Оценка позы человека как задача классификации

Ю. В. Проничкин [1] М. И. Кумсков[2]

В докладе рассмотрена актуальная задача машинного обучения - восстановление позы человека по его изображению. Предложен новый способ постановки задачи как задачи классификации каждого пикселя изображения. В предложенной постановке реализовано решение, проведено сравнение с существующим регрессионным подходом.

Ключевые слова: оценка позы человека, свёрточные нейронные сети, обнаружение ключевых точек

1. Актуальность задачи

Ключевым шагом к пониманию людей на изображениях и видео является точная оценка позы. Учитывая одно изображение RGB, мы хотим определить точное расположение в пикселях важных ключевых точек тела. Достижение понимания позы человека и артикуляции конечностей полезно для задач более высокого уровня, таких как распознавание действий, а также служит фундаментальным инструментом в таких областях, как взаимодействие человека с компьютером и анимация.

В статье [1] был предложен метод оценивания результатов и прогресса в обучении спортсменок в женских видах спорта при выполнении ими упражнений на основе системы захвата движений (motioncapture), и построении модели «проволочного человечка» -Stickman-a, являющейся моделью позы гимнастки. Такая модель позволяет проводить сравнение позы ученицы относительно эталонной позы. Сравнивание производится попарно, по всем кадрам видеозаписи выполнения упражнения ученицы и видео наставника.

[1] *Проничкин Юрий Викторович* — аспирант каф. вычислительной математики мех.-мат. ф-та МГУ, e-mail: uramur@mail.ru.

Pronichkin Iurii — graduate student, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of computational mathematics.

[2] *Кумсков Михаил Иванович* — д.ф.-м.н., профессор каф. вычислительной математики мех.-мат. ф-та МГУ, e-mail: kumskov@mail.ru.

Kumskov Mikhail — Doctor of Physico-mathematical Sciences, Lomonosov Moscow State University, Faculty of Mechanics and Mathematics, Chair of computational mathematics.

2. Существующие подходы к решению

Поза человека - набор из m пар координат (xi, yi) ключевых точек (например голова, кисти, плечи и т.д.) для каждого человека. Формально, кроме оценки позы (pose estimation) эта задача также эквивалентна задаче обнаружения ключевых (keypoint detection), использующаяся, например, для определения таких точек на лицах с последующей идентификацией человека.

Самый простой способ оценки позы - предсказывать эти координаты напрямую, т.е., имея на вход картинку, нейронная сеть должна выдавать для каждой опорной точки пару координат. Такой способ действительно просто реализуем, но с ним возникнут проблемы, как только на изображении окажется более двух человек, тем не менее есть способы, решающие эту проблему [2].

Задача оценки позы схожа с задачей семантической сегментации, в более современных работах был предложен новый способ. Он заключается в следующем: нейронная пытается предсказывать "карту уверенности позы"(pose heatmap), которая позднее декодируется в сами координаты [3].

Для каждого изображения карта уверенности - тензор размера Ho x Wo x Co, где Ho, Wo разрешение карты, а Co - m, то есть количество узлов (joints, keypoints), каждый элемент тензора - уверенность в то, что в соответствующем пикселе изображения находится та или иная ключевая точка.

Обучающий набор данных состоит из пар (Xi, Yi), в котором Xi - изображения, Yi - целевая карта уверенности. В современных алгоритмах для обучения используется карта, полученная следующим способом:

$$Y_{ij,xy} = \sum_{k\,=\,0}^{n} e^{-\frac{\left(x-x_{j_k}\right)^2+\left(y-y_{j_k}\right)^2}{2\sigma^2}}$$

эта величина моделирует плотность многомерного нормального распределения с независимыми одинаково-распределенными компонентами.

Здесь x_{j_k}, y_{j_k} - известные узлы (joints) на предварительно размеченном изображении, суммирование происходит по всем таким узлам (если например на изображении несколько человек), σ - параметр, обычно его берут равным 1.5, j - номер узла (например голова, левая рука, правая рука и т.д.). k - номер человека на изображении.

