Team: igen

Balázs Frey (balazs.frey@gmail.com)
Gábor Révy (revy.gabor@gmail.com)
Dániel Szekeres (szekeresdani@gmail.com)
Dániel Gábor Vankó (vankodaniel97@gmail.com)

1st Milestone Documentation

Homework
Practical deep learning
(BMEVITMAV45)

YOLOCR

Conception

Our main goal is to create a optical character recognition based on YOLO[1][2][3] architecture. We would like to achieve the best possible speed and accuracy by refining YOLO.

Data for training

We use SynthText and its pregenerated dataset, which includes approximately 800 thousands synthetic scene-text images and annotations with word-level and character-level bounding-boxes. However, these bounding-boxes are given as simple quadrilaterals and needs converting to rectangles.

Why SynthText?

There are a number of reasons why we chose SynthText above other datasets. SynthText provides detailed, character-level ground-truth annotations while other only contains word-level image regions and are unsuitable for training detectors.[4] Moreover, it provides much more realistic images unlike putting words on images without any transformation.

Bounding-box transformation

We want to train our model for recognising rectangle bounding-boxes, and in order to do that we have to convert the bounding-boxes given by SynthText. We take the longest side of the given quadrilateral and it will be overlapped by our rectangle's longer side. After that we make sure that all points of the original bounding-box will be inside our new bounding-rectangle.

Our model

Output

The model's output is `[center_x, center_y, width, height, rotation]`, where

- `center_x` and `center_y` are the coordinates of the center of a character's bounding rectangle,
- `width` is the rectangle's width,
- `height` is the rectangle's height,
- `rotation` is the angle of the rectangle' rotation.

This is the reason why we have to convert SynthText's bounding quadrilaterals to rectangles.

Subgoals

We set out these subgoals which you can read below. Our model should recognise:

1. Untransformed characters before gradient background *(we use gradient to preliminary simulate shadows)*.
2. Rotated and/or distorted characters.
3. Synthetic scene-text, finally.

---

[1] https://arxiv.org/abs/1506.02640v5
[2] https://arxiv.org/abs/1612.08242v1
[3] https://arxiv.org/abs/1804.02767v1
[4] http://www.robots.ox.ac.uk/~vgg/data/scenetext/gupta16.pdf