Spatial Transformer Networks

Max Jaderberg Karen Simonyan Andrew Zisserman Koray Kavukcuoglu

Google DeepMind, London, UK

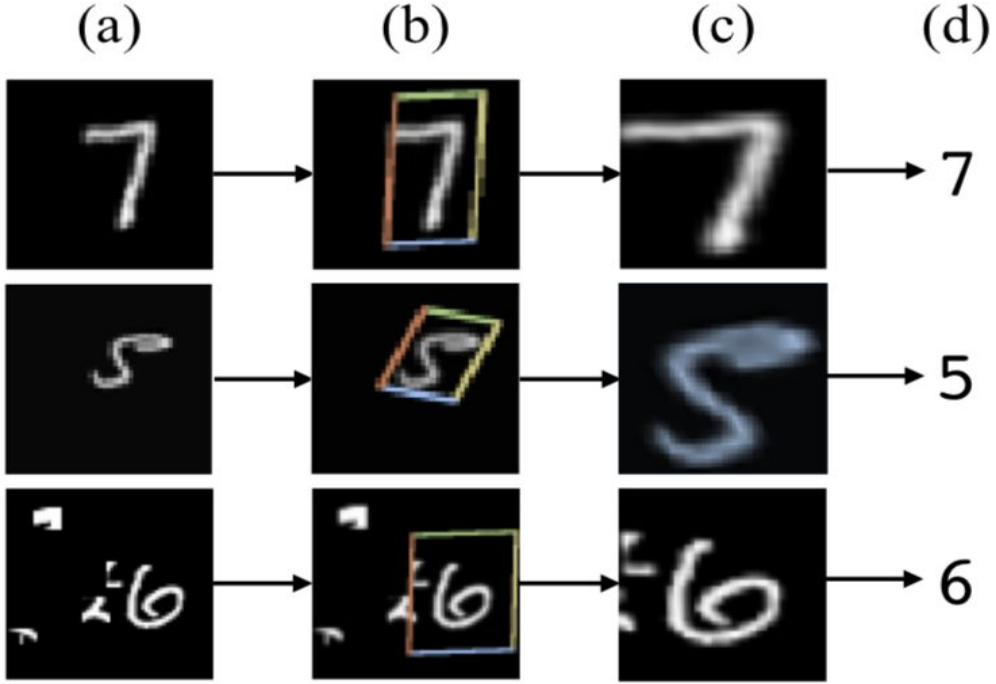
{jaderberg,simonyan,zisserman,korayk}@google.com

Горячко Виктор

Содержание.

1. Мотивация.
2. Строение Spatial transformer.
   a. Localisation net.
   b. Grid generator.
   c. Sampler.
3. Эксперименты.
   a. Distorted MNIST.
   b. Street View House Numbers.
   c. Fine-Grained Classification.
   d. MNIST Addition.
   e. Co-localisation.
   f. Higher Dimensional Transformers.
4. Заключение.

Мотивация.



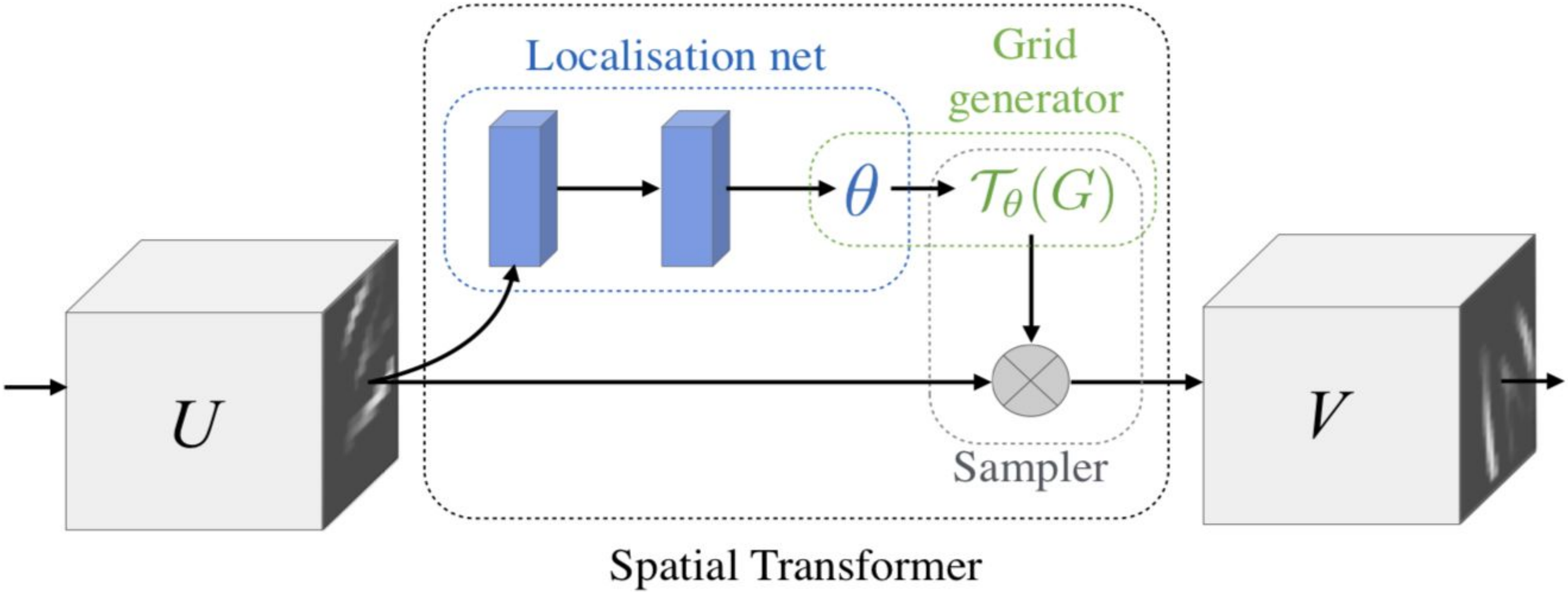
CNN are not actually invariant to large transformations of the input data

The pooling operation used in convolutional neural networks is a big mistake and the fact that it works so well is a disaster. (Geoffrey Hinton, Reddit AMA)

