МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра математического обеспечения и применения ЭВМ

ОТЧЕТ

по научно-исследовательской работе

Тема: Использование методов ML в задачах визуальной одометрии

Студент гр. 3303 _____ Гэ Као

Руководитель _____ Кринкин К.В.

Санкт-Петербург
2018

1. Наименование научной работы

Использование методов ML в задачах визуальной одометрии

2. Исследовательская часть

2.1. Постановка задачи исследования

Алгоритм глубокого обучения (такой как LSTM, сеть CNN) используется для одновременной оценки позы 6-DoF монокулярной камеры и глубины сцены.

2.2. Актуальность работы

Вся сетевая структура включает в себя сеть оценки позы и сеть оценки глубины. Сеть оценки поз представляет собой глубоко повторяющуюся сверточную нейронную сеть (RCNN), которая реализует оценку монокулярной позы от конца до конца, состоящую из выделения признаков на основе сверточных нейронных сетей и моделирования временных рядов на основе повторяющихся нейронных сетей (RNN). Сеть оценки глубины генерирует плотные карты глубины, главным образом, на основе архитектуры кодера-декодера.

2.3. Проведённые исследования

Эта статья оценивает сквозную оценку позы и оценку глубины сцены монокулярной камеры путем создания сети глубокого обучения. Нейронная сеть состоит из сети оценки позы и сети оценки глубины. Как показано, каждая из двух сетей глубины принимает непрерывное монокулярное изображение в качестве входных данных и создает позу 6-DoF и глубину в качестве выходных сигналов соответственно.

В отличие от модели CNN, RNN может сохранять свою скрытую память состояний с течением времени и иметь петлю обратной связи между ними, поэтому он может использовать временную зависимость последовательности изображений, чтобы найти входные и предыдущие состояния соединения в последовательности. Эта структура хорошо подходит для решения проблем оценки позы, включающих модели временных рядов (модели движения) и ограничения контекста (последовательности изображений). Исходя из этого, в данной статье предлагается структура оценки кадра на основе RCNN.

Архитектура системы показана на рисунке 1. Пример размера тензора, приведенный на рисунке, основан на размере изображения набора данных KITTI. Система принимает серию видео или монокулярных изображений в качестве входных данных для изучения и прогнозирования жестовых переходов между 2 кадрами $t \rightarrow t + 1$. В частности, тензор изображения, сформированный путем укладки двух последовательных изображений в CNN, сначала вводится для генерации эффективного признака, используемого для оценки монокулярной позы, затем обучение хронированию выполняется через RNN, и обучение хронированию, наконец, выполняется для каждого изображения кадра. Выполните оценку осанки. Преимущество архитектуры на основе RCNN состоит в том, что комбинация CNN и RNN позволяет сети оценки позы монокулярного пространства выполнять извлечение признаков и моделирование синхронизации.

