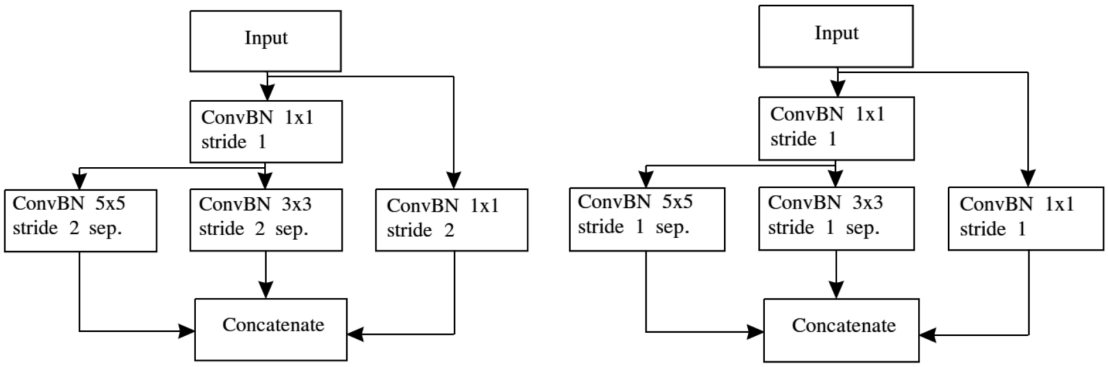
3) объединение результатов применения сверток разного размера на одном уровне параллелизма позволяет извлечь признаки различного масштаба.

Основываясь на этих принципах, сформированы два основных блока сверточной нейронной сети: Pooling-блок и MobileInception-блок.

Pooling-блок (рис. 3, а) осуществляет двукратное понижение пространственных размеров входных данных. Использование свертки 1×1 в остаточной связи позволяет снижать пространственные размеры входных данных. Помимо этого количество фильтров 1×1 в сверточном слое, а также в разделяемых свертках в глубину слоя 3×3 и 5×5 определяют глубину выходных данных. Глубина выходных данных увеличивается в четыре раза на первых двух Pooling-блоках сети и в три раза на третьем и четвертом блоках, что позволяет снизить вычислительную сложность.

В свою очередь MobileInception-блок (рис. 3, б) извлекает признаки входных данных различного масштаба. Как и в предыдущем блоке, размер выходных данных определяется количеством фильтров. Свертка 1×1 на входе блоков применена для снижения размерности и оптимизации последующего вычисления разделяемой свертки в глубину слоя с большим размером фильтра. Общая архитектура сети приведена в табл. 3.

Программная реализация алгоритма классификации на основе разработанной архитектуры, процедуры обучения и логического вывода осуществляется с использованием нейросетевой библиотеки Keras. Для этого требуется описание архитектуры нейронной сети в терминах последовательности связанных слоев и их параметров, а также описание параметров алгоритма обучения. Параметры обучения приведены в табл. 1.



а) *б)*

Рис. 3. Базовая комбинация фильтров: а) структура Pooling-блока, б) структура MobileInception-блока

Таблица 3

Общая архитектура сверточной нейронной сети

Название слоя	Размер входного тензора	Размер фильтра
ConvBn3x3 stride2	224x224x3	3x3x3x64
Pooling блок	112x112x64	1x1x64x128 3x3x64 separable 5x5x64 separable
2xMobileInception блок	56x56x256	
Pooling блок	56x56x256	1x1x256x256 3x3x128 separable 5x5x128 separable
3xMobileInception блок	28x28x512	
Pooling блок	28x28x512	1x1x512x256 3x3x256 separable 5x5x256 separable
2xMobileInception блок	14x14x768	
Pooling блок	14x14x768	1x1x768x512 3x3x512 separable 5x5x512 separable
MobileInception блок	7x7x1536	
AvgPooling7x7	7x7x1536	~
Dropout 0.35	1x1x1536	~
Dense softmax	1x1x1536	1536x6

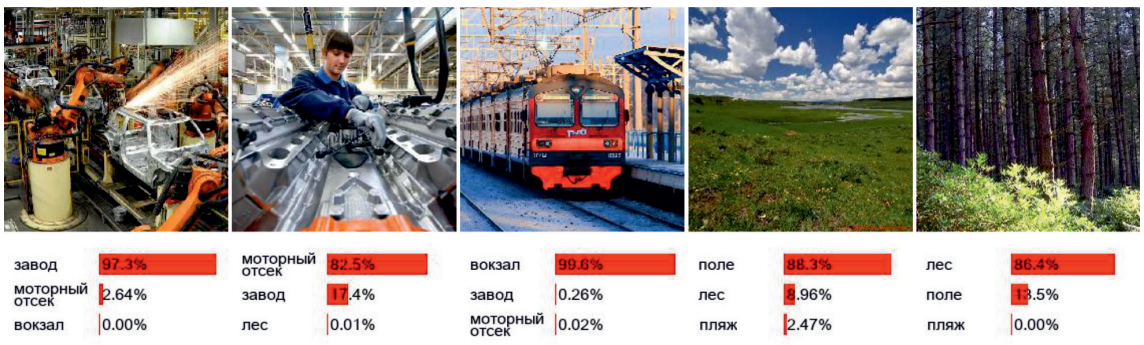


Рис. 4. Примеры классификации типов сцен

Результаты исследования и их обсуждение

В результате обучения разработанной архитектуры на наборе данных Places365 точность классификации на тестовой выборке составила 79,7%, а время обработки 11,6 мс. Примеры классификации типов сцен приведены на рис. 4 с указанием принадлежности изображения внешней среды к трем классам, вероятность которых наиболее высокая. Сравнивая полученный результат с результатами, приведенными в табл. 2, можно сказать, что точность разработанной архитектуры сопоставима с точностью архитектуры Inception ResNet v2 и значительно превышает точность MobileNet v2. По времени исполнения разработанная архитектура уступает MobileNet v2, при этом превосходит результат Inception ResNet v2.

Заключение

В рамках данного исследования выполнен анализ особенностей современных архитектур, предназначенных для классификации изображений. Разработана собственная архитектура сверточной нейронной сети для решения задачи классификации типов внешних сред мобильного робота с учетом ограничений по вычислительным ресурсам, обеспечивающая точность классификации, сопоставимую с точностью существующих архитектур при большем быстродействии.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-58-76003.

Список литературы

1. Zhou B. et al. A. Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence. 2017. URL: https://ieeexplore.ieee.org/document/7968387/ (accesed 13.07.2018).

2. Simonyan K., Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014. URL: https://arxiv.org/abs/1409.1556 (accesed 13 July, 2018).

3. Szegedy C. et al. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 2818–2826.

4. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. P. 770–778.

5. Szegedy C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI. 2017. T. 4. P. 12.

6. Howard A.G. et al. Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861. 2017. URL: https://arxiv.org/abs/1704.04861 (accesed 13.07.2018).

7. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012. P. 1097–1105.