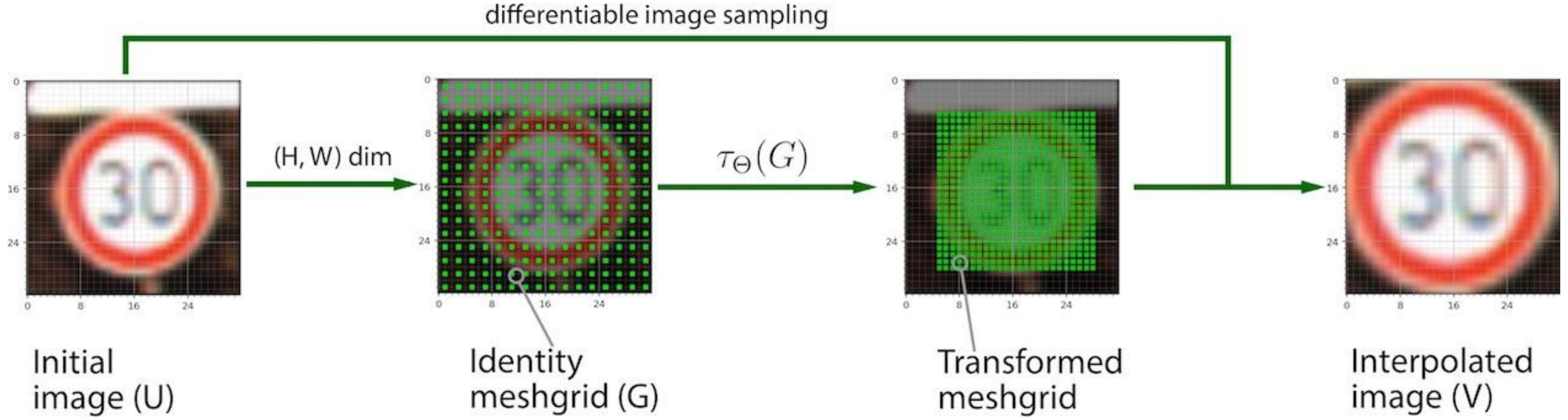
Spatial Transformer Networks can be used for:

- image classification
- co-localisation
- spatial attention

Строение Spatial transformer.



Применение STN преобразования в 4 шага при известной матрице линейных преобразований θ.

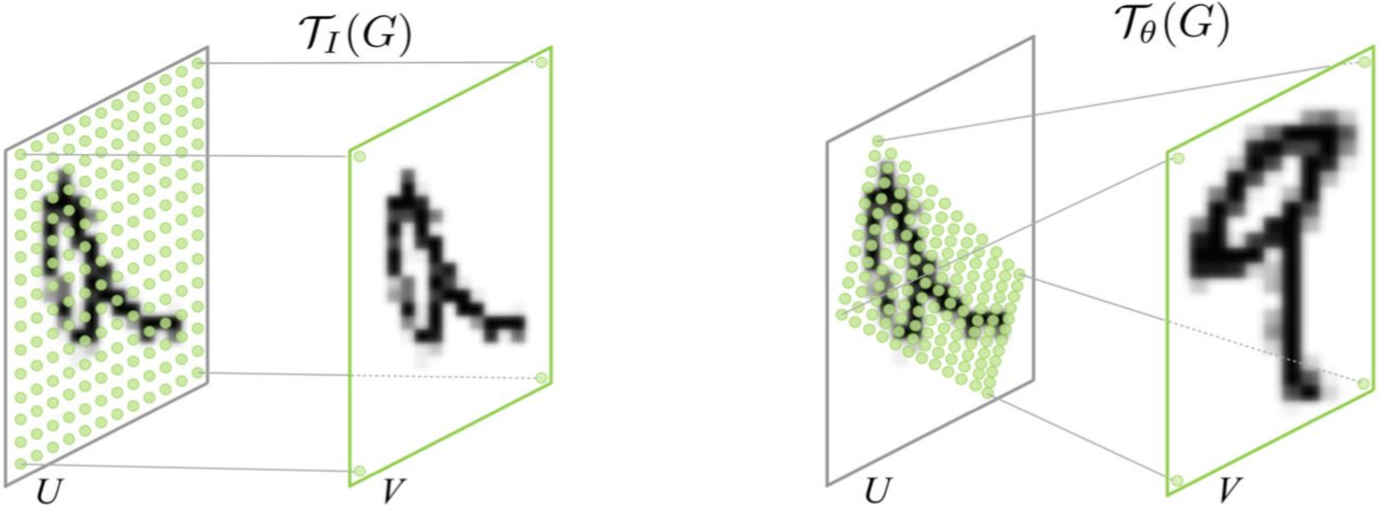


Localisation net.

- **input**: feature map U of shape (H, W, C)
- **output**: transformation matrix θ
- **architecture**: fully-connected network or ConvNet as well.

Grid generator.

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



(a) (b)

Sampler.

