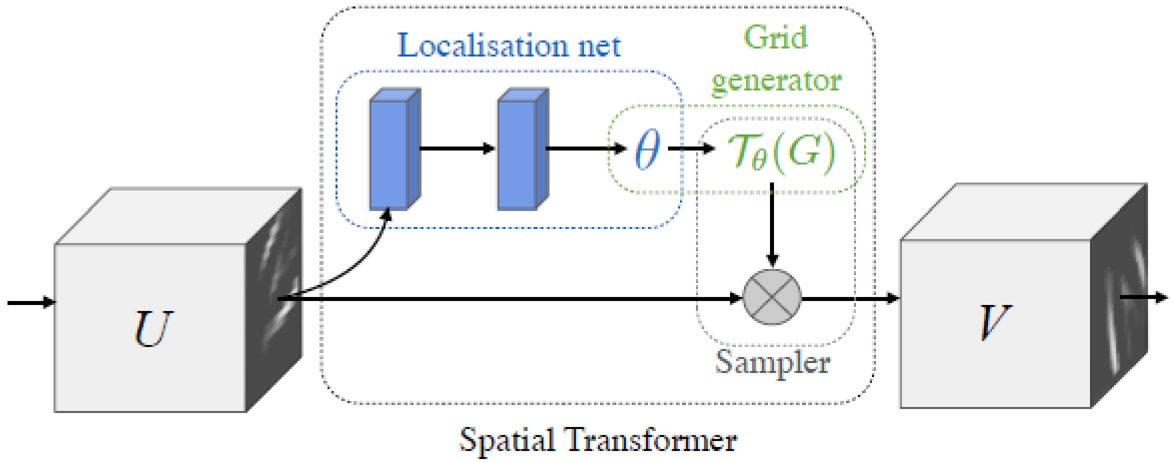


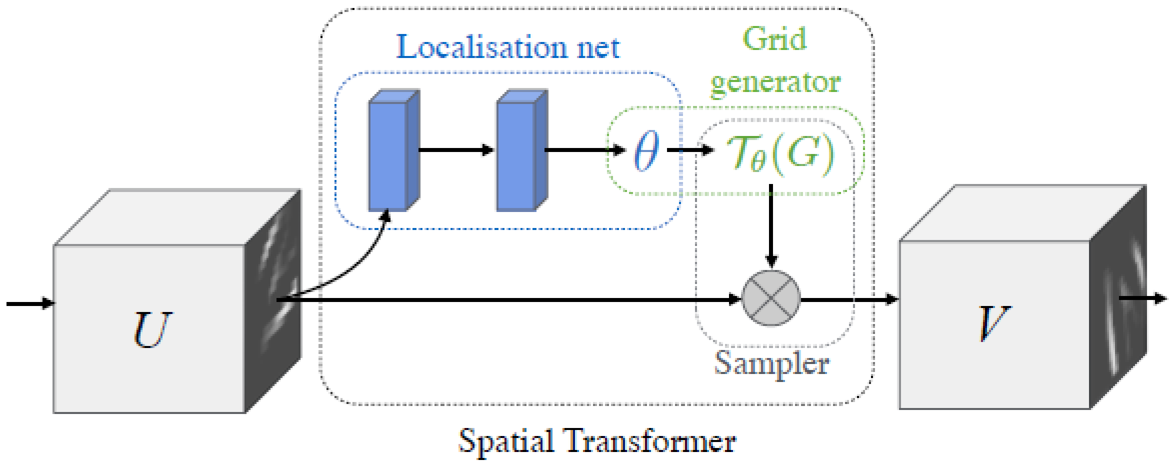
1、Spatial Transformer Networks

2、ASTER: An Attentional Scene Text Recognizer with Flexible Rectification

主讲人：杜臣

2018.07.21

Spatial Transformer Networks

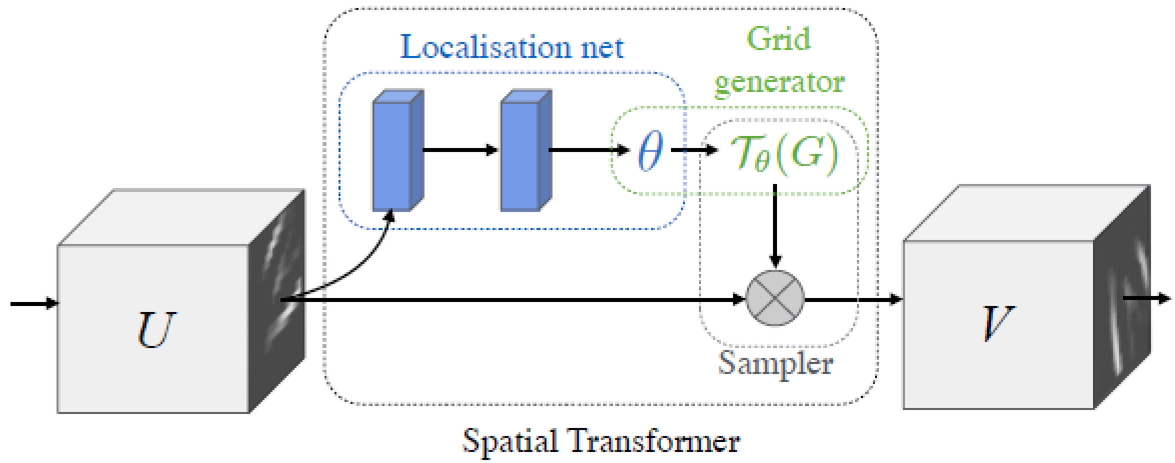


1、Localisation Network

2、Parameterised Sampling Grid

3、Differentiable Image Sampling

Localisation Network

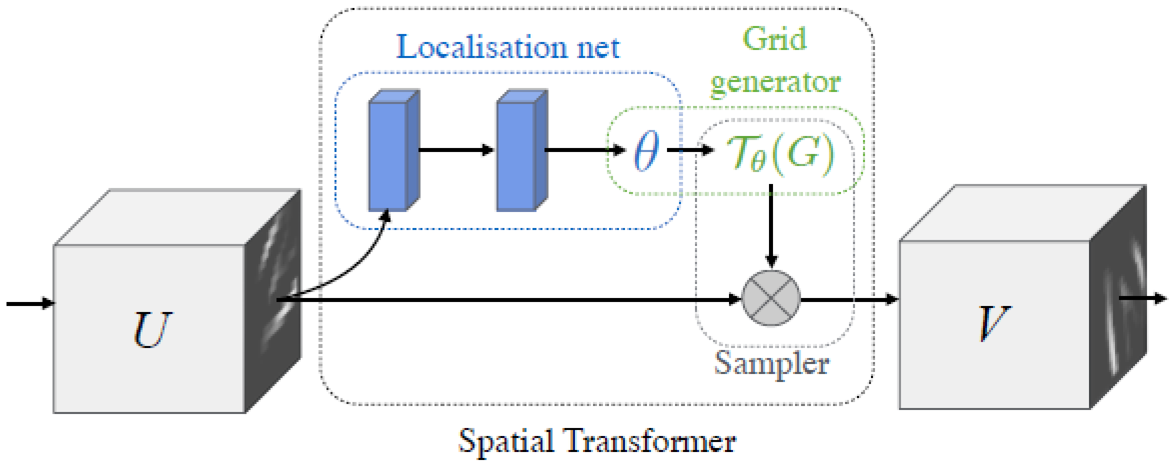


3.1 Localisation Network

The localisation network takes the input feature map $U \in \mathbb{R}^{H \times W \times C}$ with width W, height H and C channels and outputs θ, the parameters of the transformation $\mathcal{T}_\theta$ to be applied to the feature map: $\theta = f_{\text{loc}}(U)$. The size of θ can vary depending on the transformation type that is parameterised, *e.g.* for an affine transformation θ is 6-dimensional as in (1).

The localisation network function $f_{\text{loc}}()$ can take any form, such as a fully-connected network or a convolutional network, but should include a final regression layer to produce the transformation parameters θ.

Parameterised Sampling Grid



section). By *pixel* we refer to an element of a generic feature map, not necessarily an image. In general, the output pixels are defined to lie on a regular grid $G = \{G_i\}$ of pixels $G_i = (x_i^t, y_i^t)$, forming an output feature map $V \in \mathbb{R}^{H' \times W' \times C}$, where H' and W' are the height and width of the grid, and C is the number of channels, which is the same in the input and output.

