Adverse Effects of Image Tiling on Convolutional Neural Networks

G. Anthony Reina(✉) and Ravi Panchumarthy

Artificial Intelligence Products Group, Intel Corporation, Hillsboro, OR, USA
g.anthony.reina@intel.com
https://ai.intel.com/

Abstract. Convolutional neural network models perform state of the art accuracy on image classification, localization, and segmentation tasks. A fully convolutional topology, such as U-Net, may be trained on images of one size and perform inference on images of another size. This feature allows researchers to work with images too large to fit into memory by simply dividing the image into small tiles, making predictions on these tiles, and stitching these tiles back together as the prediction of the whole image.

We compare how a tiled prediction of a U-Net model compares to a prediction that is based on the whole image. Our results show that using tiling to perform inference results in a significant increase in both false positive and false negative predictions when compared to using the whole image for inference. We are able to modestly improve the predictions by increasing both tile size and amount of tile overlap, but this comes at a greater computational cost and still produces inferior results to using the whole image.

Although tiling has been used to produce acceptable segmentation results in the past, we recommend performing inference on the whole image to achieve the best results and increase the state of the art accuracy for CNNs.

1 Introduction

Since their resurgence in 2012 convolutional neural networks (CNN) have rapidly proved to be the state-of-the-art method for computer-aided diagnosis in medical imaging and have led to improved accuracy in classification, localization, and segmentation tasks [1,2]. However, memory constraints have often limited training on large 2D and 3D images due to the size of the activation maps held for the backward pass during gradient descent [3]. Two methods are commonly used to manage these memory limitations: (1) images are often downsampled to a lower resolution and/or (2) images are broken into smaller tiles [4].

Fully convolutional networks are a natural fit for tiling methods because they can be trained on one image size and perform inference on another. These networks can perform inference on arbitrarily large images by breaking the large image into smaller, overlapping tiles [5]. However, we question whether this overlapping tiles approach is indeed as accurate as simply performing inference on

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11383, pp. 25–36, 2019.
https://doi.org/10.1007/978-3-030-11723-8_3

the whole image. In this report, we design an experiment where the whole image can fit within the memory limitations and compare whole image inference to the overlapping tiles approach.

2 Methods

2.1 Brain Tumor Segmentation Dataset (BraTS)

The brain tumor segmentation (BraTS) challenge created a publicly-available multi-institutional dataset for benchmarking and quantitatively evaluating the performance of computer-aided segmentation algorithms to detect gliomal brain tumors from MRI [6–9]. In this study we use the 2018 BraTS dataset which is comprised of pre-operative MRI scans from 285 patients at 19 institutions (https://www.med.upenn.edu/sbia/brats2018/data.html). The scans were performed on 1T, 1.5T, or 3 T multimodal MRI machines and all the ground truth labels were manually annotated by expert, board-certified neuroradiologists.

2.2 U-Net

We implemented a U-Net topology which predicts tumor segmentation masks from the raw MRI slices [5] (Fig. 1). U-Net is a fully convolutional network based on an encoder-decoder architecture. Because of its design U-Net is agnostic to image size. Training and inference can be performed on different sized images.

Our model takes as input a single T2 Fluid Attenuated Inversion Recovery (FLAIR) slice from the BraTS dataset and outputs an equivalently-sized