x coordinate in $\tau_\Theta(G)$

parameters of sampling kernel

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

pixel in a channel ***c***

channels

value at location ***(n, m)*** in channel ***c*** of input ***U***

interpolation kernel

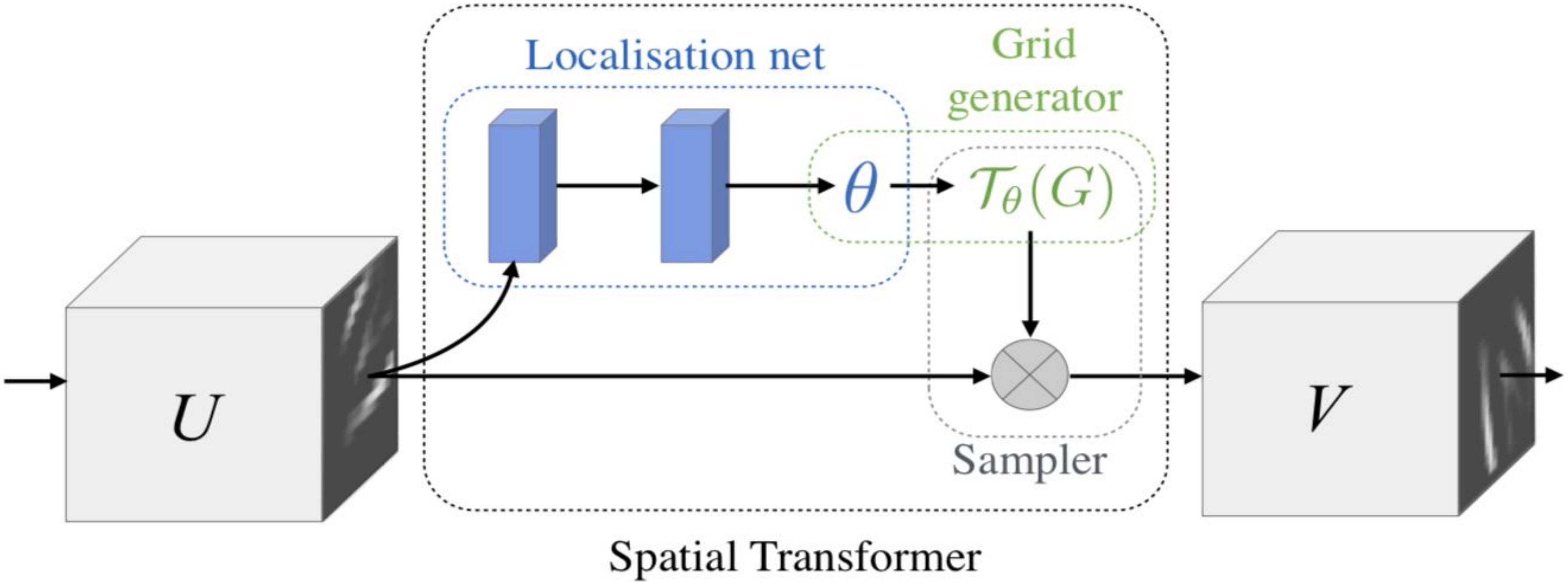
Sampler.

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

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

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

Строение Spatial transformer.



Эксперименты.

Projective transformation (Proj)

$$\begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}$$

Эксперименты.

16-point thin plate spline transformation (TPS)

$$I_f = \iint_{\mathbb{R}^2} (f_{xx}^2 + 2f_{xy}^2 + f_{yy}^2)\,dx\,dy$$

$$f(x, y) = a_1 + a_x x + a_y y + \sum_{i=1}^{p} w_i U\left(\|(x_i, y_i) - (x, y)\|\right)$$

$$U(r) = r^2 \log r. \quad \sum_{i=1}^{p} w_i x_i = \sum_{i=1}^{p} w_i y_i = 0 \quad \sum_{i=1}^{p} w_i = 0$$

Эксперименты.

where $K_{ij} = U(\|(x_i, y_i) - (x_j, y_j)\|)$, the ith row of P is $(1, x_i, y_i)$,

We will denote the $(p+3) \times (p+3)$ matrix of this system by L;

$$\begin{bmatrix} K & P \\ P^T & O \end{bmatrix} \begin{bmatrix} w \\ a \end{bmatrix} = \begin{bmatrix} v \\ o \end{bmatrix}$$

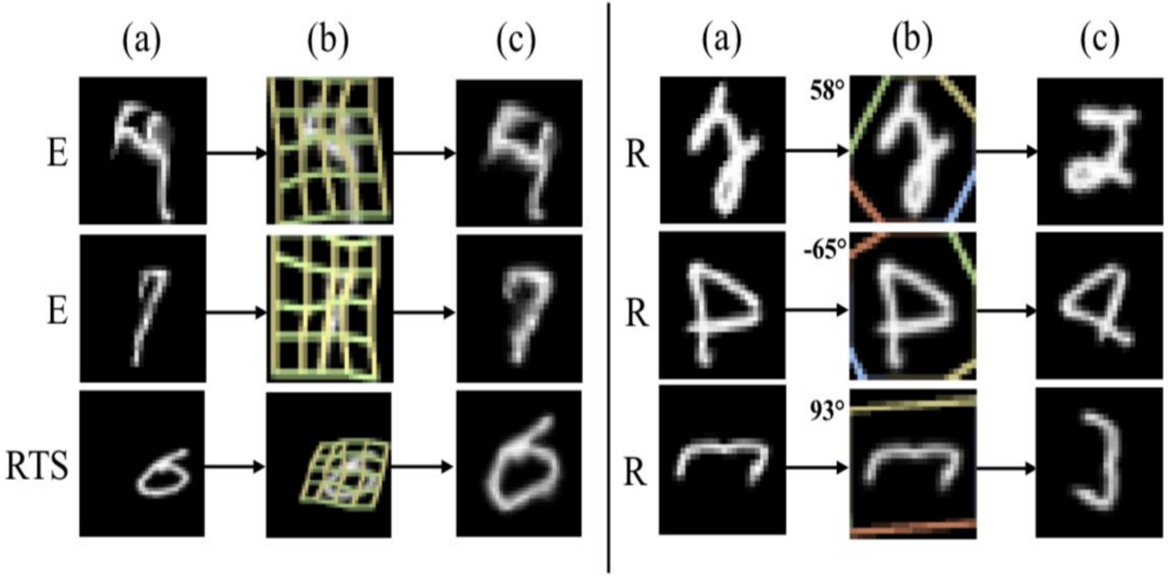
$$I_f \propto v^T L_p^{-1} v = w^T K w$$

Эксперименты.

Distorted MNIST

Model		R	RTS	P	E
FCN		2.1	5.2	3.1	3.2
CNN		1.2	0.8	1.5	1.4
ST-FCN	Aff	1.2	0.8	1.5	2.7
	Proj	1.3	0.9	1.4	2.6
	TPS	1.1	0.8	1.4	2.4
ST-CNN	Aff	0.7	0.5	0.8	1.2
	Proj	0.8	0.6	0.8	1.3
	TPS	0.7	0.5	0.8	1.1

