Collaborative slide screening for the diagnosis of breast cancer metastases in lymph nodes

Gianluca Gerard GIANLUCA.GERARD01@UNIVERSITADIPAVIA.IT
Marco Piastra MARCO.PIASTRA@UNIPV.IT
Laboratorio di Visione Artificiale e Multimedia, Dipartimento di Ingegneria Industriale e dell'Informazione
Università degli Studi di Pavia, Via Ferrata 5, 27100 Pavia, ITALY

Editors: Under Review for MIDL 2019

Abstract

In this paper we assess the viability of applying a few-shot algorithm to the segmentation of Whole Slide Images (WSI) for human histopathology. Our ultimate goal is to design a deep network that could screen large sets of WSIs of sentinel lymph-nodes by segmenting out areas with possible lesions. Such network should also be able to modify its behavior from a limited set of examples, so that a pathologist could tune its output to specific diagnostic pipelines and clinical practices. In contrast, 'classical' supervised techniques have found limited applicability in this respect, since their output cannot be adapted unless through extensive retraining. The novel approach to the task of segmenting biological images presented here is based on *guided networks*, which can segment a *query* image by integrating a *support* set of *sparsely* annotated images which can also be extended at run time. In this work, we compare the segmentation performances obtained with guided networks to those obtained with a Fully Convolutional Network, based on fully supervised training. Comparative experiments were conducted on the public Camelyon16 dataset; our preliminary results are encouraging and show that the network architecture proposed is competitive for the task described.

Keywords: fully convolutional network, few-shot learning, meta-learning, sparse annotation, lymph nodes, camelyon16, histopathological images

1. Introduction

Breast cancer is the most common form of cancer among women in the Western world. The prognosis depends on whether the cancer has spread to other organs. Sentinel lymph nodes are, in fact, the organs which are primarily reached by metastasizing cancer cells and therefore their diagnosis is of critical importance to decide patients treatment. In clinical practice, the preparation of diagnostic samples is conducted through a pipeline in which slices are cut from the sentinel lymph nodes, fixed on glass slides, then stained and finally digitized to obtain *Whole Slide Images* (WSIs). These WSIs are visually inspected by human pathologists to achieve the required diagnosis.

Our goal is to design an automated segmentation method that could help the pathologist in screening those WSI areas which are actually worth an accurate inspection. To do so we plan to apply a novel deep learning method that could correct and adapt its behavior based on a very limited set of examples. Deep learning, in fact, has achieved remarkable successes in the classification and segmentation of biomedical images (Ronneberger et al., 2015).

However supervised deep learning, which entails training on large annotated datasets of images, does not adapt easily to handle images acquired through different protocols, unless the dataset is properly extended and a full training is performed again (Shen et al., 2017).

These limitations could be overcome by a method allowing to correct the resulting output through a limited set of annotated images selected by the pathologist and supplied at run time. In this perspective, our objective is to achieve a method for the automatic segmentation of lymph-nodes, that could satisfy the following requirements:

- be collaborative and easy to use;
- achieving state-of-the-art accuracy;
- requiring minimal maintenance.

Such tool should be used as a support for automatic screening of WSIs in actual clinical environments.

2. Related Work

Transfer learning is currently the mainstream approach for training deep learning models on a limited set of annotated examples. Fine-tuning pre-trained initial weights avoids having to re-learn the network weights from scratch and can reduce the time to converge by orders of magnitude. However, the benefit of using pre-trained weights greatly decreases as the original task which the network was trained to solve diverges significantly from the target one (Yosinski et al., 2014). Furthermore, although the number of examples required for fine-tuning might decrease training times by 2 or 3 orders of magnitude, obtaining sensible results may still need thousands of annotated examples. In the line of principle, it would be extremely useful for practical application the possibility of adapting the network behavior, even after extensive training, by just providing a handful of additional, selected images.

Meta-learning has been proposed to acquire knowledge from a limited set of examples. Early work dates back to the late 1980s, Schmidhuber (1987), and early 1990s, Bengio et al. (1991), but it is only recently (Lake et al., 2016) that meta-learning was advocated as key to achieve human-level intelligence[1].