For clarity of exposition, assume for the moment that $\mathcal{T}_\theta$ is a 2D affine transformation $\mathtt{A}_\theta$. We will discuss other transformations below. In this affine case, the pointwise transformation is

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = \mathcal{T}_\theta(G_i) = \mathtt{A}_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \tag{1}$$

Parameterised Sampling Grid

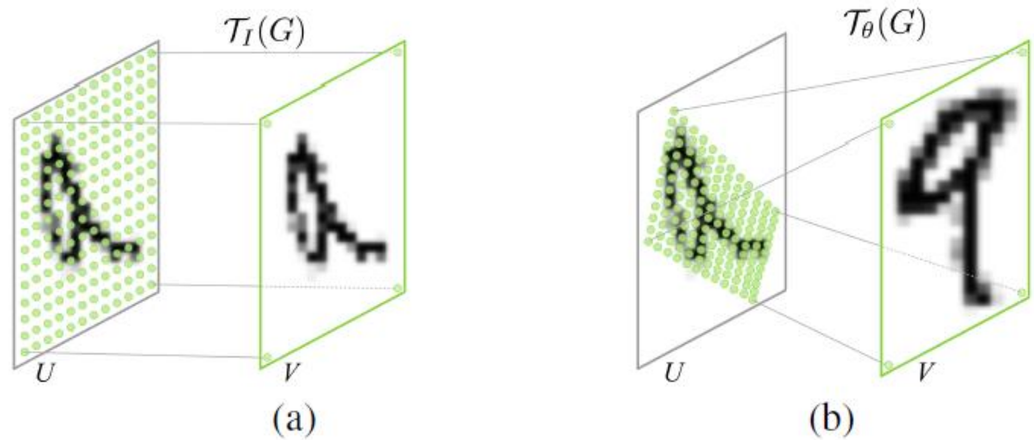


Figure 3: Two examples of applying the parameterised sampling grid to an image U producing the output V. (a) The sampling grid is the regular grid $G = \mathcal{T}_I(G)$, where I is the identity transformation parameters. (b) The sampling grid is the result of warping the regular grid with an affine transformation $\mathcal{T}_\theta(G)$.

Differentiable Image Sampling

To perform a spatial transformation of the input feature map, a sampler must take the set of sampling points $\mathcal{T}_\theta(G)$, along with the input feature map U and produce the sampled output feature map V.

Each (x_i^s, y_i^s) coordinate in $\mathcal{T}_\theta(G)$ defines the spatial location in the input where a sampling kernel is applied to get the value at a particular pixel in the output V. This can be written as

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c k(x_i^s - m; \Phi_x) k(y_i^s - n; \Phi_y) \ \ \forall i \in [1 \dots H'W'] \ \ \forall c \in [1 \dots C]$$

where Φ_x and Φ_y are the parameters of a generic sampling kernel $k()$ which defines the image interpolation (*e.g.* bilinear), U_{nm}^c is the value at location (n, m) in channel c of the input, and V_i^c is the output value for pixel i at location (x_i^t, y_i^t) in channel c. Note that the sampling is done identically for each channel of the input, so every channel is transformed in an identical way (this preserves spatial consistency between channels).

Differentiable Image Sampling

1、最近邻插值

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \delta(\lfloor x_i^s + 0.5 \rfloor - m)\delta(\lfloor y_i^s + 0.5 \rfloor - n)$$

2、线性插值

$$V_i^c = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

3、双线性插值

$$\frac{\partial V_i^c}{\partial U_{nm}^c} = \sum_n^H \sum_m^W \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|)$$

$$\frac{\partial V_i^c}{\partial x_i^s} = \sum_n^H \sum_m^W U_{nm}^c \max(0, 1 - |y_i^s - n|) \begin{cases} 0 & \text{if } |m - x_i^s| \geq 1 \\ 1 & \text{if } m \geq x_i^s \\ -1 & \text{if } m < x_i^s \end{cases}$$