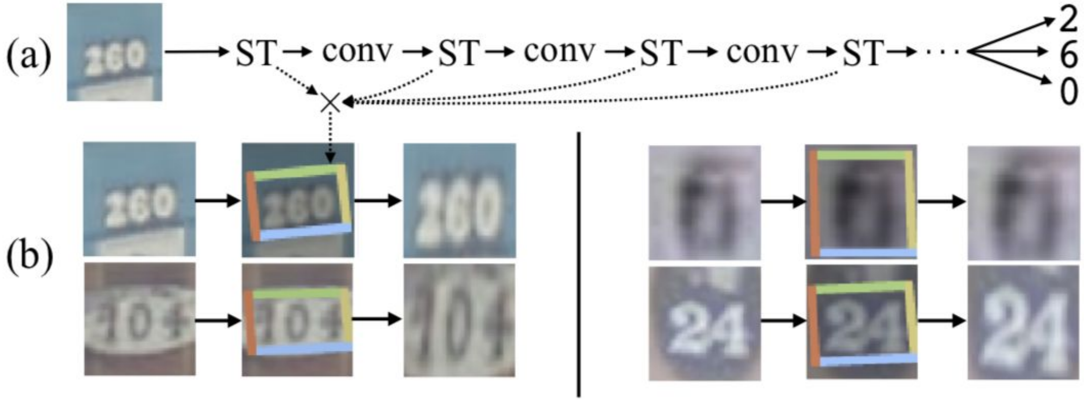
(MNIST Distortion: R, RTS, P, E)



Эксперименты.

Street View House Numbers

Model		Size 64px	Size 128px
Maxout CNN [13]		4.0	-
CNN (ours)		4.0	5.6
DRAM* [1]		3.9	4.5
ST-CNN	Single	3.7	**3.9**
	Multi	**3.6**	**3.9**



The CNN model is: conv[48,5,1,2]-max[2]-conv[64,5,1,2]-conv[128,5,1,2]-max[2]-conv[160,5,1,2]-conv[192,5,1,2]-max[2]-conv[192,5,1,2]-conv[192,5,1,2]-max[2]-conv[192,5,1,2]-fc[3072]-fc[3072]-fc[3072].

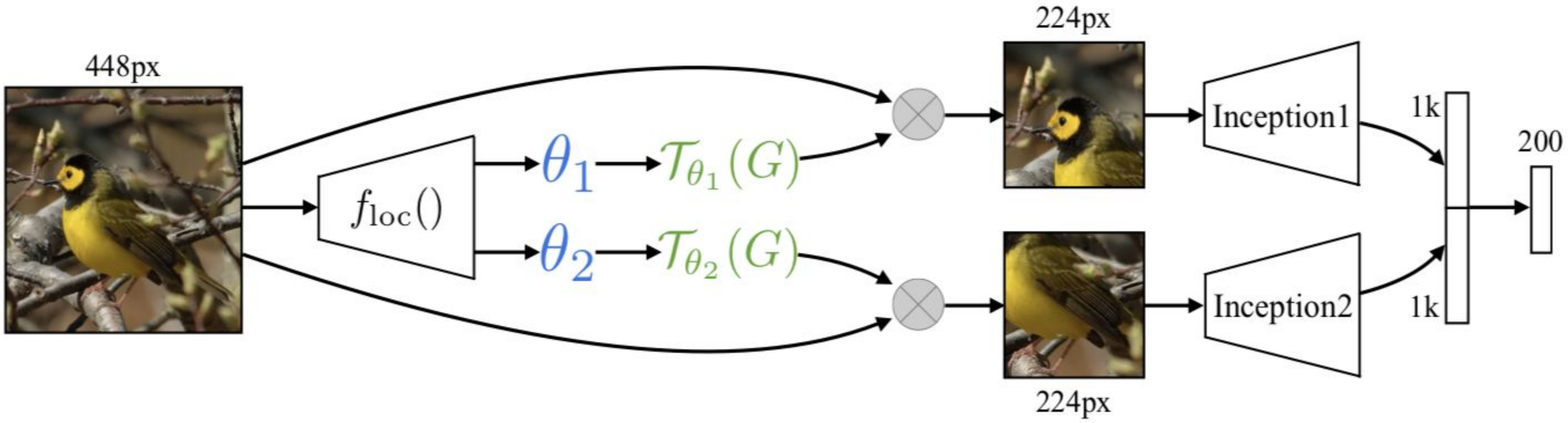
ST’s localisation network architecture is as follows: conv[32,5,1,2]-max[2]-conv[32,5,1,2]-fc[32]-fc[32].

Эксперименты.

Fine-Grained Classification

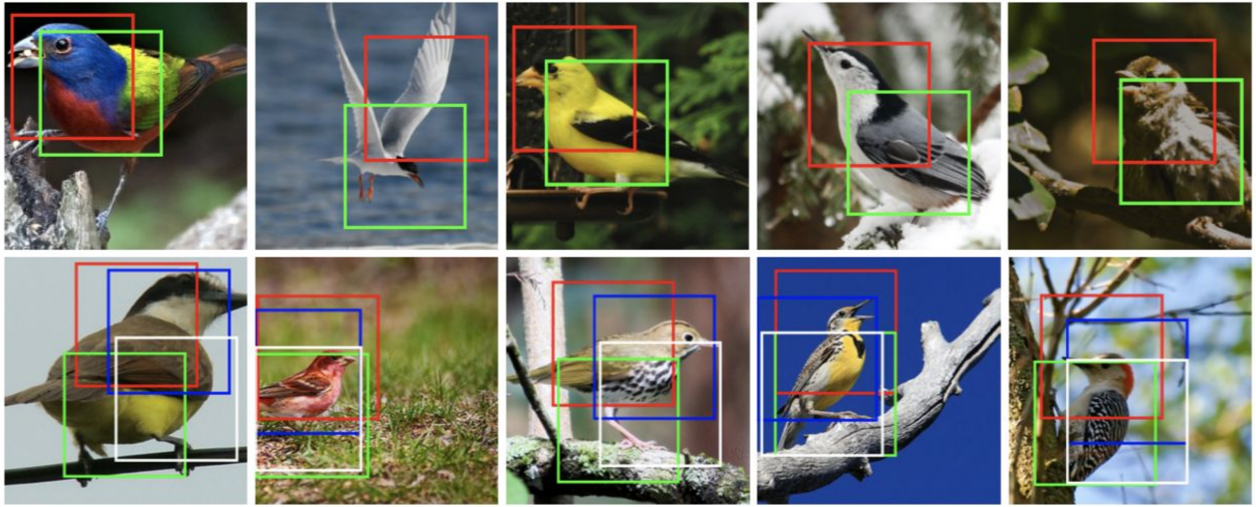
CUB-200-2011 birds dataset

CNN model – an Inception architecture with batch normalisation pre-trained on ImageNet] and fine-tuned on CUB – which by itself achieves the state-of-the- art accuracy of 82.3%



Эксперименты.