Current meta-learning systems are trained on a large set of classification problems, generated from large quantities of available annotated data, and are tested on their capability to perform classifications on new datasets with potentially new classes, which were unseen at training time. In our scenario, in contrast, the set of classes (e.g. lesion and tissue) is not subject to change over time, whereas segmentation errors could manifest over time and additional WSIs, acquired through different digital pipelines, may become available at subsequent times.

Another relevant idea for the purposes of this work is annotating images via *sparse annotations* (Glocker et al., 2013), as opposed to *dense annotations*. Sparse annotations have also been used successfully in segmentation of natural images by Xu et al. (2016), whose approach has many similarities to the few-shot segmentation method by Rakelly et al. (2018), although the former is limited to interactive segmentation only.

1. https://bair.berkeley.edu/blog/2017/07/18/learning-to-learn/

3. Methods

3.1. Few-shot learning

Few-shot learning is designed to achieve good generalization with few annotated examples. The current paradigm for few-shot learning revolves around the concept of *episode* introduced by Vinyals et al. (2016).

An s-shot, K-way episode is built by sampling a subset of K classes, possibly including the *unknown* or *background* class, from a set of classes C_{train} and then generating (Ren et al., 2018):

- A training (*support*) set $\mathsf{S} = \{(\mathbf{x}_1, \mathbf{L}_1), (\mathbf{x}_2, \mathbf{L}_2), \ldots, (\mathbf{x}_{s\times K}, \mathbf{L}_{s\times K})\}$ with s examples from each of the K classes;
- A test (*query*) set $\mathsf{Q} = \{(\mathbf{x}_1^*, \mathbf{L}_1^*), (\mathbf{x}_2^*, \mathbf{L}_2^*), \ldots, (\mathbf{x}_t^*, \mathbf{L}_t^*)\}$ of t different examples from any of the K classes.

In our case, $\mathbf{x}_i$ is an image patch with shape $[H, W, 3]$ and $\mathbf{L}_i$ represents the corresponding annotation. Annotations are sets of pixel-label pairs, (p, l), where p represents the coordinates of a pixel in the image and l is the corresponding label denoting which of the K classes the pixel belongs to. As a formula we have

$$\mathbf{L}_i = \{(p_j, l_j)\}_{j=1}^{P} \ , \ l \in \{1 \ldots K\}$$

where P is the number of annotated pixels in the image. *Sparse* annotations (Glocker et al., 2013) have small P whereas for *dense* annotations P is equal to the image size $H \times W$. Clearly, for the query set, annotations $\mathbf{L}_i^*$ are only used during the training phase and in our case these annotations must be *dense*. During the training phase, the support set S is fed to the classifier and the weights are updated to minimize the loss of the output prediction on the query set Q (see Figure 1). In few-shot learning, each support set contains just a few examples (i.e. s is small).

3.2. Sparse annotations, guided networks

In this work, we use few-shot learning to segment a new, limited dataset of sparsely-annotated histopathological patches. Figure 1 refers to the case in which $K = 2$, namely the two classes are *tissue* and *lesion*. In this figure the green dots represent sparse annotations for the lesion class whereas the red dots correspond to the tissue class[2]. Clearly, sparse annotations require significantly less time and effort to be completed in comparison to the tedious work of fully annotating a complete image. As shown in Figure 1, a sparse annotation containing multiple classes can be decomposed into multiple single-class annotations.

For few-shot segmentation, we use the *guided* network model introduced by Rakelly et al. (2018), which is shown in Figure 2. The network architecture is organized in two branches:

