Understanding when Spatial Transformer Networks do not support invariance, and what to do about it

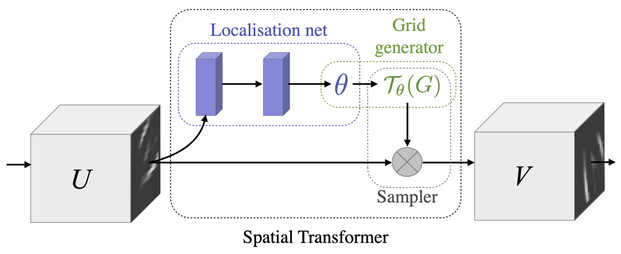
Finnveden Lukas, Jansson Ylva (yjansson@kth.se), Tony Lindeberg
KTH Royal Institute of Technology, Stockholm, Sweden.

Contributions

- We present a simple proof that STNs *do not enable invariant recognition* when transforming CNN feature maps as opposed to when transforming input images.
- We investigate alternative options for using *deeper features* to predict transformations parameters and compare those with STNs that transform CNN feature maps.
- We propose the use of *parameter sharing* between the classification network and the localisation network to enable stable training of deeper localization networks.
- We show that using deeper features is complementary to iterative methods for image alignment.

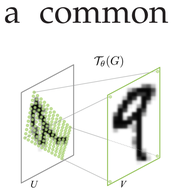
STNs and invariance

Spatial transformer networks (STNs) were introduced as an option for CNNs to *learn invariance* to image transformations.



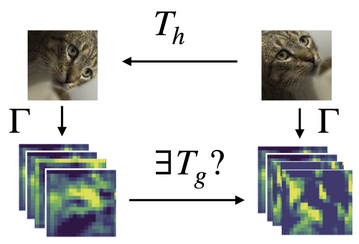
When *transforming the input*, an STN can support invariance by applying the inverse transformation to transform objects to a common pose:

$$(\Gamma \mathcal{T}_h^{-1} \mathcal{T}_h f)_c(x) = (\Gamma f)_c(x)$$



In the original paper, it was also proposed to place the ST *in the middle of a CNN* to enable the use of *deeper representations* when estimating transformation parameters. We, here, ask whether there really *exists* a transformation $\mathcal{T}_g$ dependent on $\mathcal{T}_h$ such that

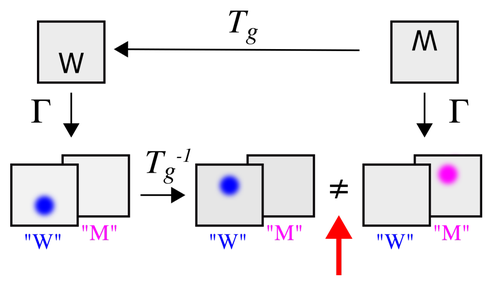
$$(\mathcal{T}_g \Gamma \mathcal{T}_h f)_c(x) = (\Gamma \mathcal{T}_h f)_c(T_g^{-1} x) \stackrel{?}{=} (\Gamma f)_c(x),$$



i.e. can an STN still support invariance when transforming *CNN feature maps*?

Intuition behind proof

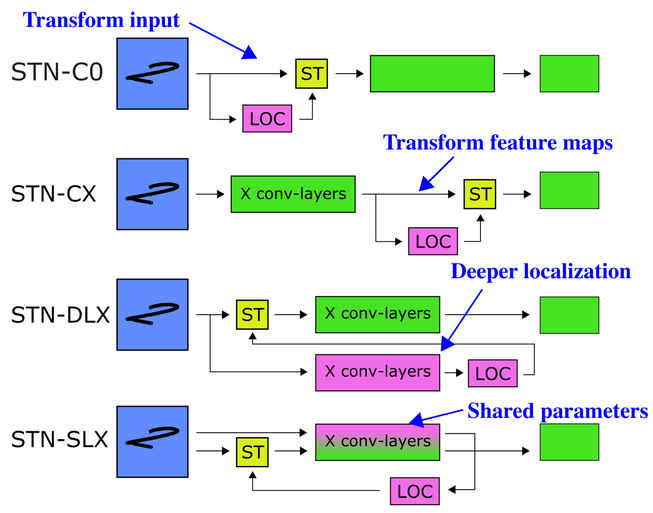
Rotation The network Γ has two feature channels "W" and "M". T_g corresponds to a rotation of 180 degrees. Since *different feature channels* respond to the transformed image as compared to the original image, it is not possible to align the feature maps by any purely spatial transformation.



Scaling When applied to a rescaled image, a filter trained to recognise an object of a smaller scale *never covers the full object of interest*. This means that the feature maps for the original and the rescaled image will *not include the same set of values* whereby alignment is clearly not possible.

STN architectures

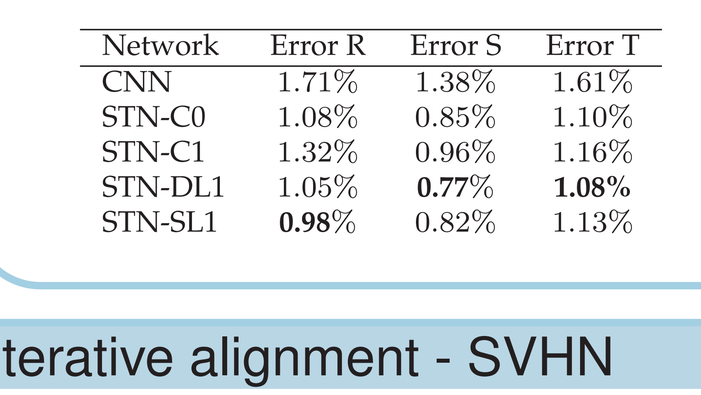
We evaluate different STN architectures to better understand the advantages of using deeper features and the consequences of transforming feature maps vs input images.



