Применение системы DeepLab для решения задачи семантической сегментации дорожных сцен*

М. Жильцов, В. Кустикова
zhiltsov.max35@gmail.com/valentina.kustikova@itmm.unn.ru
Россия, Институт информационных технологий, математики и механики
Нижегородский государственный университет им. Н.И. Лобачевского

Рассматривается задача семантической сегментации изображений, содержащих дорожные сцены. Дается обзор методов решения задачи семантической сегментации. Изучается возможность применения известной системы DeepLab, которая демонстрирует одни из лучших результатов на данных PASCAL VOC 2012, к изображениям дорожных сцен. В процессе исследования DeepLab расширяется инструкциями и скриптами по сборке, установке и запуску, приложениями для визуализации полученных результатов семантической сегментации, также выполняется оптимизация модуля работы с условными случайными полями. Эксперименты проводятся на открытом наборе данных CamVid, содержащем сцены с 32 семантическими классами. Результаты экспериментов показывают, что в ходе многоклассовой сегментации дорожной сцены качество работы системы сопоставимо с современными методами, в среднем отличается на 2–2.5%, при этом на некоторых классах дорожных объектов выигрыш составляет от 6.8 до 21.6%.

Ключевые слова: *семантическая сегментация изображений, глубокое обучение, сверточные нейронные сети, условные случайные поля, система DeepLab*

1. Введение

Задача семантической сегментации изображений является актуальной и практически значимой. Решение данной задачи позволяет получить информацию о контексте сцены. Семантическая сегментация находит широкое применение в системах компьютерного зрения и обработки изображений таких, как системы помощи водителю (Advanced Driver Assistance Systems, ADAS) и системы обработки медицинских изображений.

В настоящее время одними из наиболее перспективных методов при решении задачи являются методы глубокого обучения [27], [32], основанные на построении глубоких нейронных сетей. Цель настоящей работы состоит в том, чтобы исследовать возможности открытой системы DeepLab [9], основанной на применении глубокого обучения, для сегментации сцен дорожного движения.

Работа построена следующим образом: ставится задача семантической сегментации, дается краткий обзор методов ее решения, приводится общая схема решения задачи с использованием системы DeepLab, описываются полученные результаты на данных CamVid [4].

2. Задача семантической сегментации

Задача семантической сегментации предполагает, что имеется некоторое изображение и для каждой единицы квантования (пикселя, области или набора областей) изображения необходимо определить классы изображенных объектов. Как следствие, результатом сегментации является изображение, представляющее собой "карту" индексов классов объектов. Необходимо отметить, что в отличие от задачи сегментации изображения, которая позволяет разбить изображение на области, задача семантической сегментации еще отвечает на вопрос, какому классу каждая из областей принадлежит.

3. Методы решения задачи

К числу классических методов сегментации изображений относятся *метод сегментации по водоразделам* (watershed) [14], *метод сдвига среднего* (mean shift) [7], *метод разреза графа* (graph cut) [26] и другие. Результаты сегментации могут использоваться для семантической сегментации. Так, в [19] предлагается метод, основанный на применении *условных случайных полей* (Conditional Random Fields, CRFs) к результатам сегментации исходного изображения с использованием метода разреза графа. В [32]] развиваются идеи применения CRFs. Для полного решения задачи семантической сегментации требуется только информация о присутствующих классах объектов. Также разрабатываются методы группы "мешка слов" (bag-of-words) [11].

В настоящее время распространение получили методы, основанные на построении *сверточных нейронных сетей* (Convolutional Neural Network, CNN). Авторы [2] и [8] предлагают новые архитектуры сетей для решения задачи семантической сегментации дороги и дорожных сцен. В [6] комбинируются идеи построения сверточных нейронных сетей и графических моделей. При этом графические модели позволяют генерировать небольшое количество гипотез касательно сегментации, а сверточная сеть — извлечь признаки из гипотез и выпол-

Работа опубликована при финансовой поддержке РФФИ, грант №16-07-20482

нить грубую сегментацию. Наряду с этим, существуют работы [20], [22], которые используют "*перенос обучения*" (transfer learning) для глубоких нейросетевых моделей. Цель состоит в том, чтобы обучить глубокое представление (веса нейронной сети) для решения одной задачи и использовать его при решении другой. Примерами таких моделей могут служить сети AlexNet [17], GoogLeNet [30], VGG–16 [28], ResNet [12], которые сначала обучаются на данных ImageNet [15], потом у них модифицируются последние слои под конкретную задачу, и полученные сети дообучаются на данных интересующей задачи. Также выделяются методы, комбинирующие построение глубоких нейронных сетей с применением моделей CRFs [21]. В системе DeepLab реализован такой комбинированный подход [5], [23].