2. Dense annotations can be seen in background as a reference but are not used in the "Task Representation" branch.

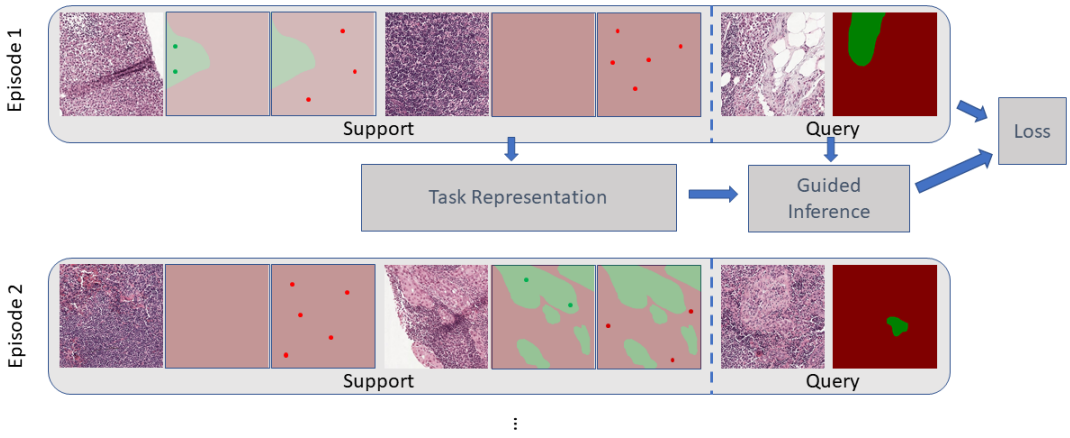

Figure 1: This figure represents episodic learning. During the first episode the network learns from the 2-shot 2-way support set to segment the query image; the prediction thus obtained is compared against the actual (dense) segmentation to compute the loss and back-propagate the errors for end-to-end learning. The training then moves on to the next episode until a prefixed number of episodes has been considered or some other convergence criteria have been satisfied.

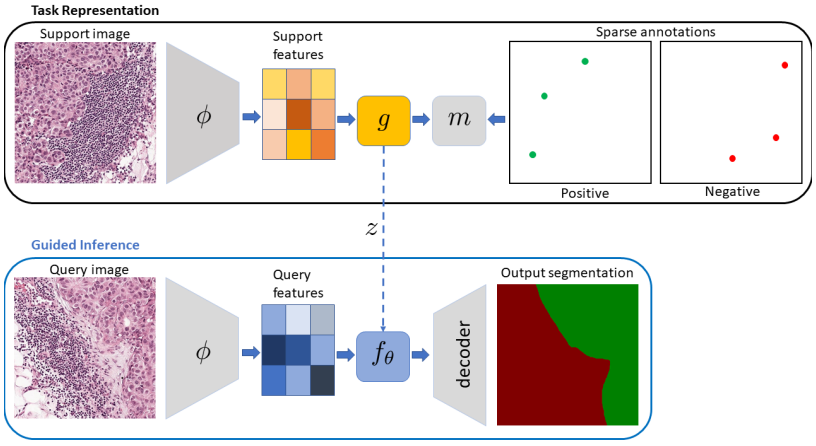

Figure 2: Late-fusion implementation of the guided network (see 3.3.1 for details). The support set (upper branch) with the corresponding positive and negative sparse annotations provide the input to the top branch of the network which derives a "Task Representation", z (the "latent representation"), that is sent as an input to the "Guided Inference" branch. This is the part of the network which is responsible to produce the output segmentation.

- In the "Task Representation" branch, the support set with sparse annotations is fed as input and the network extracts a *latent representation* $\mathbf{z}$ of the task;
- the latent representation $\mathbf{z}$ is passed to the "Guided Inference" branch to direct the output segmentation of the query image.

The network weights are learned end-to-end by computing the loss between the predicted and the actual (dense) segmentation of the query image (see Figure 1). Once trained the network needs no further optimization and, importantly for us, can incorporate additional annotations to alter the task or correct errors. According to Rakelly et al. (2018)

> a "guided" model is both able to make predictions on its own and incorporate expert guidance for directing the task or correcting errors.

3.3. Adapting the model to histopathological images

3.3.1. Dataset selection and preparation

To create the support and query set, for both training and validation, the Camelyon16 dataset of histopathological images was used. From each WSI in that dataset, we extracted patches of size 448x448 with a stride of 224. A subset of 81 WSIs containing lesions were used; in this subset, images with id from 1 to 60 were used for training, while images with id from 61 to 81 were used for validation.

