АЛГОРИТМ СЕМАНТИЧЕСКОЙ СЕГМЕНТАЦИИ ИЗОБРАЖЕНИЙ В РЕЖИМЕ РЕАЛЬНОГО ВРЕМЕНИ

Д.С. Григорьев, В.Г. Спицын
Томский политехнический университет
trygx@tpu.ru

Введение

Практический интерес к интеллектуальным системам наблюдения, беспилотным средствам передвижения и робототехнике сформировал потребность в быстродействующих алгоритмах визуального анализа сцены, способных работать в узких рамках технических характеристик мобильных вычислительных устройств.

Семантическая сегментация является задачей кластеризации частей изображения в соответствии с принадлежностью к определенному классу объектов. Данные алгоритмы находят применение при решении задач обнаружения и распознавания дорожных знаков, распознавания нейронных структур на снимках с электронного микроскопа, а также управлении автономным транспортным средством [1,2].

Высокую точность в распознавании объектов показывают глубокие сверточные нейронные сети, демонстрируя лучшие результаты на различных задачах распознавания образов, в сравнении с другими алгоритмами, такими как машины опорных векторов, условные случайные поля, случайные леса и пр. Тем не менее многие модели сетей содержат большое количество настраиваемых параметров ($\times 10^6$) что ограничивает их применимость только высокопроизводительном оборудовании. В случае, когда необходима работа алгоритма в режиме реального времени и низкое энергопотребление подобные модели не могут применяться. Поэтому ставиться задача разработки алгоритма, основанного на применении компактной архитектуры сети, позволяющей обеспечить работу в режиме реального времени, с малым количеством вычисляемых параметров.

Наборы данных и метрики

Для проведения численных экспериментов были использованы общедоступные наборы изображений для семантической сегментации. Первый набор – Cityscapes [3]. Набор содержит 5000 аннотированных изображений, из которых доступно 2975 в качестве обучающей выборки, 500 для валидации, 1525 в качестве тестовой выборки, и 19 категорий объектов. Cityscapes содержит множество различных дорожных сценариев, часто показывающие множество пешеходов и велосипедистов (рис. 1).

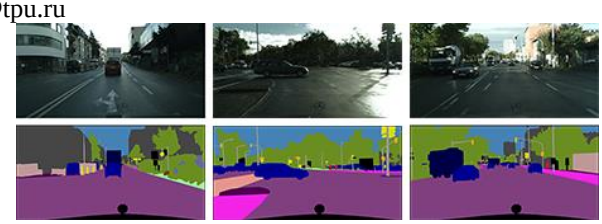

Рис. 1. Примеры изображений из набора Cityscapes вместе с их аннотированными вариантами

Второй набор PASCAL VOC 2012 [4] состоит из 1464 аннотированных изображений для обучения, 1449 для валидации и 1456 для теста, включает в себя 21 класс объектов среди которых – 20 соотносятся с объектами переднего плана, и 1 – для маркировки заднего плана как отдельного класса.

Рис. 2. Примеры изображений из набора PASCAL VOC 2012 вместе с их аннотированными вариантами

Для оценки корректности работы сети на представленных наборах данных использовались следующие метрики:

1. Mean accuracy:

$$Mean\ accuracy = \frac{1}{k}\sum_{i=1}^{n}\frac{n_{ii}}{t_i};$$

2. Mean intersection over union (MIoU):

$$MIoU = \frac{1}{k}\sum_{i=1}^{n}\frac{n_{ii}}{t+\sum_{j=1}^{k}n_{ji}-n_{ii}};$$

3. Frequency weighted intersection over union (FWIoU):

$$FWIoU = \left(\sum_{p=1}^{k}t_p\right)^{-1}\sum_{i=1}^{n}\frac{n_{ii}}{t+\sum_{j=1}^{k}n_{ji}-n_{ii}};$$

Где $k \in \mathbb{N}$ - общее количество классов, $n_{ij} \in \mathbb{N}_0$ $i, j \in 1,...,k$ - соответственно количество пикселей принадлежащих классу i, и соотнесенных к классу j. $t_i = \sum_{j=1}^{k} n_{ij}$ - общее количество пикселей принадлежащих классу i.