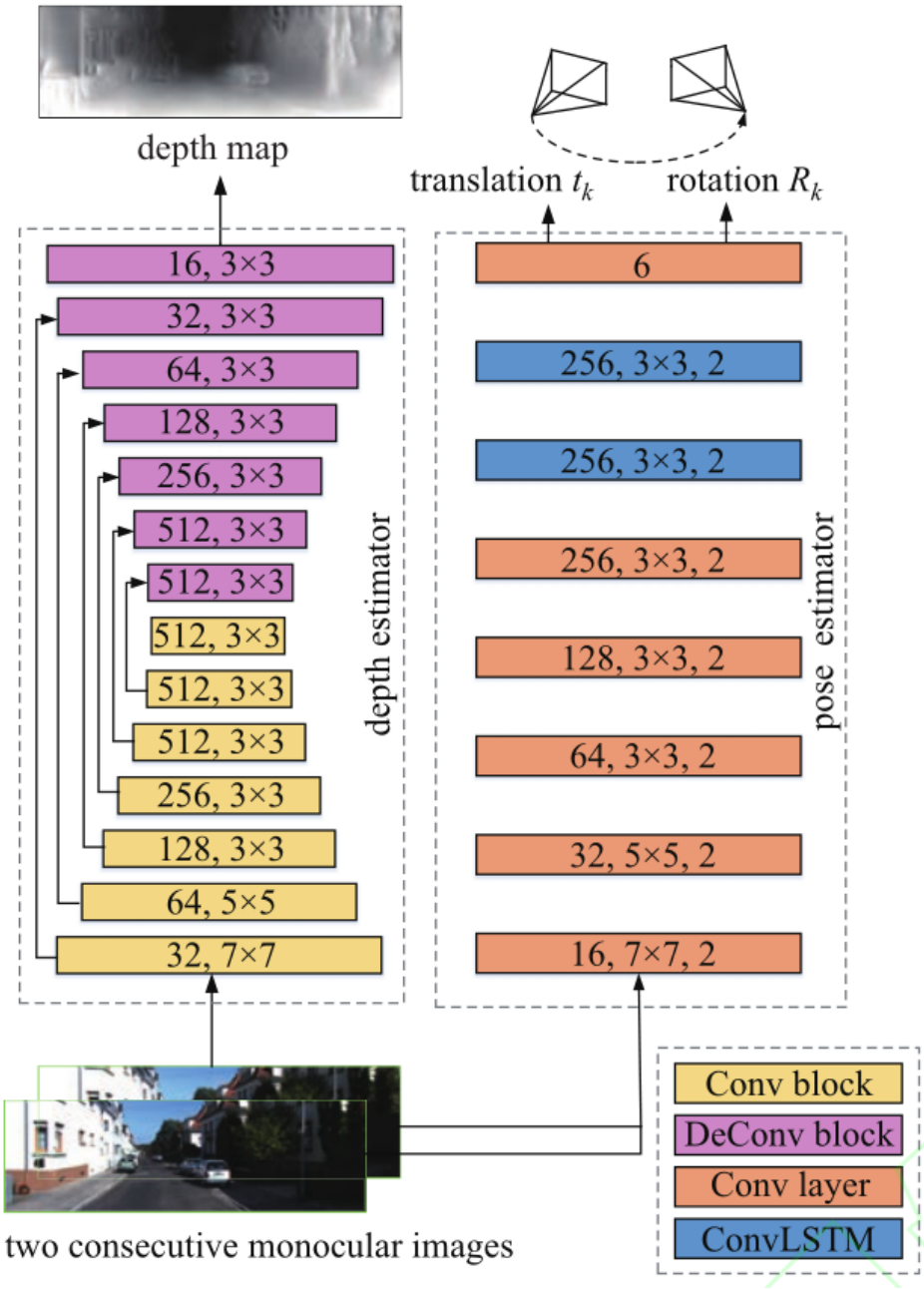

Рисунок 1. Архитектура оценки позы и оценки глубины

(1) Извлечение функции на основе CNN

Чтобы изучить эффективные особенности проблемы визуального одометра, в этой статье предлагается новая структура CNN для извлечения признаков двух последовательных монокулярных изображений RGB. Конкретная конфигурация CNN показана на рисунке 2, который имеет 5 слоев свертки. Размер ядра свертки в сети постепенно уменьшается с 7 × 7 до 3 × 3 для захвата мелких объектов. В то же время количество каналов (то

есть количество ядер свертки) увеличивается слой за слоем для изучения различных функций.

CNN использует необработанные изображения RGB вместо предварительно обработанных объектов (таких как оптические изображения или изображения глубины) в качестве входных данных для описания изображений RGB с высокой размерностью как векторов объектов с низкой размерностью, чтобы облегчить последующее моделирование синхронизации.

