Deep Volumetric Feature Encoding for Biomedical Images

Brian Avants[1,2(✉)], Elliot Greenblatt[2], Jacob Hesterman[2], and Nicholas Tustison[1]

[1] Department of Radiology and Medical Imaging, University of Virginia, Charlottesville, VA, USA
stnava@gmail.com
[2] Invicro, LLC, Boston, MA, USA

Abstract. Deep learning research has demonstrated the effectiveness of using pre-trained networks as feature encoders. The large majority of these networks are trained on 2D datasets with millions of samples and diverse classes of information. We demonstrate and evaluate approaches to transferring deep 2D feature spaces to 3D in order to take advantage of these and related resources in the biomedical domain. First, we show how VGG-19 activations can be mapped to a 3D variant of the network (VGG-19-3D). Second, using varied medical decathlon data, we provide a technique for training 3D networks to predict the encodings induced by 3D VGG-19. Lastly, we compare five different 3D networks (one of which is trained only on 3D MRI and another of which is not trained at all) across layers and patch sizes in terms of their ability to identify hippocampal landmark points in 3D MRI data that was not included in their training. We make observations about the performance, recommend different networks and layers and make them publicly available for further evaluation.

Keywords: Code: 3D VGG-19 · Landmarks · Key-point detection · Deep features

1 Introduction

Feature detection for pattern matching in images has a long history in computer vision, dating at least to the 1950s [11]. Perhaps the most well-known of these approaches is the scale invariant feature transform (SIFT) [5] which, as is typical of many of these methods, uses engineered features to localize salient features in images. These key-points are then filtered and matched in order to compute a geometric correspondence between image sets with little computational overhead. As such, SIFT is widely adopted as a core tool in industrial applications of computer vision.

While SIFT and related methods are powerful, their extension to 3D biomedical imaging has not, as yet, met with the same level of adoption and success.

Ž. Špiclin et al. (Eds.): WBIR 2020, LNCS 12120, pp. 91–100, 2020.
https://doi.org/10.1007/978-3-030-50120-4_9

Rister, et al. [10] extended SIFT to 3D but found that, although it performed well within-subject, it did not reach usability in inter-subject registration. This finding suggests that more general approaches – or different feature sets – may be of value.

New approaches to feature matching and registration of biomedical imaging data are also needed to handle its ever increasing diversity and magnitude. Furthermore, the desire to integrate imaging with other forms of data (e.g. genomics) leads to additional motivation to develop fast, general purpose search and matching based on biomedical (often volumetric) image features.

Feature-based matching, in this context, provides a powerful solution that, like SIFT, may be less sensitive than dense registration methods to occlusion, noise, resolution and modality. Furthermore, feature extraction methods may be more memory efficient which is of tremendous value when full datasets (e.g. CLARITY images) cannot be stored in memory without special handling [6].

Pre-trained convolutional networks, such as VGG-19 [14], have proven to be powerful feature encoders with applications in a variety of areas [4,16]. Distances between activation maps from intermediate layers of deep architectures (deep features) are effective metrics in domains beyond the original application area and have transformed practice in super-resolution, key point matching [8] and semantic segmentation. The use of these resources is relatively limited in the biomedical domain because these features are typically derived from 2D datasets.

The current paper is motivated by the desire to build a library of general purpose, pre-trained deep networks for volumetric feature encoding that may be used for transfer learning within the context of regression, classification, super-resolution and matching problems. ModelsGenesis [17] (Generic Autodidactic Models for 3D Medical Image Analysis) has similar goals to ours. However, in contrast to our focus on regression, Models Genesis leverages encoding/decoding (U-net like) architectures frequently used in segmentation tasks. Our contributions include: (1) an extension of the long-term proven 2D VGG-19 features to 3D, (2) approximation of these features with 3D regression networks and (3) comparison of the derived feature spaces to intrinsically 3D regression networks, including one that undergoes no training at all. The analysis contrasts the value of these networks' features at different layer depths – and with different input patch sizes – in terms of landmark matching in 3D MRI of the hippocampus [1].

2 Methods

We develop five different 3D networks based on established approaches. Two of these networks do not require additional training and extend 2D VGG 19 to 3D. Two others are based on regressing against VGG-19 activations. The final network is trained to solve a completely separate regression and classification problem and is treated as a fixed, intrinsically 3D pre-trained network space. Later, we compare the features generated by these networks in terms of their ability to perform a pure feature-based landmark matching problem. All of the work below is implemented with `R` [7] and `tensorflow` within `ANTsR`.