4. Система DeepLab

Система DeepLab построена на базе широко известной библиотеки глубокого обучения Caffe [3] и библиотеки с реализацией CRFs [16]. По существу функционал Caffe дополнен новыми слоями, скриптами и приложениями, необходимыми для решения рассматриваемой задачи. В системе добавлено порядка 10 типов слоев, среди которых слои для чтения различных форматов данных, слой DenseCRF для работы с условными случайными полями без подключения дополнительных библиотек, интерполяционный слой.

В ходе настоящего исследования система DeepLab дополнена инструкциями и скриптами по сборке, установке и запуску, приложениями для визуализации полученных результатов семантической сегментации, также выполнена оптимизация работы с файловой системой в реализации CRFs [10].

5. Общая схема решения задачи

Общая схема решения задачи семантической сегментации изображений с использованием системы DeepLab состоит из двух основных этапов.

1. *Получение грубой "карты" сегментов.* На данном этапе исходное изображение "пропускается" через обученную глубокую нейронную сеть. В результате на выходе формируется набор "карт" достоверностей того, что каждый пиксель принадлежит определенному классу объектов (рис. 1, столбец 2). При этом количество "карт" совпадает с числом классов. Далее, если выбрать для каждого пикселя класс, например, с максимальным значением достоверности, то можно уже после этого этапа сформировать "карту" индексов классов и получить грубую карту сегментов на изображении.
2. *Уточнение границ сегментов с помощью CRFs.* Применение CRFs позволяет улучшить результаты сегментации, полученные на предыдущем этапе, поэтому не является обязательным. Алгоритм CRFs работает на множестве "карт" достоверностей, полученных на первом этапе, и формирует результирующую "карту" индексов классов (рис. 1, столбец 3).

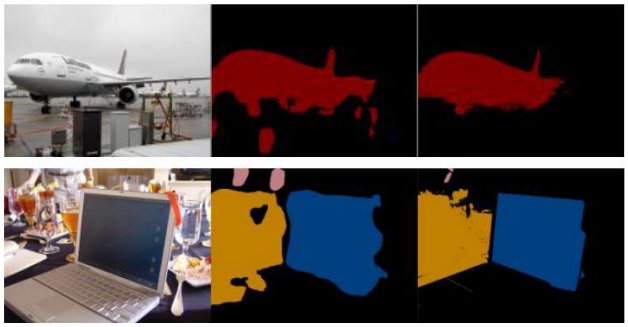

Рис. 1: Примеры изображений из набора PASCAL VOC 2012 [24].

Решение задачи с использованием описанного подхода требует наличия обученной модели глубокой нейронной сети. Обучение модели осуществляется стандартными средствами библиотеки Caffe. Для этого требуется описание архитектуры нейронной сети в терминах последовательности связанных слоев и их параметров, а также описание параметров алгоритма обучения. Предоставляемые авторами DeepLab модели ориентированы на семантическую сегментацию изображений, содержащих 21 класс объектов, которые описаны в наборе данных PASCAL VOC 2012 [24].

6. Вычислительные эксперименты

6.1 Тренировочные и тестовые данные

Для оценки качества работы системы DeepLab на задаче сегментации дорожных сцен используется набор данных CamVid [4]. В наборе CamVid имеется 32 семантических класса из категорий "движущиеся объекты", "дорога", "верх изображения" (небо, тоннель и т.п.), "зафиксированные в пространстве объекты" (здание, дерево, светофор и т.п.). Основное назначение этого набора – полная сегментация дорожной сцены. Набор содержит 701 изображение размером до 960×720 с попиксельной разметкой, из них для экспериментов было отобрано 233 в качестве тестовой выборки.

6.2 Показатель качества

Для оценки качества сегментации используется среднее значение метрики *IoU* (Intersection over Union) [28] среди 11 основных классов – "авто", "небо", "дорога", "забор", "здание", "дерево", "тротуар", "велосипедист", "пешеход", "дорожный знак", "столб".

6.3 Модели глубоких нейронных сетей

При проведении экспериментов используется модифицированная модель MSc-LargeFOV нейросети Deeplab-MSc-LargeFOV, основанная на VGG-16. Основные отличия от VGG-16 состоят в следующем:

1. Увеличение максимального размера входного изображения до необходимого в зависимости от набора данных.
2. Замена полностью связных слоев на сверточные.

Следует также отметить, что Deeplab-MSc-LargeFOV рассматривает входное изображение на разных уровнях детализации. Выход сети получается объединением результатов с разных уровней.