ASTER: An Attentional Scene Text Recognizer with Flexible Rectification

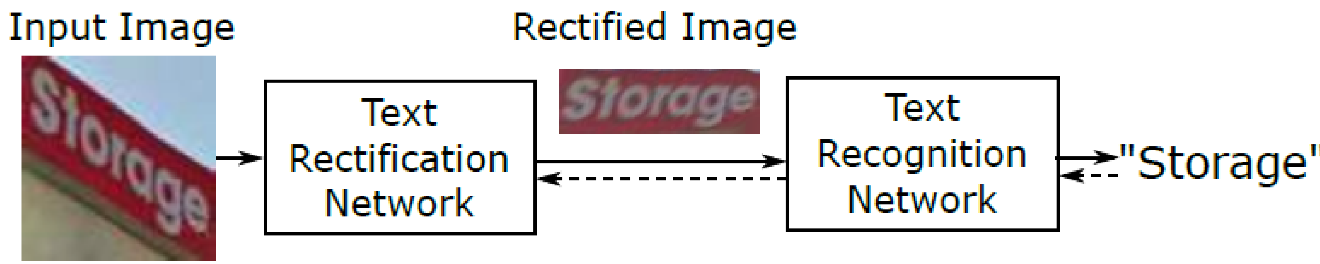


Fig. 2. Overview of the proposed model. Dashed lines show the flow of gradients.

1、Rectification Network

  1、Localisation Network

  2、Grid Generator

  3、Sampler

2、Recognition Network

背景

Fig. 1. Examples of irregular text.

场景文字识别中经常出现弯曲、仿射变换等形变的字符串。

采用一个矫正网络加一个识别网络实现倾斜场景文本的识别。

Rectification Network

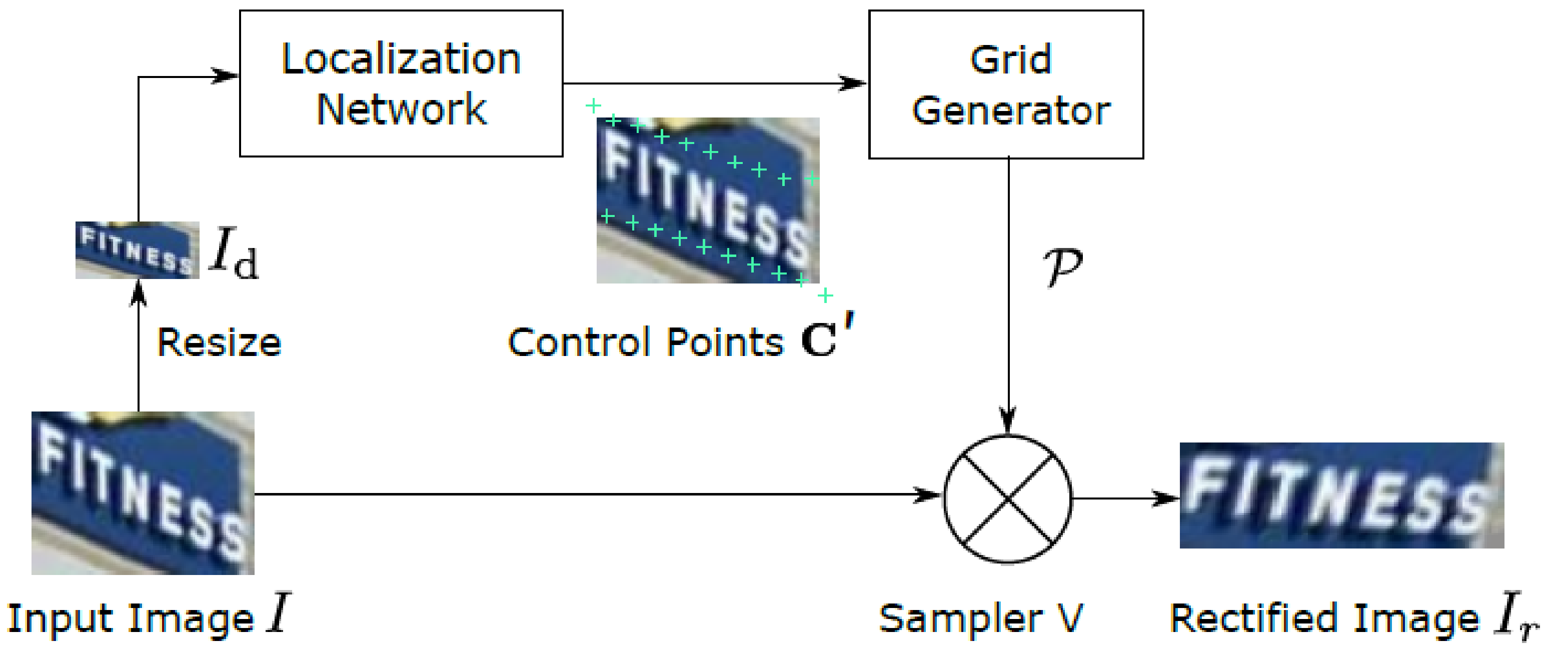


Fig. 4. Structure of the rectification network.