The set of patches obtained from the selected images needed filtering due to the presence of background – just white slide or containing a very small amount of tissue – or other non-tissue artifacts. This was done by a logit of the RGB image converted to HSV with the following formula

$$logit(p(foreground)) = -78.6801 + 0.237\bar{V} + 0.9713\bar{S} + 15.6831V_{fg} \tag{1}$$

where $\bar{V}$ and $\bar{S}$ are the mean values of the corresponding channels in the patch, and V_{fg} is the percentage of foreground in the V-channel as computed by the Otsu filter. The parameters in (1) were found by applying a logistic regression to the problem of reliable classification of foreground vs. background. Patches having a $p(foreground) < 0.9$ where discarded.

After filtering, the resulting dataset was imbalanced toward the tissue class with 241543 patches marked as tissue and 31538 patches marked as lesion. Dataset re-balancing was performed by selecting a subset of tissue patches at random[3].

3.3.2. VGG-16 pre-training

As it will be explained below, the image encoders used for implementing our network are derived from the VGG-16 architecture (Simonyan and Zisserman, 2014). The publicly-available weights for VGG-16 originating from the training on the ImageNet dataset (Russakovsky et al., 2015) were adopted as the starting point for our work. Subsequently, those weights were fine-tuned using the set of filtered patches described in the previous section.

3. The training set was composed by 23577 patches classified as lesion and 23737 classified as tissue. The validation set was composed by 7961 patches classified as lesion and 7801 classified as tissue

For the purpose of VGG-16 fine-tuning, each patch needed to be labeled as belonging to one specific class. In our case, each patch was classified to be either lesion or tissue by relying on the dense annotations associated; a patch was labeled to be lesion if at least one pixel in the center window of size 224x224 was annotated as lesion, otherwise it was classified as tissue (Liu et al., 2017).

3.3.3. Guided network training and validation

With reference to Figure 2, in the network architecture of choice guidance is extracted by the guide $g((\mathbf{x}, \mathbf{L}))$ as a latent representation $\mathbf{z}$, whereas inference is carried out by $f_\theta(\mathbf{x}^*, \mathbf{z})$ using $\mathbf{z}$ alone. Two different configurations of the guided network have been used for the experiments:

- *Late Fusion.* This is the network represented in Figure 2, which is also the reference implementation in Rakelly et al. (2018). The visual features are extracted from the images by $\phi(\mathbf{x})$. The annotations $\mathbf{L}^q$, where $q \in \{+, -\}$ (i.e. foreground/background), are mapped in the feature layer coordinates with the map m. The map m is a fixed bilinear interpolation for downsampling implemented as fractionally strided convolution (Dumoulin and Visin, 2016). The images features $\phi(\mathbf{x})$ and the output of $m(\mathbf{L}^q)$ are then fused with an element-wise multiplication ψ

  $$g_{late}(\mathbf{x}, \mathbf{L}^+, \mathbf{L}^-) = \psi(\phi(\mathbf{x}), m(\mathbf{L}^+), m(\mathbf{L}^-))$$

  With *late fusion* new annotations can be added interactively during inference, making real-time collaborative segmentation possible.

- *Early Fusion.* In this configuration, not represented in figures, the support images and the annotations are concatenated in a channel-wise fashion and used as input to the support feature extractor ϕ. This approach, which is similar to the stacking of images with "positive" and "negative clicks" by Xu et al. (2016), has the disadvantage that the encoder ϕ used for the support set must be different from the one used for the query set. The network training is also substantially slower, over 5 times slower, than with late-fusion.

The encoders ϕ used in both configurations are based on a VGG-16 network in which the top, fully-connected layers were replaced by fully convolutional layers as in Shelhamer et al. (2016). The decoder consists of a bilinear interpolator for upsampling implemented as fractionally strided convolution, where the initial filter was a bilinear upsampling kernel[4]. The advantage of such architecture is that the decoder is just another convolution layer hence allows performing backpropagation.