2.1 Network 1: 3D VGG-19 - No Training

There is a long line of research revolving around the use of randomly selected features in machine learning. Such features are unbiased and, at large scale and for general purpose application, provably good. Convolutional neural network architectures may encode valuable feature representations even without training [12]. While evidence of this has existed for some time, recent work has put the claims on more solid foundation [2]. Ramanujan et al., for instance, successfully "validate the unreasonable effectiveness of randomly weighted neural networks for image recognition" [9]. Following this work, we include an untrained 3D VGG-19 architecture among the networks we test. This network uses layers with randomly initialized weights for encoding. Its architecture is the same as the following two networks and includes groups of two (in shallower parts of the network) and four convolutional layers with filters of size $3 \times 3 \times 3$ followed by 3D max pooling. The number of filters increases dyadically with depth, except in the last block. See [14] for details. Our 3D variant is identical to an expanded 2D version with the exception that the input layer, for our 3D version, is single-channel and, of course, has more parameters (60,058,688) in line with its increased dimensionality.

2.2 Network 2: Transfer Learning from 2D VGG-19 to Pseudo-3D VGG-19

Both work in segmentation [13] and video has demonstrated the ability to transfer weights from multi-channel 2D convolutional filters into 3D. As in prior work, we adapt the *keras vgg19* imagenet weights from its canonical 2D implementation into a 3D single channel variant with all filter sizes, filter counts and biases the same. Two observations are key here. First, while acknowledging the limitations of this assumption, we treat the x-y spatial orientation of the original VGG-19 filters as rotatable into y-z and x-z planes where channel information occupies the orthogonal dimension. Second, this results in three variants of the network, one for each orientation. When applied to perform inference on new data, the outputs of each oriented network should be either concatenated or averaged. In the evaluation study described below, we concatenate features before use in landmark matching. See Fig. 1 for an overview of the approach to transferring 2D VGG to pseudo-3D space.

2.3 Network 3: Direct 3D VGG Learning of Pseudo-3D VGG-19 Activations

This network shares the same architecture as the prior two. Our task, here, is to train a 3D VGG-19 to approximate the sum of the oriented pseudo-3D VGG-19 outputs. By outputs, here, we mean the activations at the deepest layer of the VGG-19 architectures which is known as `conv5_4`, where 5 and 4 indicate

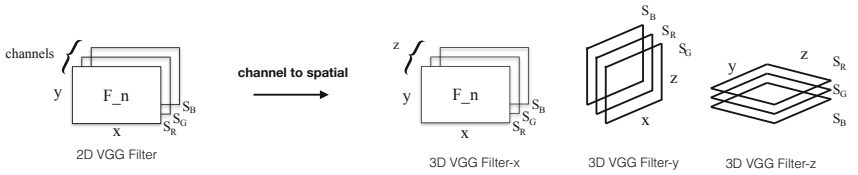

Fig. 1. Multi-channel VGG-19 (2D) to single-channel pseudo-3D VGG-19 filter transfer. The transfer operation results in 3 variants, one for each orientation, x, y, z. See the code for additional details.

its position is at the 4th convolutional layer of the 5th VGG block. The loss function, then, is:

$$\frac{1}{3}\sum_{i=1}^{3} \|\phi_{54}(X) - \phi_{54}^{i}(X)\|^2$$

where ϕ is the 3D-VGG-19 network, ϕ^i is the oriented pseudo-3D VGG-19 network, X denotes a tensor input and $\|\cdot\|$ is the Frobenius norm. The output layer is denoted by the *54* subscript.

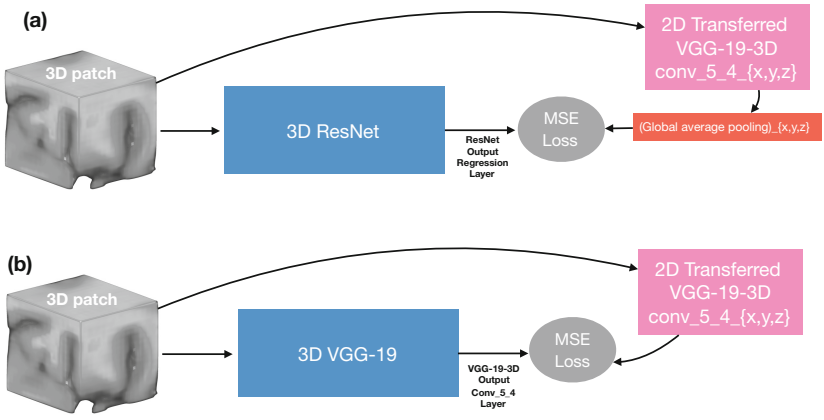

Fig. 2. Two approaches to learning the 3D distance space induced by the 2D to 3D VGG19. (a) 3D Resnet learning of 2D-to-3D VGG-19 activation maps. The loss function averages over the oriented outputs from the 2D to 3D transferred features. (b) 3D VGG-19 learning of 2D-to-3D VGG-19 activation maps. The loss function averages over the oriented outputs.