Начальные веса сети MSc-LargeFOV инициализируются значениями, полученными в ходе обучения соответствующей сети Deeplab-MSc-LargeFOV на данных PASCAL VOC 2012. Для использования сети на наборе данных CamVid заменяется последний слой сети в соответствии с количеством семантических классов в дорожных сценах (32 класса) и производится дообучение на тренировочном множестве изображений с дорожными сценами.

6.4 Результаты

Ниже представлены полученные результаты экспериментов (таблица 1). Несложно видеть, что увеличение количества классов в процессе сегментации ведет к усложнению задачи, причем для разных классов это отражается по-разному. Так, для классов, представленных в таблице, можно наблюдать сильное расхождение в значениях точности. На некоторых классах качество работы системы крайне низкое, к таким относятся классы «столб» — 5.4% и «дорожный знак» — 17.1%. На других, напротив, качество сегментации достаточно высокое: класс «небо» — 90.7%, «здание» — 83.5%, «дерево» — 75.6%. В некоторых случаях (класс «забор») наблюдается улучшение на 6.8—21.6% по сравнению с существующими подходами. При этом в среднем отличие по 11 классам не превышает 2–2.5%.

Таблица 1: Качество сегментации на избранных семантических классах данных CamVid без применения CRFs.

Метод	IoU (class accuracy), %			
	Класс “авто”	Класс “дорога”	Класс “забор”	Среднее значение
MSc-LargeFOV	70.7	88.8	54.4	59.0
Ladicky et al. [18]	78.7	93.9	47.6	62.5
Sturgess et al. [29]	72.7	95.3	45.7	59.2
Tighe and Lazebnik [31]	78.1	96.0	32.8	62.5

Время работы сети составляет ∼5 минут, то есть на обработку одного изображения требуется ∼1.2 секунды. Время обучения составляет ∼8 часов. В ходе экспериментов использовалась система с процессором Intel(R) Core i5-2430M @ 2.4GHz, 6GB RAM и графическим процессором NVIDIA Tesla X2070.
Одним из основных компонентов системы DeepLab является модуль DenseCRF. Применение CRFs к результатам сегментации, полученным с использованием сети, не приводит к значительным изменениям показателей качества. В среднем они находятся в пределах 1%. При этом для каждого отдельного класса результаты могут как улучшиться, так и ухудшиться. Время работы этого компонента системы для описанного набора данных составляет ∼40 минут, или 10 секунд на одно изображение.
Таким образом, данную сеть можно использовать для сегментации сложной дорожной сцены с высокой точностью сегментации для ряда классов дорожных объектов.

7. Заключение

В данной работе рассмотрена задача семантической сегментации дорожных сцен. Для ее решения применена широко известная система DeepLab. Система основана на обучении глубоких сверточных нейронных сетей с целью получения грубой карты сегментов и построении условных случайных полей для уточнения границ этих сегментов. Проведены эксперименты на открытом наборе данных CamVid, содержащем сцены с 32 семантическими классами. В ходе экспериментов рассматривались 11 наиболее значимых классов. Результаты экспериментов показали, что в случае многоклассовой сегментации дорожной сцены качество работы системы сопоставимо с современными методами, в среднем отличается на 2–2.5%, при этом на некоторых классах дорожных объектов выигрыш составляет от 6.8 до 21.6%.
В дальнейшем планируется сравнить качество работы DeepLab с системой SegNet [1], которая специализирована для семантической сегментации дорожных сцен. Также рассматривается перспектива применения/модификации указанных систем для решения более узкой задачи — детектирования дорожных препятствий.

8. Благодарности

Работа выполнена в лаборатории «Информационные технологии» Института информационных технологий, математики и механики ННГУ им. Н.И. Лобачевского при поддержке компании Itseez.

Литература

[1] Badrinarayanan V., Handa A., Cipolla R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling. – URL: https://arxiv.org/pdf/1511.00561.pdf. – 2015.

[2] Brust C.-A., Sickert S.,Simon M., Rodner E., Denzler J. Convolutional Patch Networks with Spatial Prior for Road Detection and Urban Scene Understanding. – 2015. – URL: http://arxiv.org/abs/1502.06344.

[3] Caffe Framework URL: http://caffe.berkeleyvision.org.

[4] CamVid Dataset URL: http://mi.eng.cam.ac.uk/ research/projects/VideoRec/CamVi_d

[5] Chen L.-Ch., Papandreou G., Kokkinos I., Murphy K., Yuille A.L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs // ICLR. – 2015. – URL: http://arxiv.org/abs/1412.7062.

[6] Cogswell M., Lin X., Purushwalkam S., Batra D. Combining the Best of Graphical Models and ConvNets for Semantic Segmentation. – 2012. – URL: http://arxiv.org/abs/1412.4313.