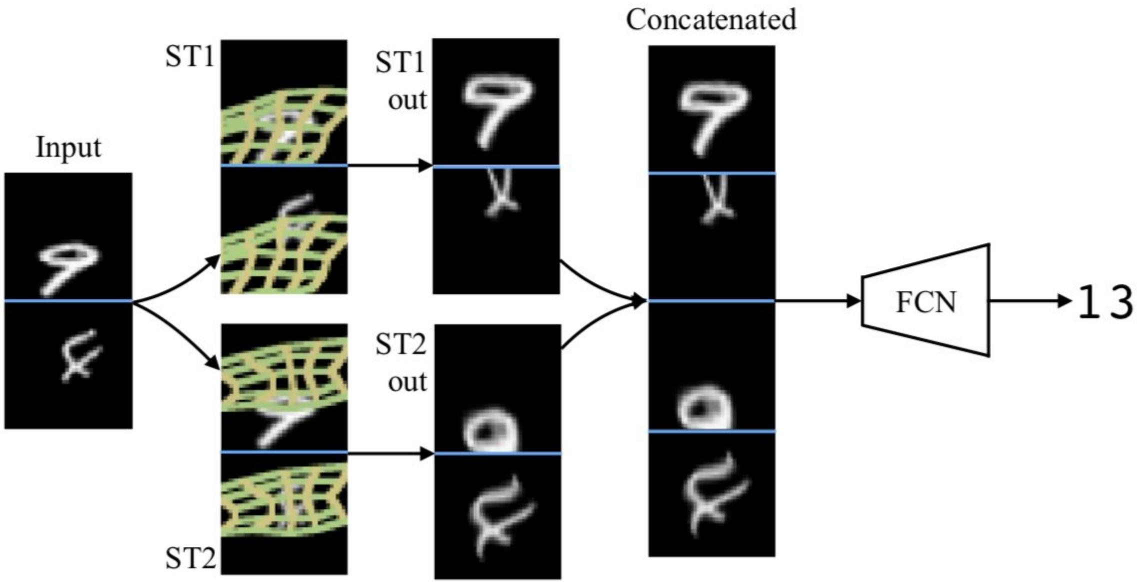
Model		
Cimpoi '15 [5]		66.7
Zhang '14 [40]		74.9
Branson '14 [3]		75.7
Lin '15 [23]		80.9
Simon '15 [30]		81.0
CNN (ours)	224px	82.3
2×ST-CNN	224px	83.1
2×ST-CNN	448px	83.9
4×ST-CNN	448px	**84.1**



Эксперименты.

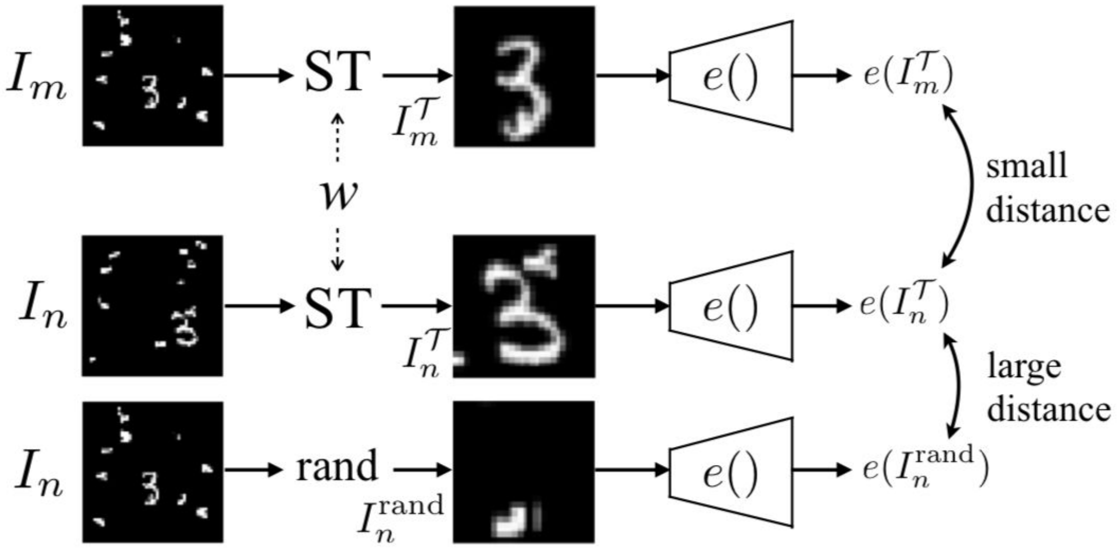
MNIST Addition

Model		RTS
FCN		47.7
CNN		14.7
	Aff	22.6
ST-FCN	Proj	18.5
	TPS	19.1
	Aff	9.0
2×ST-FCN	Proj	5.9
	TPS	5.8