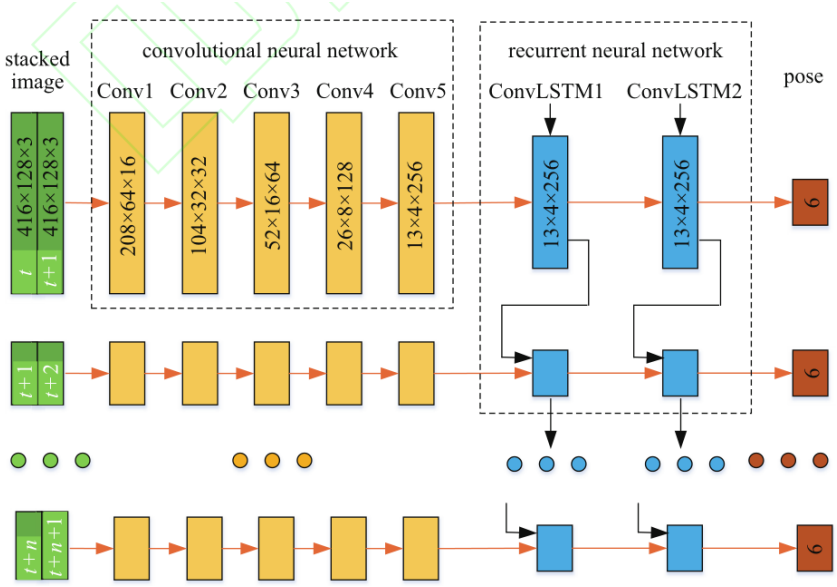

Рисунок 2 Архитектура предлагаемой сети оценки позы на основе RCNN

(2) RNN-моделирование временных рядов

После использования CNN для извлечения эффективных признаков, RNN вводится для моделирования временных отношений в последовательности изображений. Для проблемы использования CNN только для обнаружения и использования корреляции между изображениями, снятыми на длинных дорожках, в этой статье будет использоваться сверточная долговременная память, которая может изучать долговременные зависимости, вводя элементы памяти и ячейки. , ConvLSTM) сеть [1] в качестве основы RNN этой статьи, для достижения моделирования временных рядов изображений последовательности или видео. RNN может определить, отбрасывать или сохранять те предыдущие скрытые состояния, где сохраненное состояние будет использоваться для обновления текущего состояния и изучения модели движения.