3. Предлагаемый подход

Во-первых, как и в задаче сегментации к целевому тензору добавляется канал, отвечающий за фон, т.е. такая матрица, элементы которой

являются вероятностями отсутствия всех узлов. Во-вторых, меняется целевая карта уверенности (k пробегает узлы, l пробегает позы):

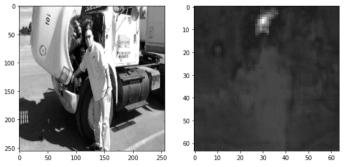
$$Y_{ij,xy} = \sum_k I\left[x = x_{j_k}\right] I\left[y = y_{j_k}\right] Y_{i(k+1),xy} = 1 - \sum_j \sum_l I\left[x = x_{j_l}\right] I\left[y = y_{j_l}\right]$$

И самое главное, сама задача меняется на задачу многоклассовой классификации. В такой постановке выход нейронной сети для каждого пикселя карты уверенности - это набор k + 1 честной вероятности, а не уверенности.

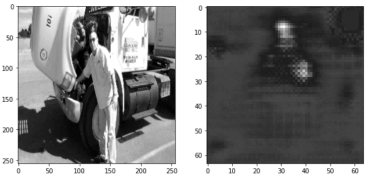
Как и во многих моделях задачи обнаружения объектов, здесь возникает проблема - дисбаланс классов - фона намного больше, чем всего остального. Эту проблему можно также решать с помощью техники hard negative mining [4], что и было сделано в этой работе. Применение этой техники дало результат - на выходе начали появляться достаточно хорошие карты уверенности.

Дальнейшее изучение поведения карт уверенности привело к следующей идее - выбрать функцию потерь, которая бы сбалансировала потерю на фоне и на не фоне, т.е. чтобы эти значения были одного порядка.

На рисунке ниже показаны карты уверенности модели после 15 эпох обучения, слева в предлагаемой классификационной постановке, справа в существующей регрессионной



Регрессия



Классификация

Работа выполнена при финансовой поддержке Минобрнауки РФ в рамках реализации программы Московского центра фундаментальной и прикладной математики по соглашению №075-15-2019-1621», Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект» и гранта РФФИ 19-07-00752

Список литературы

[1] Маъмуров Б.Б., Кумсков М.И., Проничкин Ю.В, Ильясова А.О, “Конструирование позы человека по фотографии на основе нейронных сетей глубокого обучения”, *Дурдона*, Материалы конференции Инновационные технологии в спорте и физическом воспитании под-

растающего поколения (нашриёти Бухара, Республика Узбекистан, 2019), **18**, Тезисы докладов, 2019, 63–65.

[2] Alexander Toshev, Christian Szegedy, *DeepPose: Human Pose Estimation via Deep Neural Networks*, arXiv: abs/1312.4659.

[3] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, *Convolutional Pose Machines*, arXiv: abs/1602.00134v4.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, *SSD: Single Shot MultiBox Detector*, arXiv: abs/1512.02325v5.

Human pose estimation as a classification problem
Pronichkin I.V., Kumskov M.I.

The report considers an urgent problem of machine learning – the human pose estimation based on his image. A new way of setting the problem as the problem of classifying each pixel of the image is proposed. In the proposed formulation, the solution is implemented, a comparison with the existing regression approach is made.

Keywords: human pose estimation, convolutional neural networks, keypoint detection.

References

[1] Mamurov B.B., Kumskov M.I., Pronichkin I.V., Ilyasova A.O., "Constructing a Human Pose from a Photo Based on Deep Learning Neural Networks", *Durdona*, Conference materials Innovative technologies in sports and physical education of the younger generation (Bukhara Publishing House, Republic of Uzbekistan, 2019), **18**, Тезисы докладов, 2019, 63–65 (In Russian).

[2] Alexander Toshev, Christian Szegedy, *DeepPose: Human Pose Estimation via Deep Neural Networks*, arXiv: abs/1312.4659.

[3] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, Yaser Sheikh, *Convolutional Pose Machines*, arXiv: abs/1602.00134v4.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, *SSD: Single Shot MultiBox Detector*, arXiv: abs/1512.02325v5.