Neural encoding with affine feature response transforms

Umut Güçlü, Thirza Dado
Radboud University, Donders Institute for Brain, Cognition and Behaviour
u.guclu@donders.ru.nl, thirza.dado.1@donders.ru.nl

***Abstract*— We introduce affine feature response transforms (AFRT; \ ˈe-fərt \) a new family of neural encoding models based on spatial transformer networks (STNs). AFRT drastically simplifies the state-of-the-art neural encoding models by factorising receptive fields into a sequential affine component with five out-of-the-box interpretable parameters and response components with a small number of feature weights per voxel.**

I. Introduction

Despite their success, development of neural encoding models with deep convolutional neural networks faces a severe limitation due to the high dimensionality of the model parameters when fitting a single voxel response. That is, each voxel is modeled as a function of all possible spatial locations of the stimulus. This would not only be problematic because of the extremely high number of computations needed, but also because the problem becomes very ill-posed which can lead to overfitting. To overcome this, we have to resort to very strong regularization methods which, when coupled with the high dimensionality, makes the estimated models very difficult to interpret. Alternatively, spatial transformer networks (STNs) do differentiable affine transformations on images which can be parametrized with only five out-of-the-box interpretable values (XY-translation/scale, planar rotation) [4]. Here, we present affine feature response transforms (AFRT; \ ˈe-fərt \) which replace convolutions with spatial transformations in neural encoding models to overcome the aforementioned problems.

II. Methods

We reanalyzed part of the dataset that was originally published by [3], consisting of fMRI measurements of the visual area V1 during the presentation of visual stimuli from ImageNet (one subject, training set = 1200, test set = 50). The AFRT model consists of three main components: (i) *n* affinity layers for *n* voxel responses that transform each input image to match the effective receptive field (RF) shape of each voxel, (ii) the AlexNet model for object recognition for high-level feature extraction, and (iii) a dense layer that transforms the extracted RF feature representations into voxel responses. As a baseline, we used an encoding model comprising AlexNet for stimulus-feature transformation and ridge regression for feature-response transform [2]. In other words, while the baseline model extracts convolutional feature maps and fits a linear model thereon to predict voxel responses, AFRT affine transforms the inputs into the effective RF size of the feature model and fits a dense layer on individual feature activations.

III. Results

We quantified the performance of the models as the Pearson product-moment correlation coefficient between observed and predicted responses. The AFRT model performed equally well as the baseline model. Table 1 shows the breakdown of model performance when different AlexNet layers were used as extracted features.

	correlation		**percentage**	
	AFRT	Baseline	AFRT	Baseline
L1	0.19	0.20	0.49	0.32
L2	0.17	0.15	0.21	0.06
L3	0.13	0.13	0.11	0.12
L4	0.12	0.14	0.10	0.08
L5	0.12	0.15	0.09	0.10
L6	n/a	0.15	n/a	0.09
L7	n/a	0.14	n/a	0.11
L8	n/a	0.14	n/a	0.11

IV. Conclusion

The AFRT model performed equally well as the baseline model with fewer parameters, making it more efficient and potentially interpretable. Future work will further investigate the interpretability of the learned affine parameters (as a proxy for pRF estimation) and mixing weights of the features (as a proxy for pRF visualization) as well as integrating spatial transformers into neural decoding models such as [1, 5].

References

[1] Thirza Dado, Yağmur Güclütürk, Luca Ambrogioni, Gabrielle Ras, Sander Erik Bosch, Marcel van Gerven, and Umut Güçlü. Hyperrealistic neural decoding: Linear reconstruction of face stimuli from fmri measurements via the gan latent space. *bioRxiv*, 2020.

[2] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[3] Tomoyasu Horikawa and Yukiyasu Kamitani. Generic decoding of seen and imagined objects using hierarchical visual features. *Nature communications*, 8(1):1–15, 2017.

[4] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Ko-ray Kavukcuoglu. Spatial transformer networks. *arXiv preprint arXiv*: 1506.02025, 2015.

[5] Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güclütürk, Marcelvan Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video reconstruction from brain activity. *bioRxiv*, 2021