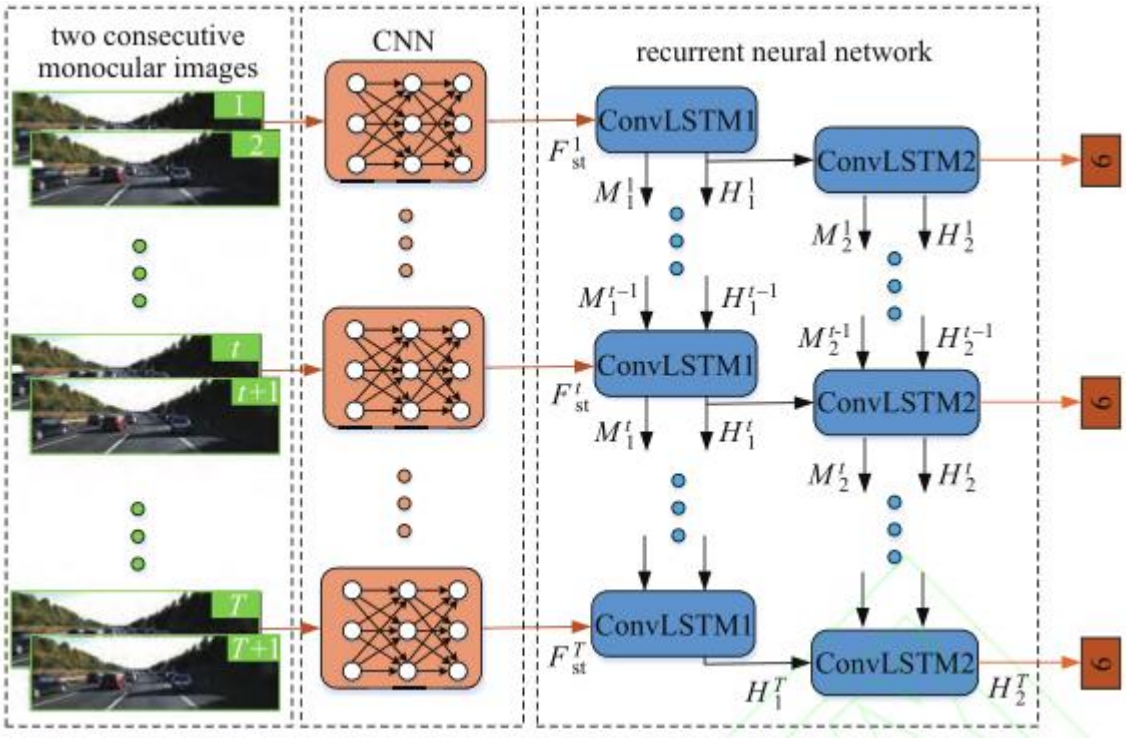

Рис.3 Архитектура уровня ConvLSTM

В этом методе глубокий RNN составляется путем объединения двух слоев ConvLSTM. Скрытое состояние ConvLSTM является входом другого ConvLSTM, как показано на рис 3. В момент времени t используется функция вывода пятого сверточного уровня (conv5). F_{st}^{t} , где *st* (пространственно-временное) представляет пространство-время, ячейка памяти M_1^{t-1}, полученная в вышеуказанный момент, скрытое состояние H_1^{t-1} используется в качестве входа уровня ConvLSTM1, а затем выводится уровень от H_1^t до ConvLSTM2. В то же время ConvLSTM2 имеет скрытое состояние слоя H_1^t , ячейку памяти предыдущего времени M_2^{t-1} и скрытое время предыдущего времени H_2^{t-1} в качестве входных данных, и его скрытое состояние H_2^t отправляется в слой свертки для генерации окончательной оценки позы.

Расширьте обычный LSTM, заменив оператор свертки (обозначенный как *) произведением Адамара (обозначенный как ◦), чтобы учесть временную корреляцию признаков CNN, введенных перед кадром t в динамической модели. Взяв уровень ConvLSTM1 в качестве примера, один блок ConvLSTM1 в кадре t можно описать как

$$I_1^t = \sigma(H_1^{t-1} * W_i^h + F_{st}^h * W_i^f + B_i) \quad (1)$$

$$A_1^t = \sigma(H_1^{t-1} * W_a^h + F_{st}^t * W_a^f + B_a) \quad (2)$$

$$O_1^t = \sigma(H_1^{t-1} * W_o^h + F_{st}^t * W_o^f + B_o) \quad (3)$$

$$G_1^t = \tanh(H_1^{t-1} * W_g^h + F_{st}^t * W_g^f + B_g) \quad (4)$$

$$M_1^t = F_1^t \circ M_1^{t-1} + I_1^t \circ G_1^t \quad (5)$$

$$H_1^t = O_1^t \circ \tanh M_1^t \quad (6)$$

Где σ и tanh - функции активации S и гиперболического тангенса. В уравнениях (1) - (6) $\{W_i^h, W_a^h, W_o^h, W_g^h, W_i^f, W_a^f, W_o^f, W_g^f\}$ и $\{B_i, B_a, B_o, B_g\}$ представляют веса и смещения в соответствующих сверточных слоях. Установите параметры ядра; I_1^t, A_1^t и O_1^t - входные данные для кадра ***t***, строб забытия и выходной шлюз; G_1^t, M_1^t, H_1^t - входная модуляция, единица хранения и скрытое состояние. Все они представлены трехмерным тензором размером 3 × 3 × 256 в сети оценки поз. Эта статья будет использовать двухслойный сверточный блок LSTM, чтобы узнать временную корреляцию многомерных объектов (3 × 3 × 256).

2.4. Результаты исследования

Для оценки предлагаемой системы CNN-LSTMVO в этой статье используется структура TensorFlow для реализации структуры сети и обучения с использованием графического процессора NVIDIA GTX 1080Ti. Тестовый ноутбук оснащен процессором NVIDIA GeForce GTX 1050Ti и процессором Intel Core i5 2,5 ГГц. В то же время, эта статья в основном использует набор данных KITTI для обучения и тестирования, а также использует набор данных Make3D [4] для оценки способности модели к обобщению базы данных.