The input pair $(\mathbf{x}, \mathbf{L})$ to the guide g is decomposed across receptive fields (represented as boxes in color inside each map in Figure 2) so that each receptive field $(\mathbf{x_j}, \mathbf{L_j})$ is associated a latent representation $\mathbf{z}_j = g((\mathbf{x}_j, \mathbf{L}_j))$, called *local latent representation*. Since we need to learn a visual segmentation which is position invariant, all local latent representations $\mathbf{z}_j$ are pooled together to provide the global task representation $\mathbf{z} = m_P(\{\mathbf{z}_j : \forall j\})$. The adopted pooling function m_P was average pooling.

4. See http://avisynth.nl/index.php/Resampling.

4. Experimental results and discussion

We tested the algorithm above with late and early fusion. Tests were performed with 1 and 5 shots per episode, with 5 points and 10 points sparse annotations and with dense annotations as well.

The results for the late-fusion strategy are summarized in Table 1 and for the early-fusion variant in Table 2 . For both fusion strategies the values reported were obtained in

Table 1: Guided networks with late fusion of features and annotation masks

Shots	**Annotation**	**Accuracy Tissue**	**Accuracy Lesion**	**IOU Tissue**	**IOU Lesion**
1	5 points	0.9557	0.9497	0.8850	0.9010
1	10 points	0.9439	0.9416	0.8854	0.8994
1	dense	0.9552	0.9643	0.8983	0.9260
5	5 points	0.9308	0.9773	0.8861	0.9290
5	**10 points**	**0.9611**	**0.9575**	**0.9036**	**0.9337**
5	dense	0.9499	0.9634	0.9048	0.9354

Table 2: Guided networks with early fusion of support images and annotations

Shots	**Annotation**	**Accuracy Tissue**	**Accuracy Lesion**	**IOU Tissue**	**IOU Lesion**
1	5 points	0.9427	0.9582	0.7484	0.8675
1	10 points	0.9160	0.9499	0.8049	0.8922
1	dense	0.9466	0.9660	0.8389	0.8946
5	5 points	0.9051	0.9427	0.7946	0.8967
5	**10 points**	**0.9291**	**0.9614**	**0.7854**	**0.8821**
5	dense	0.9717	0.9681	0.8964	0.9324

the best run, measured by the overall accuracy, across 104000 iterations[5]. We report, for each s-shot P-points configuration, the results obtained of per-class accuracy and per-class Intersection Over Union (IOU).

For a comparative evaluation, we trained a FCN-32s as described in Shelhamer et al. (2016) in fully supervised mode and with dense annotations. In this latter case, we obtained a per-class accuracy of 0.9509 on tissue and 0.9517 on lesions, an IOU of 0.8878 on tissue and an IOU of 0.9210 on lesions. As it can be seen, the late model with 5 shots and 10 annotation points (in bold in Table 1) is better than a similar configuration using early fusion (in bold in Table 2) and is competitive with a standard FCN-32s network with dense annotations, according to the results reported above.

5. Some of the early-fusion experiments did not run for the entire 104000 iterations due to time constraints (early-fusion experiments take much longer than late-fusion experiments). In such cases we could complete 88000 iterations for early-fusion 1-shot, 10 points annotation, 72000 for 1-shot, dense annotation, and 56000 for 5-shots, 10 points annotation