STN-CX transforms CNN feature maps, which prevents proper invariance. STN-SLX similarly to STN-DLX has a deeper localization network, but *shares parameters* between the classification and localization networks. This enables more stable training.

Qualitative results - MNIST

STN-SL1 finds a canonical pose for both rotated and scaled images. STN-C1 fails to compensate for rotations and scalings. This is because a spatial rotation/scaling is not enough to align the feature maps of a transformed image with those of the original.



Performance - MNIST

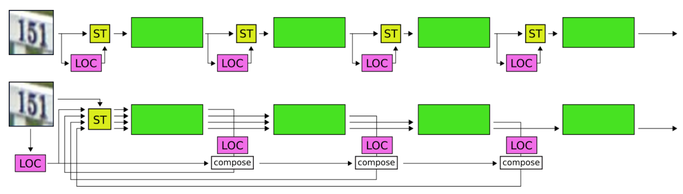
For rotations (R) and scalings (S), STN-C1 has better classification performance than a standard CNN but worse than all networks that transform the input.

For *translations (T)*, pose alignment of feature maps is possible for all network versions (because of the translation covariance of CNNs). In accordance with theory, the differences in performance are here much smaller.

Network	Error R	Error S	Error T
CNN	1.71%	1.38%	1.61%
STN-C0	1.08%	0.85%	1.10%
STN-C1	1.32%	0.96%	1.16%
STN-DL1	1.05%	**0.77%**	**1.08%**
STN-SL1	**0.98%**	0.82%	1.13%

Iterative alignment - SVHN

On the Street View House Numbers (SVHN) dataset our iterative STN with parameter sharing



improves results compared to iteratively transforming *feature maps* as proposed by Jaderberg et. al.

Network	Error (ours)	Jaderberg et. al
CNN	3.88%	4.0%
STN-C0-large	3.69%	3.7%
STN-C0123	3.61%	3.6%
STN-SL0123	**3.49%**	-

Deeper features - SVHN

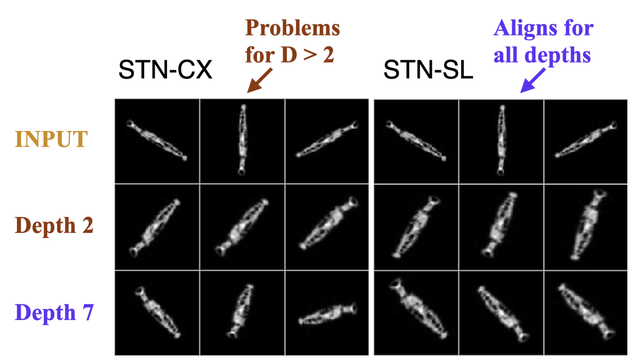
When *transforming feature maps*, using deeper features does not give any large improvement and can decrease performance. A *deeper localisation network* also does not give clear improvements with depth. When transforming the input and *sharing parameters*, there is a clear advantage of using deeper features.

Depth	STN-CX	STN-DLX	STN-SLX
$X = 0$	3.81%	-	-
$X = 3$	3.70%	3.48%	3.54%
$X = 6$	3.91%	3.75%	3.29%
$X = 8$	4.00%	3.76%	**3.26%**

When combined with *iterative alignment*, two of the parameter-shared networks are further improved (STN-SL3: 3.33%, STN-SL6: 3.18%) while the other networks achieve similar or worse accuracy.

Plankton dataset

On the Plankton dataset, when transforming feature maps at depth 2, the localisation network *does* rotate them as if trying to achieve invariance. However, when transforming the feature maps at depth 7, the localisation network entirely stops compensating.



Since invariance cannot be achieved at either depth, transforming the depth-2 feature maps still results in a higher classification error (22.3%) than a similar network that transforms the input (21.5%).

Summary and conclusions

A spatial transformation of a CNN feature map can, in the general case, *not align the feature maps* of a transformed image with those of its original.

This implies that STNs that transform feature maps *do not enable invariant recognition*. This inability is clearly visible in practice and *negatively impacts classification performance*.

We advocate *using deeper features* to estimate the transformation in STNs while still *transforming the input image*.

When training especially deep localization networks, *sharing parameters* between the classification and the localisation network increases stability.

References

T. S. Cohen, M. Geiger, and M. Weiler, "A general theory of equivariant CNNs on homogeneous spaces," in Advances in Neural Information Processing Systems, 2019, pp. 9142–9153.

R. K. Cowen, S. Sponaugle, K. Robinson, J. Luo, "PlanktonSet 1.0: Plankton imagery data collected from ..." [Online]. Available: https://accession.nodc.noaa.gov/0127422, 2015

M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 2017–2025.

Y. Jansson, M. Maydanskiy, L. Finnveden, and T. Lindeberg, "Inability of spatial transformations of CNN feature maps to support invariant recognition," arXiv preprint arXiv:2004.14716, 2020.

C.-H. Lin and S. Lucey, "Inverse compositional spatial transformer networks," in CVPR, 2017, doi:10.1109/CVPR.2017.242, pp. 2568–2576.

Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist/

Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011.