We train this network using a V100 NVIDIA GPU on tasks 1 through 10 of the medical decathlon dataset [15] which includes a variety of 3D CT and MRI images from the brain, heart, liver, prostate, lung, pancreas, spleen and colon. We use the `tensorflow` ADAM optimizer with learning rate 1e−4. Patch sizes of 32^3 are extracted from datasets that permit this dimensionality. Otherwise,

patches of size 16^3 are used. Each patch is scaled to $[-127.5, 127.5]$. Batch sizes of 32 were employed. We trained on 64,000 patches and validated on patches extracted from left out images. Note that validation was used to guide the point at which we extracted the best weights from the training history. At convergence (67 epochs), the overall correlation (in validation data) between the real and predicted activation maps reached 0.792 with a training error reduction (from initialization) of a factor of 2.2. See Fig. 2 for an overview of this training paradigm.

2.4 Network 5: Direct 3D ResNet Learning of Pseudo-3D VGG-19 Activations

This comparison network is similar to the prior one but, here, we employ our 3D variant of the ResNet architecture [3]. ResNet is a classification or regression network and its output dimensions do not match that of 3D VGG19. To overcome this barrier – and still allow ResNet to predict pseudo-3D VGG encodings – we add a global average pooling layer to the output of each ϕ^i_{54}. This leads to a 512 vector regression target for each oriented network. The ResNet can directly learn this encoding using the loss function:

$$\frac{1}{3}\sum_{i=1}^{3} \|\psi(X) - \phi^i_{54g}(X)\|^2$$

where ψ is the ResNet and ϕ^i_{54g} is the ϕ^i_{54} output followed by global average pooling. The norm is Euclidean.

We train this ResNet (25,851,112 parameters) in the same manner as the prior network. We use the same patches, optimizer parameters and convergence criterion. At convergence (28 epochs), the overall correlation (in validation data) between the real and predicted activation maps reached 0.881 with a training error reduction (from initialization) of a factor of 1.7.

2.5 Network 5: Pre-trained ResNet Network

Our last comparison network is a ResNet with 25,851,112 parameters that predicts age, gender and data collection site based on T1-weighted neuroimaging (also known as `brainAge`). This network was trained on a dataset of control subjects where each image is bias corrected and affinely registered to a template image. Training data include:

- Dallas Lifespan Brain Study (DLBS): n = 275 (lifespan);
- Human Connectome Project (HCP): n = 1245 (young control);
- Information eXtraction from Images (IXI): n = 563 (lifespan);
- Nathan Kline Institute Rockland (NKI): n = 1260 (lifespan);
- Open Access Series of Imaging Studies (Oasis-2): n = 433 (lifespan, 18–93);
- Southwest University Adult Lifespan Dataset (SALD): n = 494 (young control).

As this network is not a primary topic of this work, we leave further details of training and performance to its online documentation (see `ANTsR` brain age network documentation and application). Nevertheless, the network achieves – in completely independent validation data from sites not included in training – an age prediction absolute error of 3.4 years over the lifespan and 88% accuracy for gender classification. This suggests that the network is encoding "real" information about shape and structure in human neuroimages and, as such, is sufficient to use as the source of deep features.

2.6 Evaluation Strategy in Terms of Landmark Matching in 7T Hippocampus Data

We obtained public 7T T2-weighted MRI of the human hippocampus ($n = 34$) [1] as a resource for anatomical labels and point-wise landmarks. Each image was labeled by a manual rater with two anatomically identified points at the head and tail of the hippocampus. We selected these two points for their relative saliency within the structure of these images. One subject (`001`) was arbitrarily selected as the template image. All other subjects serve as testing data where the task is for the underlying matching algorithm to use deep features to identify the anatomically homologous landmark points in the target image.

Figure 3 shows example activation maps from this network demonstrating that the ResNet layers capture shape variation associated with the example input patch. Such activation maps are the source of the feature distances that will drive the automated landmark matching. Similar maps are generated by each of the candidate networks.

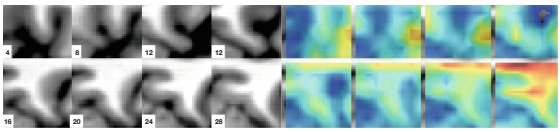

Fig. 3. 3D Activation maps for an example patch input for the brain-age network. The patch is shown at left and activation is at right. Slices within the 3D patch are indexed in the lower left of each image panel.