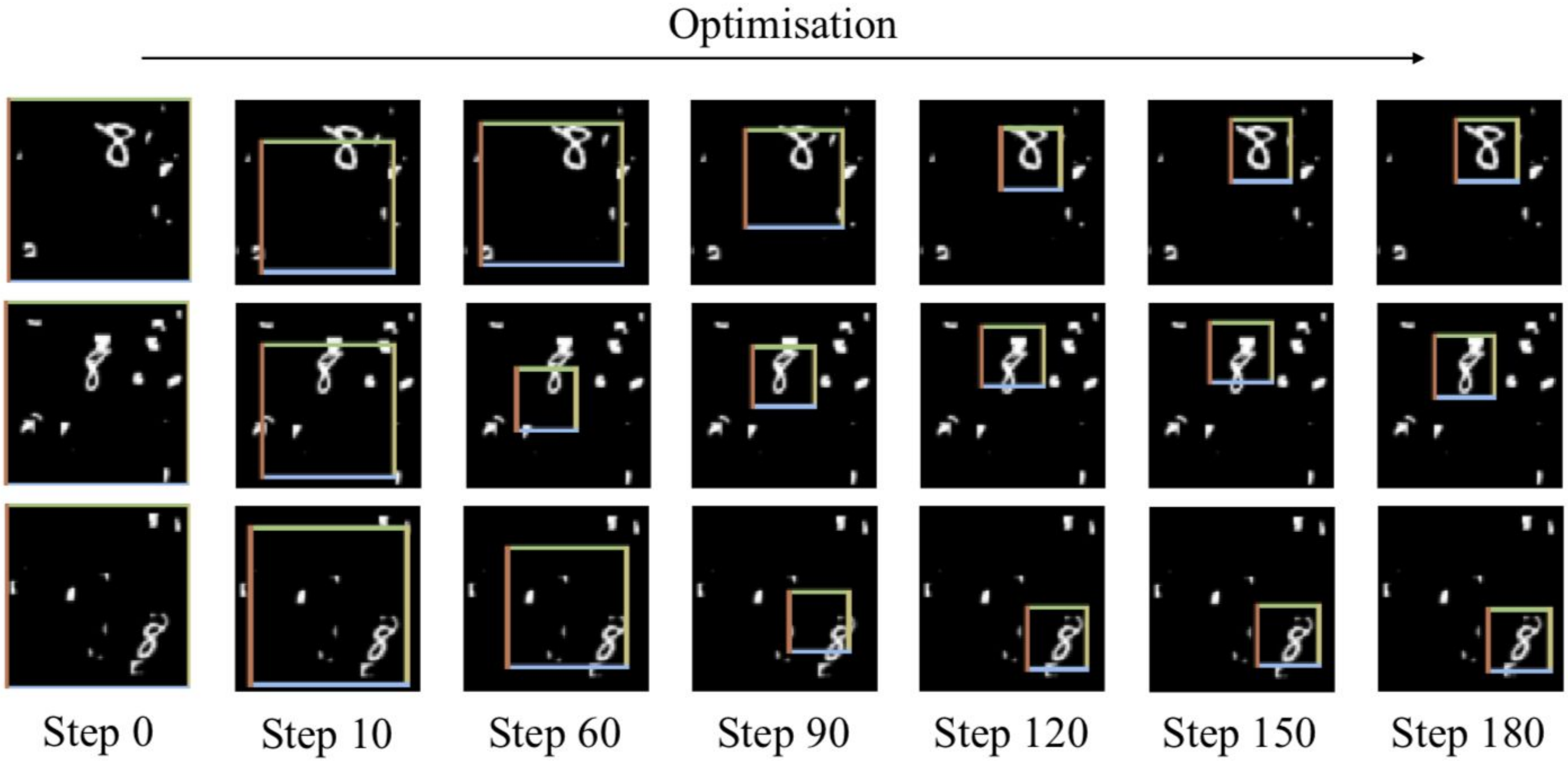
Эксперименты. Co-localisation.

Class	MNIST Distortion T	TC
0	100	81
1	100	82
2	100	88
3	100	75
4	100	94
5	100	84
6	100	93
7	100	85
8	100	89
9	100	87

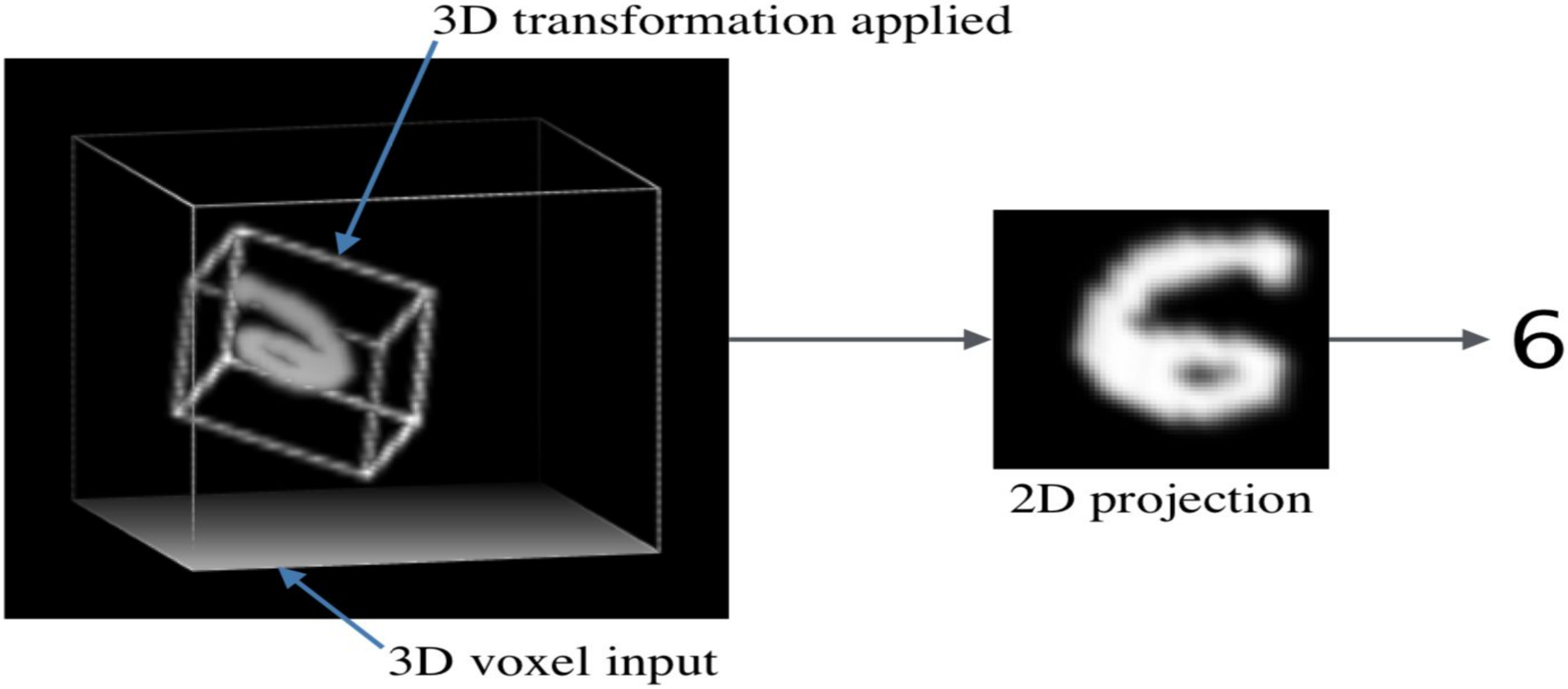


$$\sum_{n}^{N} \sum_{m \neq n}^{M} \max(0, \|e(I_n^{\mathcal{T}}) - e(I_m^{\mathcal{T}})\|_2^2 - \|e(I_n^{\mathcal{T}}) - e(I_n^{\text{rand}})\|_2^2 + \alpha)$$