Rectification Network

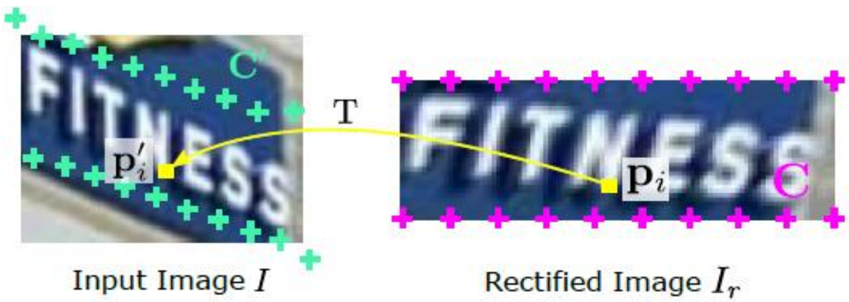
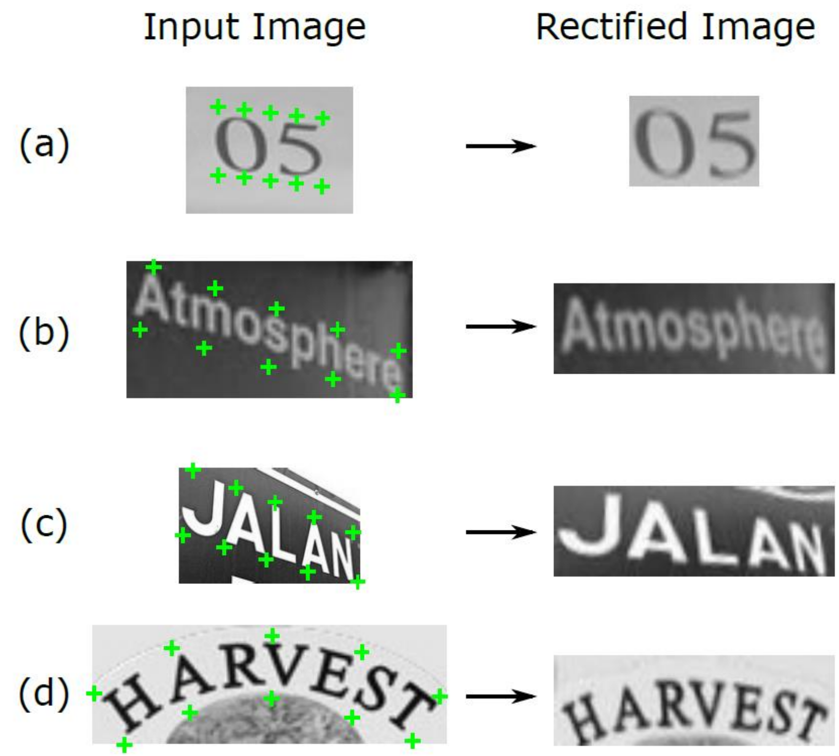
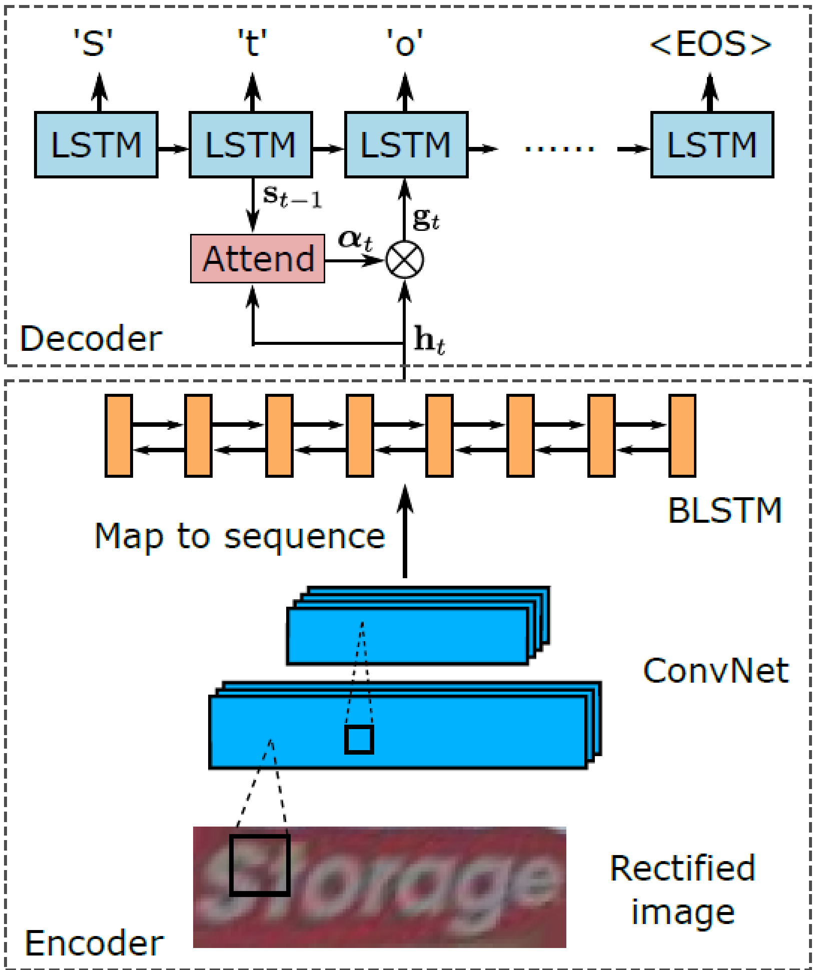