Архитектура сети и обучение

Архитектуры решающие задачи семантической сегментации основываются на применении архитектуры, в которой полносвязные слои заменены слоями свертки. Примерами могут служить сети типа SegNet и FCN [1,5]. Несмотря на эффективное распознавание, архитектура этих сетей содержит большое число вычисляемых параметров, что ограничивает их применение в системах реального времени.

Отличительной особенностью этих сетей является использование топологии «энкодер-декодер». Энкодер представляет собой обыкновенную сверточную сеть, обученную для классификации входного образа, тогда как декодер выполняет интерполяцию выхода энкодера.

В качестве активационной функции использовался PReLU (Parametric ReLU). Дополнительные численные эксперименты с еще одной разновидностью функции активации ELU не показали увеличения точности:

$$f(x)=\begin{cases} x & x>0 \\ \alpha\left(\exp(x)-1\right) & x\le 0 \end{cases}, \quad f'(x)=\begin{cases} 1 & x>0 \\ f(x)+\alpha & x\le 0 \end{cases},$$

где гиперпараметр $\alpha>0$.

Для формирования оптимальной архитектуры сети предложено использовать формат, энкодер-декодера составленный из блоков (22 и 10 соответственно). Размерность входного слоя 512×512. Каждый блок состоит из трех слоев свертки: фильтр 1×1, для сокращения размерности, и уменьшения количества вычислений. В этом случае фильтр свертки 1×1 заменяется на увеличенный 2×2 с параметром смещения равным 2. Блок представляет собой сверточный слой с фильтрами различного типа в зависимости от расположения:

1. Обычная свертка;
2. Дилатационная свертка;
3. Ассимметричная свертка (представляет собой последовательность фильтров с размерами $1\times n$ и $n\times1$ соответственно).

Размерность выходного слоя $C\times512\times512$, где *C*-количество классов.

Обучение и тестирование разработанной архитектуры осуществлялось при помощи открытой библиотеки Caffe и 2x NVIDIA 980. В качестве алгоритма оптимизации использовался SGD со следующими параметрами. Параметр скорости обучения вычислялся следующим образом:

$$base\_learinig\_rate\times\left(1-\frac{iter}{\max\_iter}\right)^{power},$$

где *power* = 0.9, *iter*, *max_iter* – текущее и максимальное число итераций соответственно. Параметры *base_learning_rate* $5*10^{-4}$, batchsize 10, регуляризации весов (weight decay) $5*10^{-4}$.

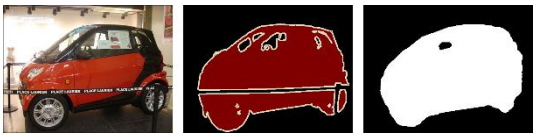

Рис. 3. Примеры результата сегментации изображений предложенной архитектурой

В результате на наборе данных PASCAL VOC 2012 по метрике MeanIoU 61.7, для Cityscapes 58.4 (Class iIoU). Производительность разработанной сети по параметру скорости работы составила в среднем 150 fps при заданном разрешении (512 * 512) изображений. Что более чем соответствует работе в режиме реального времени. Общее количество вычисляемых параметров сети составило 0.35×10^{6}.

Заключение

В результате разработанный алгоритм, основанный на применении сверточной сети, топологически представленной в виде энкодер-декодера позволяет получить приемлемые результаты в задаче семантической сегментации на 2х наборах данных, при этом обладает небольшим количеством вычисляемых параметров. В дальнейшем предполагается проведение численных экспериментов с использованием мобильных и встраиваемых платформ.

Работа выполнена в рамках Программы повышения конкурентоспособности ТПУ и при финансовой поддержке РФФИ в рамках научного проекта № 18-08-00977 А.

Список использованных источников

1. E. Shelhamer, J. Long, T. Darrell. Fully Convolutional Networks for Semantic Segmentation. [Электронный ресурс]. – URL:https://arxiv.org/abs/1605.06211
2. O. Ronneberger, P. Fischer, T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. [Электронный ресурс]. – URL:https://arxiv.org/abs/1505.04597
3. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. [Электронный ресурс]. – URL:http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
4. M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
5. V. Badrinarayanan, A. Kendall, R. Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. [Электронный ресурс]. – URL: https://arxiv.org/abs/1511.0056