Each of the networks under study has five-dimensional weights per convolutional layer. The first three dimensions are spatial. The fourth is the channel dimension. The fifth is the number of filters. In general, the number of filters increases with depth. The VGG-19 architectures start with 64 filters and end with 512. The ResNet architectures range from 64 filters to 2048 at the deepest level. We select shallow, mid-range and deep layers for comparison of performance on automated landmark identification. For VGG, we select `conv2_2`, `conv4_2` and `conv5_4` layers. For ResNet architectures, we select layers 6, 140 and 1290 as the shallow, mid and deep feature layers to evaluate. For each of

these variants, we also explore patch sizes of 12, 16, 20, 24, 28 and 32 voxels per patch axis. In total, this results in 90 different performance comparisons on the landmark matching task.

The evaluation metric is the mean Euclidean distance between the target ground truth landmark locations (in physical space) and the estimated landmark position. An overview of the paradigm is in Fig. 4. The features, for each run of the matching algorithm, are constant for the template image and are simply the deep features that arise from the patch centered at the landmark position. The best match in the target image is then identified by taking the voxel position in the target image whose feature map is closest under either the Frobenius or Euclidean norm, depending on the network. This amounts to a landmark matching process that evaluates the deep feature space as a strict similarity metric without further regularization.

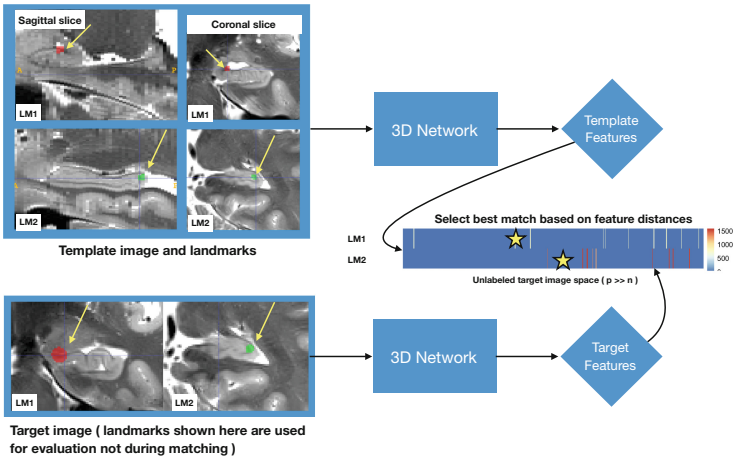

Fig. 4. Evaluation strategy for landmark localization accuracy based on deep feature matching with five different networks. In brief, the deep feature spaces are employed as similarity metrics. The matching is greedy and unconstrained. As such, this serves as a pure test of deep features' ability to match anatomy based on feature similarity.

3 Evaluation Results

Figure 5 provides an overview of results that visualizes the outcome of all 90 comparisons. Results are reported in the form of the t-statistic resulting from a pairwise t-test between the initial landmark distances and the final landmark distances after matching. We first provide general observations and then focus on the best performers.

In general, patch sizes significantly impact performance. The `ResNet` and `brainAge` networks, in particular, benefit from increasing patch sizes in the

deeper and mid-range layers, respectively. This is verified by regressing patch size against the improvement in landmark distance and – despite a small sample of only 6 patch values against which to regress – p-values < 0.005 emerge. Conversely, the `no Train`, `3D VGG` and `pseudo 3D VGG` networks show the opposite effect at the deep layers: decreasing patch size significantly improves performance. However, at shallow layers, the VGG architectures perform better with larger patch sizes.

The best networks succeed, with some configurations, at performing substantially better than chance with the lowest p-value being well below an aggressive Bonferroni correction level of $0.05/90 = 0.00056$ where we correct for all 90 test comparisons (a t-statistic of 5 with 33 degrees of freedom results in a p-value of 1.772e−05). The best result is gained by the brainAge ResNet with results of patch size 32 and the middle and deep layer being nearly equivalent with t-statistics of 5.86 and 5.85 respectively. The second best result is gained from the brain age network shallow layer with patch size 20. Interestingly, the third best result is gained by the `no training` 3D VGG19 network with a t-statistic of 5.01 at the mid-layer and with patch size 32.

Fig. 5. Landmark localization accuracy results relative to initial distances (pairwise t-test, 33 degrees of freedom). The heatmap and entries in each panel correspond to the t-statistic from the pairwise test.