Fig. 5. Text rectification with TPS transformation. Crosses are control points. The yellow arrow represents the transformation T, connecting a point $\mathbf{p}_i$ and its corresponding point $\mathbf{p}'_i$.



Recognition Network



$$e_{t,i} = \mathbf{w}^\intercal \tanh\left(\mathbf{W}\mathbf{s}_{t-1} + \mathbf{V}\mathbf{h}_i + b\right)$$

$$\alpha_{t,i} = \exp\left(e_{t,i}\right) / \sum_{i'=1}^{n} \exp(e_{t,i'})$$

$$\mathbf{g}_t = \sum_{i=1}^{n} \alpha_{t,i}\mathbf{h}_i$$

$$(\mathbf{x}_t, \mathbf{s}_t) = \mathrm{rnn}\left(\mathbf{s}_{t-1}, \left(\mathbf{g}_t, f\left(y_{t-1}\right)\right)\right)$$

$$p(y_t) = \mathrm{softmax}\left(\mathbf{W}_o\mathbf{x}_t + b_o\right)$$
$$y_t \sim p(y_t)$$

Fig. 7. Structure of the basic text recognition network.

Recognition Network

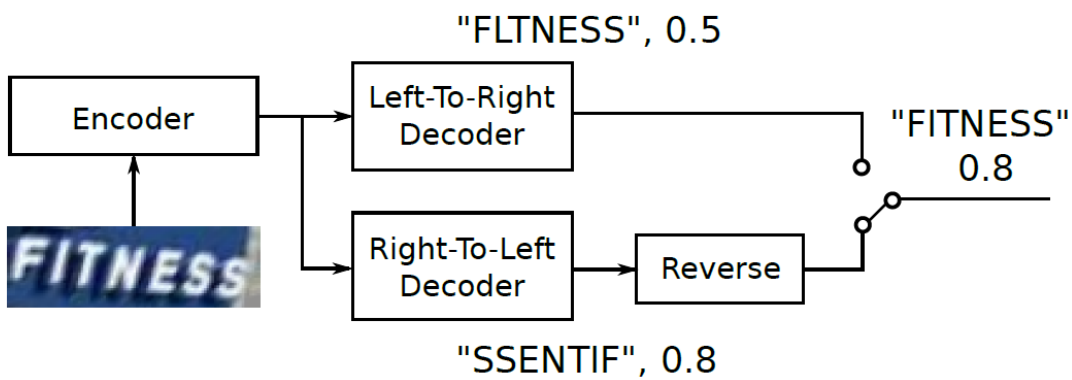


Fig. 8. Bidirectional decoder. “0.5” and “0.8” are recognition scores.

3.3 Training

The model is trained end to end under a multi-task setting, whose objective is

$$L = -\frac{1}{2}\sum_{t=1}^{T}(\log p_{\mathrm{ltr}}(y_t|I) + \log p_{\mathrm{rtl}}(y_t|I)) \tag{8}$$

References

[1] Spatial Transformer Networks

[2] ASTER: An Attentional Scene Text Recognizer with Flexible Rectification