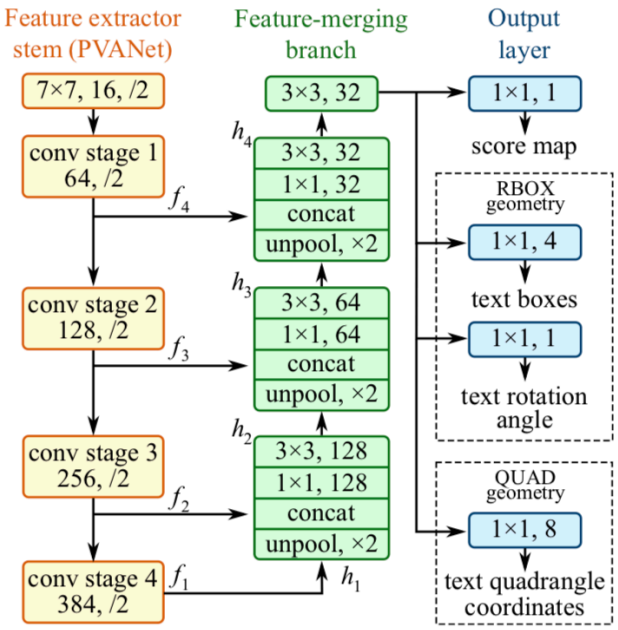
Постановка задачи сравнения нейросетевых и непрерывно-морфологических методов детекции текста

Гайдученко Н. Е., Труш Н. А, Торлак А. В., Миронова Л. Р., Акимов К. М.,
Гончар Д. А.

Модель **EAST** может быть представлена в виде трех основных частей, схема приведена на рис. 1



1. Feature extractor stem

Стержень представляет собой сверточную сеть, предварительно обученную на наборе данных ImageNet[1]. Четыре уровня feature map, обозначенных как f_i , извлекаются из стержня. Их размеры равны 1/32, 1/16, 1/8 и 1/4 от исходного изображения соответсвенно.

2. Feature-merging branch

Далее мы постепенно объединяем их, пользуясь формулой

$$g_i = \begin{cases} \text{unpool}(h_i) & \text{if} \quad i \leq 3 \\ \text{conv}_{3\times3}(h_i) & \text{if} \quad i = 4 \end{cases} \tag{1}$$

$$h_i = \begin{cases} f_i & \text{if} \quad i = 1 \\ \text{conv}_{3\times3}(\text{conv}_{1\times1}([g_{i-1}; f_i])) & \text{otherwise} \end{cases} \tag{2}$$

где g_i - основание, а h_i - влитая feature map (оператор [x;x] является конкатенацией)

На каждом этапе слияния - feature map с предыдущего этапа проходит через unpool слой, увеличивающий ее размер в 2 раза. Далее она конкатенируется с текущей feature map. Затем слой *conv1×1*[2] сокращает число каналов и объем вычислений, за которым следует слой *conv3×3*, который объединяет информацию для окончательного получения выходных данных этой стадии слияния. После последней стадии слияния слой *conv3×3* создает окончательную feature map ветви слияния и передает ее на выходной слой.

3. Output layer

Количество выходных каналов для каждой свертки показано на рис. 2

Geometry	channels	description
AABB	4	$\mathbf{G} = \mathbf{R} = \{d_i \mid i \in \{1,2,3,4\}\}$
RBOX	5	$\mathbf{G} = \{\mathbf{R}, \theta\}$
QUAD	8	$\mathbf{G} = \mathbf{Q} = \{(\Delta x_i, \Delta y_i) \mid i \in \{1,2,3,4\}\}$

Мы поддерживаем небольшое количество каналов для сверток в ветви, что добавляет небольшую часть накладных вычислительных расходов на стержень и делает сеть эффективной. Выходной слой содержит несколько операций *conv1×1* для проецирования 32 каналов feature map в 1 канал score map *Fs* и многоканальную *Fg*.
Для RBOX - геометрия представлена 4 каналами выровненной по осям прямоугольной рамкой (AABB) $\boldsymbol{R}$ и 1 каналом, который представляет угол поворота ϑ.
Для QUAD $\boldsymbol{Q}$ используется 8 чисел для обозначения сдвига координат из четырех угловых вершин $\{p_i \mid i \in \{1,2,3,4\}\}$ четырехугольника к месту расположения пикселя. Поскольку каждое смещение расстояния содержит два числа $(\Delta x_i, \Delta y_i)$, то вывод содержит 8 каналов.

---

[1] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei- Fei. Imagenet: A large-scale hierarchical image database. In Proc. of CVPR, 2009.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385, 2015.