УДК 004.932

ActSRF: How to segment images with uncertainty

В.С. Скороходов, Д.М. Дроздова, Д.А. Юдин

Московский физико-технический институт (национальный исследовательский университет)

В последнее время особую популярность в компьютерной графике представляет развитие моделей на основе дифференцируемого представления сцены [1] - по набору изображений и позиций камер генерировать сцену с других ракурсов. В оригинальной статье [1] для восстановления изображения используется MLP (multilayer perceptron) с ReLU в качестве функции активаций. На вход модель функцию позиционное кодирование от координат и направление камеры, а на выходе получают плотность и цвет. В [1] в качестве позиционного кодирования использовались гармонические функции. Но, как и в задачах обработки естественного языка, может помочь использование обучаемых параметров в позиционном кодировании. Авторы [3] предлагают использовать сложную процедуру с технологией хеширования, что позволяет увеличить качество и скорость реконструкции. В нашей модели мы воспользовались как раз таким позиционным кодированием.

Внедрение в нейросеть возможности вычисления неопределенности (uncertainty) помогает лучше понять, где алгоритм может ошибаться в своих предсказаниях. Существует несколько видов неопределенностей, мы восстанавливаем aleatoric uncertainty, которая отвечает за шум в данных. Используя идею из [4], мы добавили еще одну дополнительную голову, которая отвечает за восстановление дисперсии/неопределенности. Также из [4] мы взяли идею алгоритма для активного обучения.

Есть способы восстанавливать неопределенность для классификации или семантической сегментации (достаточно похожие задачи). Мы воспользовались идеей из [5] и добавили две дополнительные головы, выходы которых зависят только от позиции в пространстве, по аналогии с [2]. И с помощью них в каждой точке восстанавливаем набор $(mu_1, \ldots, mu_n)$ и b_sem. Считаем, что

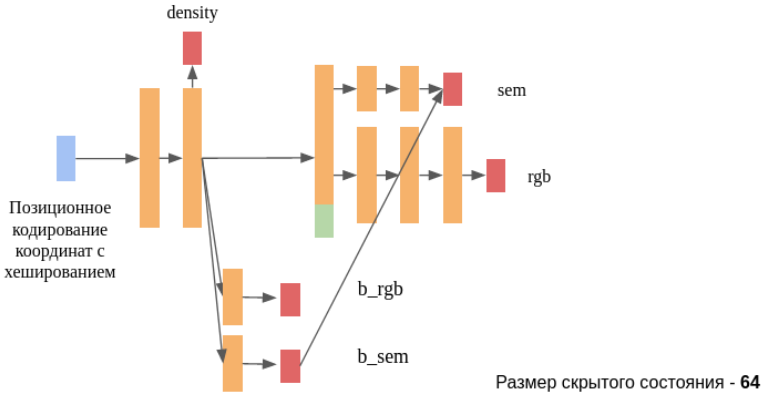
$$logit_i \sim N(sem_i, b\_sem)$$

$$\widehat{p_i} = \mathrm{E}(p_i), \text{ где } p_i = SoftMax(\{logit_1, \ldots, logit_n\})_i$$

Есть проблема в том, что мы не можем явно по формуле посчитать матожидание. Поэтому мы генерируем небольшое количество $logit$ из соответствующих распределений, считаем $SoftMax$, усредняем, и получаем оценку вероятностей. После этого считаем $CrossEntropy$. Можно подумать, что такое восстановление семантики оказывается неточным, однако в процессе обучения максимальное mu становится сильно больше 20, а b_sem менее 0.1, так что во время генерации и взятии SoftMax все равно наибольшей вероятностью будет обладать всегда только один класс. В наших экспериментах мы генерировали всего 10 * (число классов) нормальных случайных величин.

Схема модели представлена на рис.1

рис.1



Функция ошибки состоит из трех слагаемых:

$$L = (1 - \omega)L_{nerf} + \omega L_{uncert} + \lambda L_{semantic},$$

$$L_{nerf} = \frac{1}{N}\sum_{i=1}^{N} ||rgb(r_i) - \widehat{rgb(r_i)}||$$

$$L_{semantic} = CrossEntropy(\widehat{sem}, gt_{sem})$$

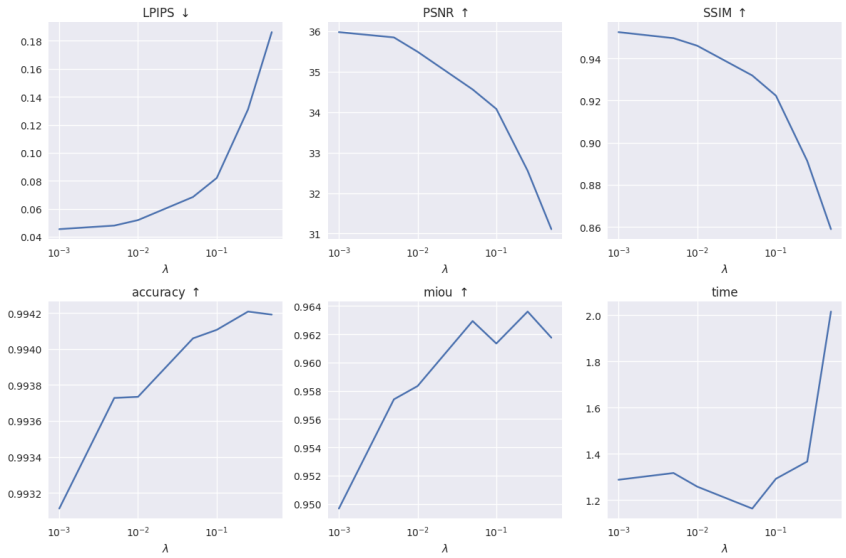
$$L_{uncert} = \frac{1}{N}\sum_{i=1}^{N} \frac{||rgb(r_i) - \widehat{rgb(r_i)}||}{2b_{rgb}^{\ 2}} + \frac{1}{2}\log b_{rgb}$$

Для обучения мы использовали датасет Replica [6] с размером картинки 640x480, который представляет из себя набор искусственно сгенерированных изображений комнат и офисов.

Эксперименты

К сожалению, пока что далеко не все эксперименты проведены. Мы сделали подбор гиперпараметра lambda рис.2, который отвечает за влияние семантического лосса, при этом без восстановления semantic uncertainty, то есть взяли b_sem = 0.

рис.2



Сейчас мы планируем понять, как влияет восстановление двух uncertainty и активное обучение. В будущем мы планируем добавить возможность восстановления epistemic uncertainty - оно направлено на измерение неопределенности модели в своем предсказании.

Финансирование

Исследование выполнено за счет гранта Российского научного фонда № 21-71-00131, https://rscf.ru/project/21-71-00131/.

Литература

1. Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.
2. Zhi, Shuaifeng, et al. "In-place scene labelling and understanding with implicit scene representation." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
3. Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." *ACM Transactions on Graphics (ToG)* 41.4 (2022): 1-15.
4. Pan, X., Lai, Z., Song, S., & Huang, G. (2022, November). ActiveNeRF: Learning Where to See with Uncertainty Estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII* (pp. 230-246). Cham: Springer Nature Switzerland.
5. Kendall, Alex, and Yarin Gal. "What uncertainties do we need in bayesian deep learning for computer vision?." *Advances in neural information processing systems* 30 (2017).
6. Straub, Julian, et al. "The Replica dataset: A digital replica of indoor spaces." *arXiv preprint arXiv:1906.05797* (2019).
7. Jiaxiang Tang, 2022, https://github.com/ashawkey/torch-ngp, Torch-ngp: a PyTorch of instant-ngp