4 Discussion

This effort evaluated several deep 3D networks as feature encoders and their use and evaluation in landmark matching. These networks are publicly available at https://figshare.com/articles/pretrained_networks_for_deep_learning_applications/7246985 and may serve purposes beyond those exhibited here. Potential applications include disease classification, dimensionality reduction and use within loss functions for problems such as image translation or super-resolution.

Several interesting findings arose from the evaluation study. First, the performance of the `no training` 3D VGG network validates, in 3D, prior claims of the potentially good performance of 2D convolutional networks with random weights [9]. Second, pre-trained ResNet architectures are able to provide valuable 3D feature encodings for landmark matching even if they are trained on very different data and problem domains than which they are being applied. Third, complex effects of patch size are apparent in these results. These may be confounded by the way in which these networks were trained although further investigation of that question will be left to future work.

The findings in this work are insufficient to determine the extent to which network depth impacts performance in landmark matching. Additional tests across many more layers – and concomitant statistical modeling of depth × filter number effects – would be needed to understand these likely complex interactions. However, network depth (anecdotally speaking) does appear to impact performance, as has been shown previously in 2D. This impact, like that of patch size, will likely vary with network architecture and problem domain.

It is a substantial challenge to identify the optimal layers, patch sizes and training paradigms for generating repurposable deep feature networks. The number of evaluation runs is inevitably large and computationally demanding when exploring deep 3D networks. If we use the field of super-resolution as an example (see [16]), we must rely on the community to employ these networks in creative ways and arrive at consensus about their usefulness. Until we have larger 3D datasets, we may not achieve the generality of VGG-19. More work is also needed to establish a general similarity metric based on deep features.

In conclusion, we are releasing this work as public domain investigation into the questions posed here that are at the interface of deep learning, image registration and biomedical applications. We must also acknowledge that this work must continue with more sophisticated matching strategies that go beyond the greedy method used here. Furthermore, we hope that more detailed, landmark-based evaluations will be performed in the future. We believe that such studies may (relative to evaluations that fixate on segmentation overlap) provide greater insight and specificity during the evaluation of similarity metrics and transformation models for medical registration problems.

References

1. Berron, D., et al.: A protocol for manual segmentation of medial temporal lobe subregions in 7 Tesla MRI. NeuroImage Clin. (2017). https://doi.org/10.1016/j.nicl.2017.05.022
2. Gaier, A., Ha, D.: Weight Agnostic Neural Networks, June 2019. http://arxiv.org/abs/1906.04358
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.90

4. Kim, J., Lee, J.K., Lee, K.M.: Accurate image super-resolution using very deep convolutional networks. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2016). https://doi.org/10.1109/CVPR.2016.182
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
6. Mang, A., Gholami, A., Biros, G.: Distributed-memory large deformation diffeomorphic 3D image registration. In: International Conference for High Performance Computing, Networking, Storage and Analysis, SC (2016). https://doi.org/10.1109/SC.2016.71
7. Muschelli, J., et al.: Neuroconductor: an R platform for medical imaging analysis. Biostatistics **20**(2), 218–239 (2018). https://doi.org/10.1093/biostatistics/kxx068. http://dx.doi.org/10.1093/biostatistics/kxx068
8. Neubert, T., Makrushin, A., Hildebrandt, M., Kraetzer, C., Dittmann, J.: Extended StirTrace benchmarking of biometric and forensic qualities of morphed face images. In: IET Biometrics (2018). https://doi.org/10.1049/iet-bmt.2017.0147
9. Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., Rastegari, M.: What's Hidden in a Randomly Weighted Neural Network? November 2019. http://arxiv.org/abs/1911.13299
10. Rister, B., Horowitz, M.A., Rubin, D.L.: Volumetric image registration from invariant keypoints. IEEE Trans. Image Process. **26**(10), 4900–4910 (2017). https://doi.org/10.1109/TIP.2017.2722689
11. Rosenfeld, A.: Picture processing by computer. ACM Comput. Surv. (CSUR) (1969). https://doi.org/10.1145/356551.356554
12. Saxe, A.M., Koh, P.W., Chen, Z., Bhand, M., Suresh, B., Ng, A.Y.: On random weights and unsupervised feature learning. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011 (2011)
13. Shan, H., et al.: 3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network. IEEE Trans. Med. Imag. (2018). https://doi.org/10.1109/TMI.2018.2832217
14. Simonyan, K., Zisserman, A.: VGG-16. arXiv preprint (2014). https://doi.org/10.1016/j.infsof.2008.09.005
15. Simpson, A.L., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms, February 2019. http://arxiv.org/abs/1902.09063
16. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/CVPR.2018.00068
17. Zhou, Z., et al.: Models genesis: generic autodidactic models for 3D medical image analysis. In: Shen, D., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 384–393. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_42