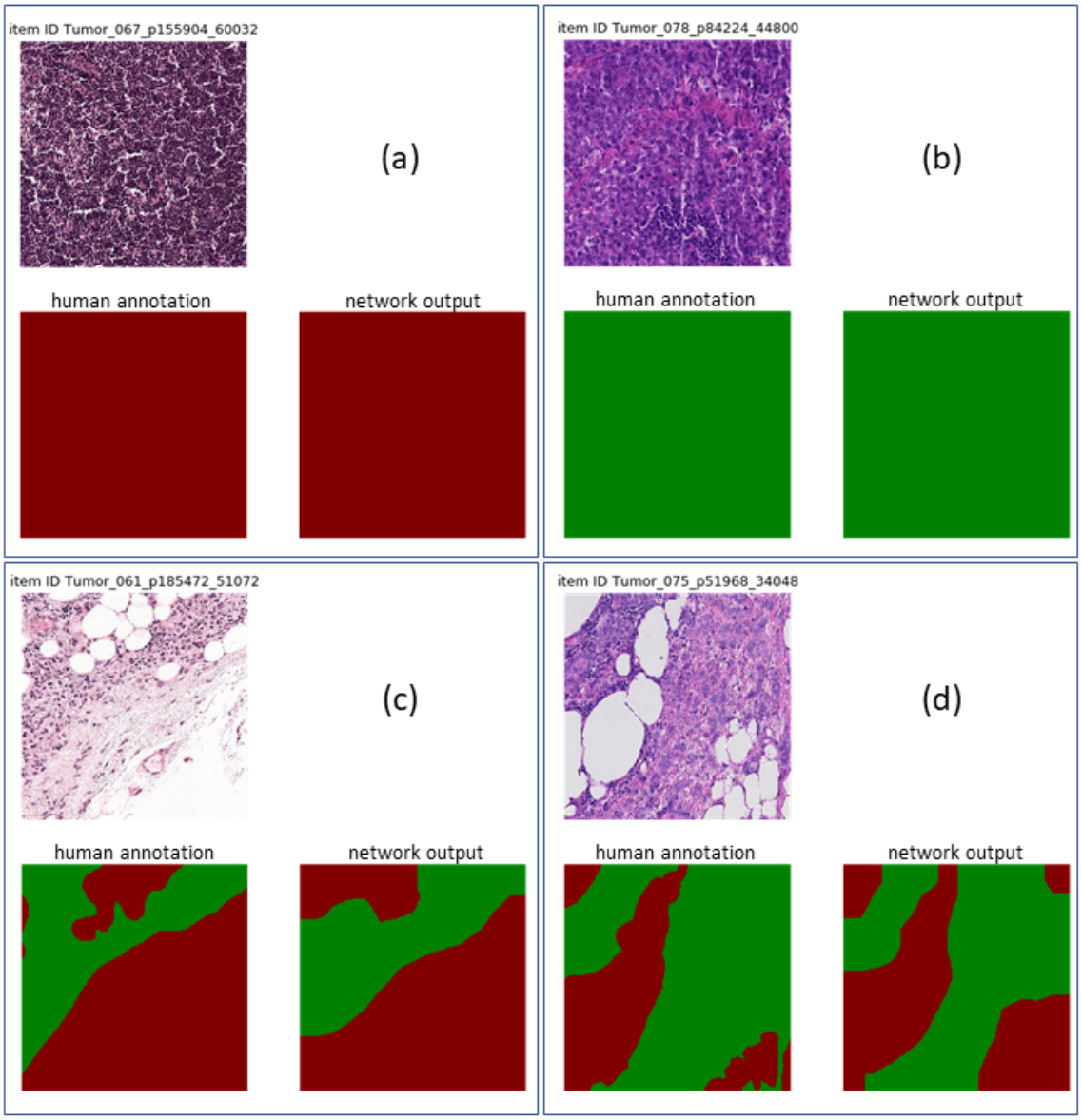

Figure 3: Comparative examples of segmentation results produced with a late fusion, 5-shots, 10 points configuration. Normal tissue is brown and lesions are green. Patches in the top row contain tissue of one class only, respectively tissue and lesion, whereas patches in the bottom row contain tissue of both classes.

5. Conclusions and future work

We have applied a guided network that performs few-shot segmentation with sparse annotations to histopathological images and we have compared the results with different configurations to a 'standard' approach using an FCN-32s architecture with dense annotations. On a balanced validation set of lesion and tissue patches, few-shot segmentation with late fusion, 5 shots and as low as 10 annotations compares favorably to the FCN results. In future work, we plan to test the few-shot segmentation algorithm on new test data extracted from Camelyon17 (Bandi et al., 2018). We will also compare the results to different deep segmentation algorithms such as dilated FCN (Garcia-Garcia et al., 2017). In addition, we plan to create different validation tests in which the segmentation produced could be modified by extending the support set according to the observations an expert pathologist.