В этой статье оптимизатор Адама используется для обучения сети, и итераций 150 000. Параметры β1 = 0,9 и β2 = 0,999, а скорость обучения составляет 0,0002. Длина последовательности входного изображения сети позы равна 2, а длина последовательности входного изображения сети глубины равна 1. Все эксперименты проводились с использованием последовательности изображений, снятых монокулярной камерой. Размер изображения, вводимого в сеть, составляет 416 × 128. Используйте различные виды методов улучшения

данных, чтобы повысить производительность и уменьшить возможные перестройки, такие как масштабирование изображения и обрезка изображения. В частности, 20% изображений выбираются случайным образом для масштабирования, а диапазон случайного масштабирования составляет [1, 1.15].

Оценка точности траектории

Чтобы оценить эффективность предложенной сети оценки поз и сравнить ее с методом SfMLearner [5], метод оценки в [5] сначала используется для оценки последовательности данных визуального одометра KITTI в последовательностях 09 и 10, где последовательность Длины 09 и 10 составляют 1591 кадр и 1201 кадр соответственно.

В этой статье система CNN-LSTMVO сравнивается с двумя вариантами монокулярного ORB-SLAM [6] и SfMLearner, двумя вариантами ORB-SLAM являются ORB-SLAM (full) и ORB-SLAM (short). ORB-SLAM (full) использует оценку положения во всех кадрах последовательности управления и позволяет обнаруживать и перемещать петлевые сигналы, в то время как ORB-SLAM (short) не включает в себя обнаружение и перенос петлевых сигналов. В этой статье абсолютная ошибка траектории (absolute trajectory error, ATE), используемая в методе SfMLearner, используется в качестве меры точности оценки позы. Как показано в Таблице 1, эффект этого метода лучше, чем у ORB-SLAM без обнаружения и перемещения с обратной связью, и он равен эффекту ORB-SLAM (full), что доказывает эффективность предложенного метода.

Таб.1 Абсолютная ошибка траектории (ATE) на одометрии KITTI Набор данных

Метод	последовательность 09 / м	последовательность 10 / м
ORB-SLAM (full)	0.014 ± 0.008	0.012 ± 0.011
ORB-SLAM (short)	0.064 ± 0.141	0.064 ± 0.130
SfMLearner	0.021 ± 0.017	0.020 ± 0.015
CNN-LSTMVO	0.014 ± 0.007	0.012 ± 0.008

ЗАКЛЮЧЕНИЕ

Предлагается новая система оценки глубины и позы монокулярного изображения, основанная на глубоком обучении без присмотра, а именно CNN-LSTMVO. Чтобы решить проблему, заключающуюся в том, что метод контролируемого визуального одометра требует больших данных обучения, а текущий метод обучения без наблюдения не может интегрировать контекстную информацию в оценку позы, предлагается синхронная поза, основанная на сверточной нейронной сети и RNN. Методы оценки и оценки глубины.

Посредством предлагаемой структуры глубокого обучения и функции ограничения этот метод может обеспечить точную оценку глубины позы и среды.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1]Shi X J, Chen Z R, Wang H, et al. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances in Neural Information Processing Systems. Canada: Neural Information Processing Systems Foundation,2015: 802-810.

[2] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving the KITTI vision benchmark suite,” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2012.

[3] A. Geiger, J. Ziegler, and C. Stiller, “Stereoscan: Dense 3D reconstruction in real-time,” in Intelligent Vehicles Symposium (IV), 2011.

[4] Saxena A, Sun M, Ng A Y. Make3D: Learning 3D scene structure from a single still image[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5): 824-840.

[5]Zhou T H, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2017: 6612-6621.

[6] Mur-Artal R, Montiel J M M, Tardos J D. ORB-SLAM: A versatile and accurate monocular SLAM system[J]. IEEE Transactions on Robotics, 2015, 31(5): 1147-1163.