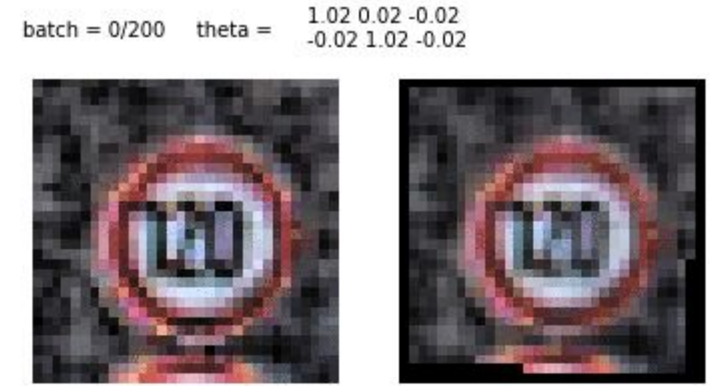
Эксперименты. Co-localisation.

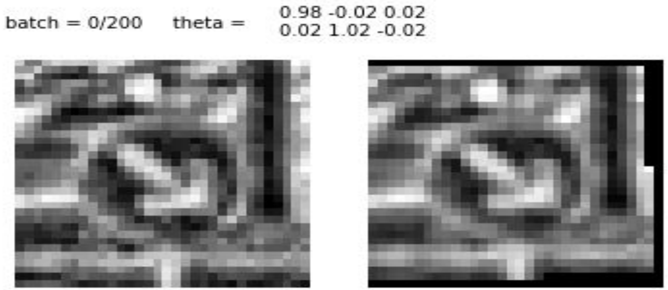


Higher Dimensional Transformers.

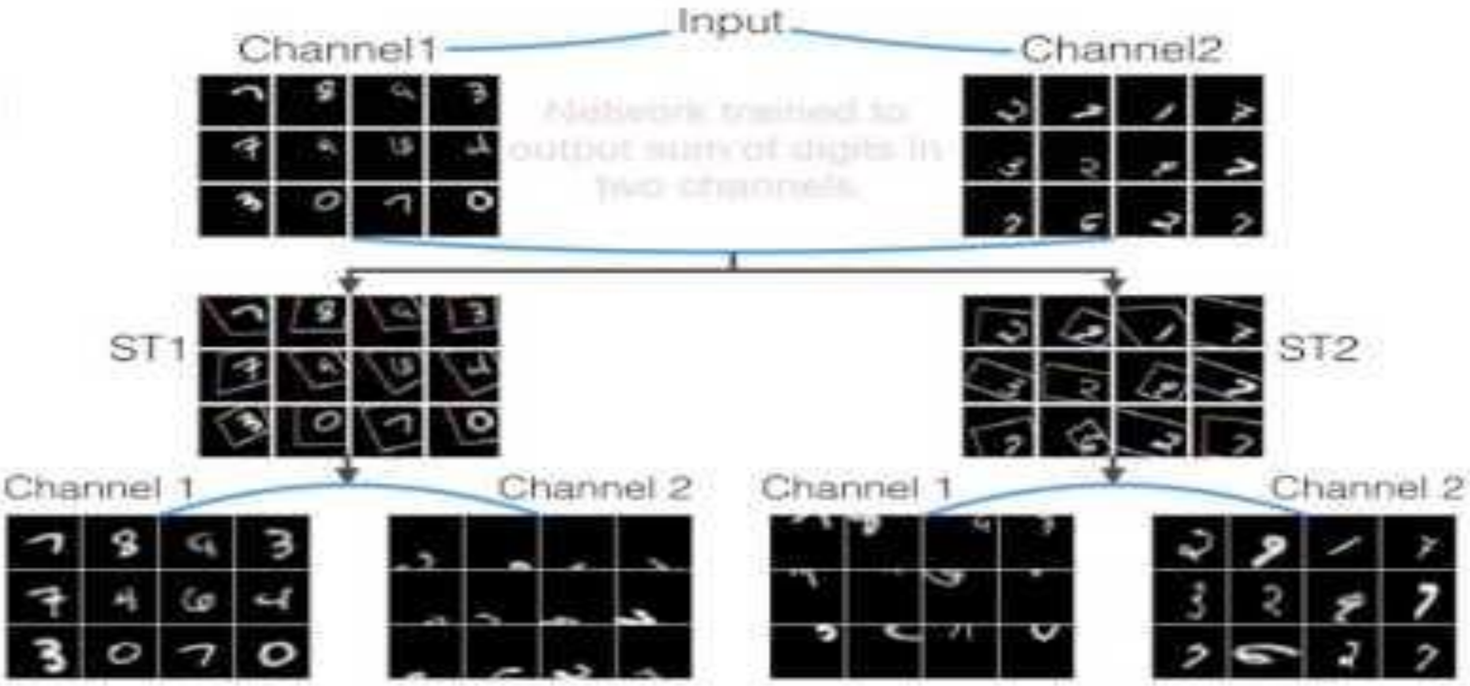


Spatial Transformer Networks with IDSIA-like classifier for German Traffic Signs Dataset classification