Acknowledgments

We thank i-data.tek for the sponsorship of this research. We also thank Prof. Patrizia Morbini of the Department of Molecular Medicine at the University of Pavia for her helpful advises on the clinical practice of diagnostic histopathology.

References

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, Quanzheng Li, Farhad Ghazvinian Zanjani, Svitlana Zinger, Keisuke Fukuta, Daisuke Komura, Vlado Ovtcharov, Shenghua Cheng, Shaoqun Zeng, Jeppe Thagaard, Anders B. Dahl, Huangjing Lin, Hao Chen, Ludwig Jacobsson, Martin Hedlund, Melih Cetin, Eren Halici, Hunter Jackson, Richard Chen, Fabian Both, Jorg Franke, Heidi Kusters-Vandevelde, Willem Vreuls, Peter Bult, Bram van Ginneken, Jeroen van der Laak, and Geert Litjens. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, pages 1–1, 2018. ISSN 0278-0062. doi: 10.1109/TMI.2018.2867350. URL https://ieeexplore.ieee.org/document/8447230/.

Y. Bengio, S. Bengio, and J. Cloutier. Learning a synaptic learning rule. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume ii, page 969. IEEE, 1991. ISBN 0-7803-0164-1. doi: 10.1109/IJCNN.1991.155621. URL http://ieeexplore.ieee.org/document/155621/.

Vincent Dumoulin and Francesco Visin. A guide to convolution arithmetic for deep learning. mar 2016. URL http://arxiv.org/abs/1603.07285.

Alberto Garcia-Garcia, Sergio Orts-Escolano, Sergiu Oprea, Victor Villena-Martinez, and Jose Garcia-Rodriguez. A Review on Deep Learning Techniques Applied to Semantic Segmentation. apr 2017. URL http://arxiv.org/abs/1704.06857.

Ben Glocker, Darko Zikic, Ender Konukoglu, David R. Haynor, and Antonio Criminisi. Vertebrae Localization in Pathological Spine CT via Dense Classification from Sparse Annotations. pages 262–270. Springer, Berlin, Heidelberg, 2013.

doi: 10.1007/978-3-642-40763-5_33. URL http://link.springer.com/10.1007/978-3-642-40763-5{_}33.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building Machines That Learn and Think Like People. apr 2016. URL http://arxiv.org/abs/1604.00289.

Yun Liu, Krishna Kumar Gadepalli, Mohammad Norouzi, George Dahl, Timo Kohlberger, Subhashini Venugopalan, Aleksey S Boyko, Aleksei Timofeev, Philip Q Nelson, Greg Corrado, Jason Hipp, Lily Peng, and Martin Stumpe. Detecting cancer metastases on gigapixel pathology images. Technical report, arXiv, 2017. URL https://arxiv.org/abs/1703.02442.

Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A. Efros, and Sergey Levine. Few-Shot Segmentation Propagation with Guided Networks. may 2018. doi: 10.1002/tox. URL http://arxiv.org/abs/1806.07373.

Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel. Meta-Learning for Semi-Supervised Few-Shot Classification. mar 2018. ISSN 2211-2855. doi: 10.5121/ijci.2015.4227. URL http://arxiv.org/abs/1803.00676.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. ISBN 9783319245737. doi: 10.1007/978-3-319-24574-4_28.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Jürgen Schmidhuber. Evolutionary principles in self-referential learning. Master's thesis, Institut f. Informatik, Tech. Univ. Munich, 1987. URL http://people.idsia.ch/~juergen/diploma.html.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. may 2016. URL http://arxiv.org/abs/1605.06211.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep Learning in Medical Image Analysis. *Annual review of biomedical engineering*, 19:221–248, 2017. ISSN 1545-4274. doi: 10.1146/annurev-bioeng-071516-044442. URL http://www.ncbi.nlm.nih.gov/pubmed/28301734http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5479722.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. sep 2014. URL http://arxiv.org/abs/1409.1556.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching Networks for One Shot Learning. jun 2016. URL http://arxiv.org/abs/1606.04080.

Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep Interactive Object Selection. mar 2016. URL http://arxiv.org/abs/1603.04042.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? nov 2014. URL http://arxiv.org/abs/1411.1792.