[7] Comaniciu D., Meer P. Mean Shift: A Robust Approach Towards Feature Space Analysis // IEEE Transactions on Pattern Analysis and Machine Intelligence. – Vol. 24, Iss. 5.– P. 603-619. URL: https://courses.csail.mit.edu/6.869/handouts/ PAMIMeanshif t.pd.

[8] Dan L., Noa G., Ethan F. StixelNet: A Deep Convolutional Network for Obstacle Detection and Road Segmentation // In the Proceedings of the 26th British Machine Vision Conference (BMVC). – 2015.

[9] DeepLab System Public Repository URL: https://bitbucket.org/deeplab/deeplab-public.

[10] DeepLab System Working Copy URL: https://github.com/ITLab-Vision/ITLab-Vision-deeplab.

[11] Freytag A., Fröhlich B., Rodner E. Efficient Semantic Segmentation with Gaussian Processes and Histogram Intersection Kernels // In the Proceedings of the 21st International Conference on Pattern Recognition. – 2012. – P. 3313-3316.

[12] He K., Zhang X., Ren Sh., Sun J. Deep Residual Learning for Image Recognition. – 2015. – URL: http://arxiv.org/abs/1512.03385.

[13] Hinton G.E. Learning Multiple Layers of Representation // Trends in Cognitive Sciences. – 2007. – Vol. 11. – P. 428-434.

[14] Image segmentation and mathematical morphology URL: http://cmm.ensmp.fr/ beucher/wtshed.html.

[15] ImageNet Database URL: http://www.image-net.org.

[16] Krähenbühl P., Koltun V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials // NIPS. – 2011. – URL: http://www.philkr.net/home/densecrf.

[17] Krizhevsky A., Sutskever I., Hinton G.E. ImageNet Classification with Deep Convolutional Neural Networks // Advances in Neural Information Processing Systems 25 (NIPS 2012). – 2012. – P. 1097–1105. – URL: http://www.cs.toronto.edu/ fritz/absps/ imagenet.pdf.

[18] Ladický Ľ. et al. What, where and how many? Combining object detectors and crfs // Computer Vision–ECCV 2010. – Springer Berlin Heidelberg, 2010. – C. 424-437.

[19] Ladicky L., Russell C., Kohli P. Associative Hierarchical CRFs for Object Class Image Segmentation // In Proceedings of the IEEE International Conference on Computer Vision. – 2009. – P. 739-746.

[20] Long J., Shelhamer E., Darrell T. Fully Convolutional Networks for Semantic Segmentation. – 2014. – URL: http://arxiv.org/abs/1411.4038.

[21] Mohan R. Deep Deconvolutional Networks for Scene Parsing. – 2014. – URL:http://arxiv.org/abs/1411.4101.

[22] Noh H., Hong S., Han B. Learning Deconvolution Network for Semantic Segmentation. – 2015. – URL: http://arxiv.org/abs/1505.04366.

[23] Papandreou G., Chen L.-Ch., Murphy K., Yuille A.L. Weakly- and Semi-Supervised Learning of a DCNN for Semantic Image Segmentation. – 2015. – URL: http://arxiv.org/abs/1502.02734.

[24] PASCAL Visual Object Classes URL: http://host.robots.ox.ac.uk/pascal/VOC.

[25] PASCAL VOC Development Kit Documentation URL: http://host.robots.ox.ac.uk:8080/pascal/VOC/ voc2012/devki t_doc.pdf.

[26] Rother C., Kolmogorov V., Blake A. "GrabCut": interactive foreground extraction using iterated graph cuts // In Proceeding SIGGRAPH'04 (ACM SIGGRAPH 2004 Papers). – 2004. – P. 309-314.

[27] Schmidhuber J. Deep Learning in Neural Networks: An Overview. – URL: http://arxiv.org/abs/1404.7828.

[28] Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Visual Recognition. – 2014. – URL: http://www.robots.ox.ac.uk/ vgg/research/ very_deepURL:.

[29] Sturgess P., Alahari K., Ladicky L., Torr P.H.S. Combining appearance and structure from motion features for road scene understanding // BMVC. – 2009.

[30] Szegedy C., et al. Going Deeper with Convolutions. – 2014. – URL: http://arxiv.org/abs/1409.4842.

[31] Tighe J., Lazebnik S. Finding things: Image parsing with regions and per-exemplar detectors //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2013. – C. 3001-3008.

[32] Vezhnevets A., Buhmann J.M., Ferrari V. Active Learning for Semantic Segmentation with Expected Change // In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2012. – P.3162-3169.