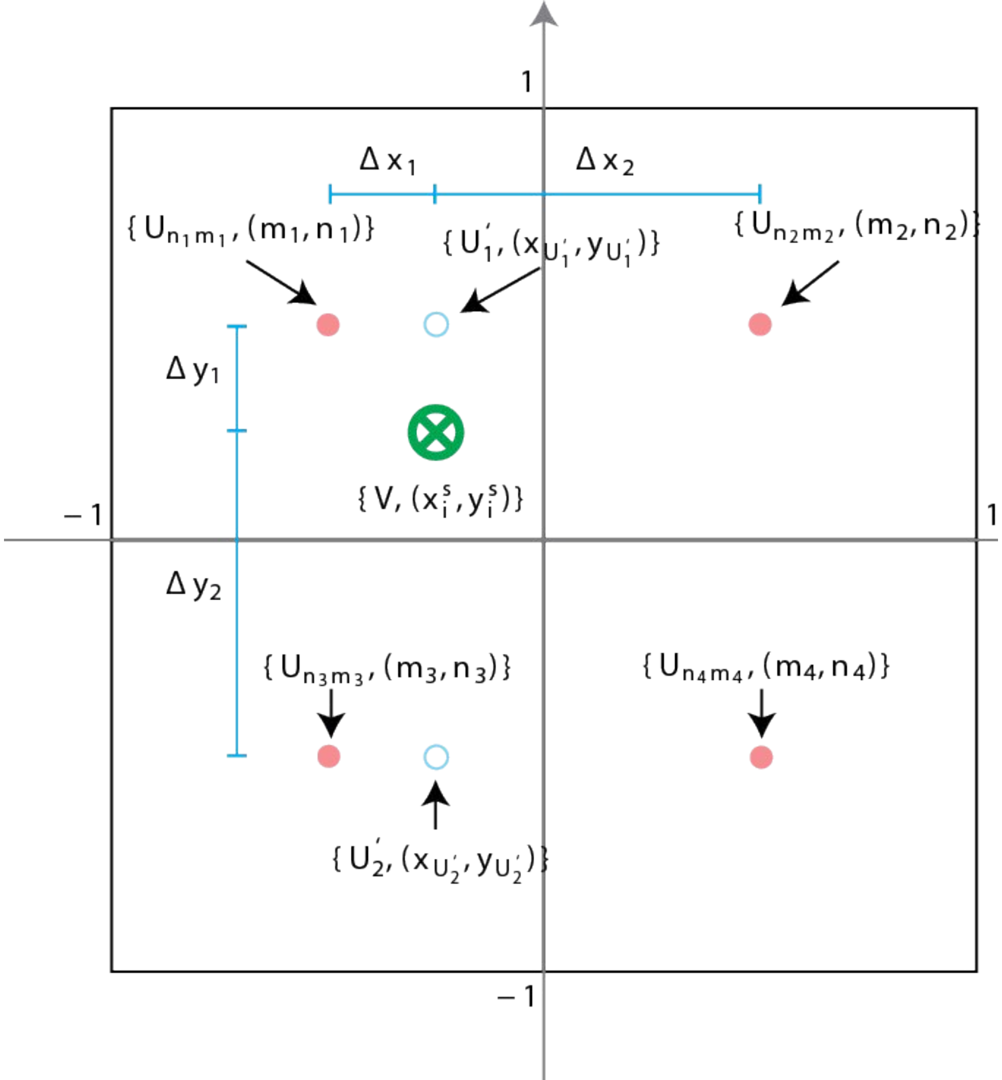


MNIST Addition



Источники.

1. https://arxiv.org/pdf/1506.02025.pdf
2. https://www.youtube.com/watch?v=Ywv0Xi2-14Y
3. https://www.youtube.com/watch?v=T5k0GnBmZVI
4. https://vision.cornell.edu/se3/wp-content/uploads/2014/09/fulltext4.pdf
5. https://cs.stackexchange.com/questions/81861/bilinear-interpolation



The procedure can be divided into three linear interpolations. First the value U_1' at position $(x_{U_1'}, y_{U_1'})$ can be computed by interpolating the values $U_{n_1m_1}$ and $U_{n_2m_2}$:

$$U_1' = \Delta x_2 \; U_{n_1m_1} + \Delta x_1 \; U_{n_2m_2}.$$

As the sum of Δx_1 and Δx_2 is equal to one, due to normalization of the axes, the above equation can be rewritten as:

$$U_1' = (1 - \Delta x_1)U_{n_1m_1} + (1 - \Delta x_2)U_{n_2m_2}.$$

The terms Δx_1 and Δx_2 can be expressed as:

$$\begin{aligned} \Delta x_1 &= |x_i^s - m_1| \\ \Delta x_2 &= |x_i^s - m_2|, \end{aligned}$$

which, substituted into the equation for U_1' yields:

$$U_1' = U_{n_1m_1}(1 - |x_i^s - m_1|) + U_{n_2m_2}(1 - |x_i^s - m_2|).$$

Similarly the value for U_2' can be computed:

$$U_2' = U_{n_3m_3}(1 - |x_i^s - m_3|) + U_{n_4m_4}(1 - |x_i^s - m_4|).$$

Once U_1' and U_2' have been computed, V can be determined by linearly interpolating U_1' and U_2':

$$V = U_1'(1 - \Delta y_1) + U_2'(1 - \Delta y_2).$$

The values for Δy_1 and Δy_2 can be expressed as follows:

$$\begin{aligned} \Delta y_1 &= |y_i^s - y_{U_1'}| = |y_i^s - n_1| = |y_i^s - n_2| \\ \Delta y_2 &= |y_i^s - y_{U_2'}| = |y_i^s - n_3| = |y_i^s - n_4|. \end{aligned}$$

Substituting the above equations and those of Δx_1 and Δx_2 into the equation for V yields:

$$\begin{aligned} V = & \; U_{n_1m_1} \cdot (1 - |x_i^s - m_1|) \cdot (1 - |y_i^s - n_1|) \\ & + U_{n_2m_2} \cdot (1 - |x_i^s - m_2|) \cdot (1 - |y_i^s - n_2|) \\ & + U_{n_3m_3} \cdot (1 - |x_i^s - m_3|) \cdot (1 - |y_i^s - n_3|) \\ & + U_{n_4m_4} \cdot (1 - |x_i^s - m_4|) \cdot (1 - |y_i^s - n_4|), \end{aligned}$$

which can be written more compactly as:

$$\begin{aligned} V &= \sum_{k=1}^{4} U_{n_km_k} \cdot (1 - |x_i^s - m_k|) \cdot (1 - |y_i^s - n_k|) \\ &= \sum_{n}^{H} \sum_{m}^{W} U_{nm} \cdot (1 - |x_i^s - m|) \cdot (1 - |y_i^s - n|). \end{aligned}$$