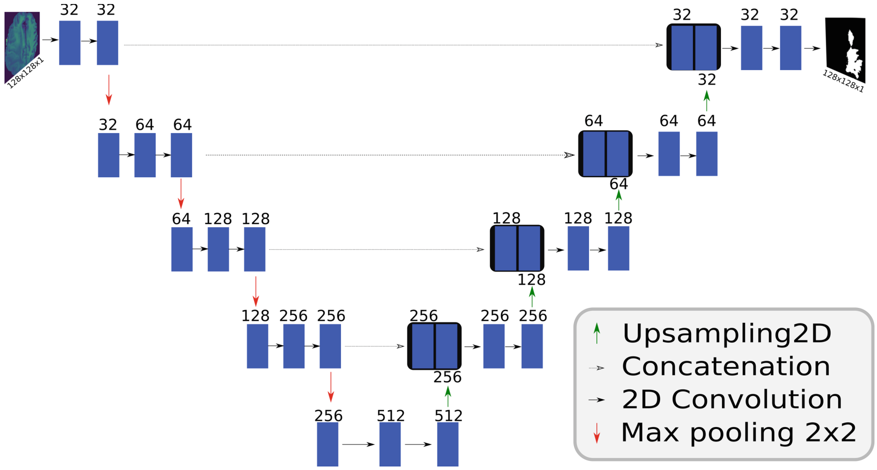

Fig. 1. The U-Net topology used in this study. We reduced the number of kernels in each layer by half from the original paper [5] and added dropout just before the 3rd and 4th max pooling layers.

mask predicting the whole tumor. The contracting path captures context (via max pooling) and the expanding path enables localization (via upsampling). Unlike the standard encoder-decoder, each feature map in the expanding path is concatenated with a corresponding feature map from the contracting path, augmenting downstream feature maps with spatial information acquired using smaller receptive fields. Intuitively, this allows the network to consider features at various spatial scales. Since its introduction in 2015, U-Net has quickly become one of the standard deep learning topologies for image segmentation. We modified the published topology by reducing the number of feature maps by half and adding dropout (0.2) just before the 3rd and 4th max pooling layers.

2.3 Training

We divided the BraTS 2018 dataset into a training/test split of approximately 85/15. Although we are considering the 2D slices from the MRI to be independent images for the model, we ensured that the 2D slices from a single study were contained in only one of the two datasets in order to prevent data leakage. There were 35,960 image/mask pairs in the training set and 8,215 in the test set. The FLAIR channels from each 2D slice were normalized by subtracting the mean pixel value of the slice and dividing by the standard deviation of the pixel values from the slice. The original slices were 240×240 pixels (*i.e.* whole image). A random crop of 128×128 pixels was taken from the normalized FLAIR slices and their corresponding ground truth masks. We performed randomized flipping (up/down and left/right) and 90 degree rotation of the training set images.

The Dice coefficient was used to measure the quality of the tumor predictions. Dice is defined as:

$$\frac{2 \times |Truth \cap Prediction| + smooth}{|Truth| + |Prediction| + smooth}$$

where $Truth$ is the ground truth tumor mask and $Prediction$ is the predicted tumor mask. A smoothing factor of 1.0 is added to both numerator and denominator for numerical stability in the case of non-existent ground truth masks.

The model was implemented in Keras 2.2 and TensorFlow 1.10. The complete source code can be found at https://github.com/NervanaSystems/topologies/tree/master/tiling_experiments. Stochastic gradient descent with the Adam optimizer (learning rate = 0.0001) was used to minimize $-\log Dice$. A batch size of 128 was used during training. We created a batch generator which randomly selected cropped images/masks from the training set for each batch. We trained 30 epochs (280 steps per epoch) and saved the model that produced the best Dice loss on the test dataset.

2.4 Tiling

Tiling is typically applied when using large images due to the memory limitations of the hardware (Fig. 2). For the current experiment we specifically chose the topology and dataset so that it would fit within 16 GB of RAM at inference time for a batch size of at least 1. Our goal is compare inference based on the whole image to inference based on an overlapping tiled version of the whole image.

To perform the overlapping tiling at inference time, our algorithm cropped 128×128 patches from the whole image at uniformly spaced offsets in both the horizontal and vertical dimensions. Inference was performed individually on the patch and the prediction mask was added to a running mean prediction of the whole tumor mask. As highlighted in Fig. 2, (*Middle*), two tiles may produce different predictions for the pixels they share in their overlap. This results in a lower confidence (green pixels) prediction for those pixels rather than the highly positive (yellow) or highly negative (blue) confidence present in the predictions of the individual tiles (and when predicting the whole image) (Fig. 2, *Right*).

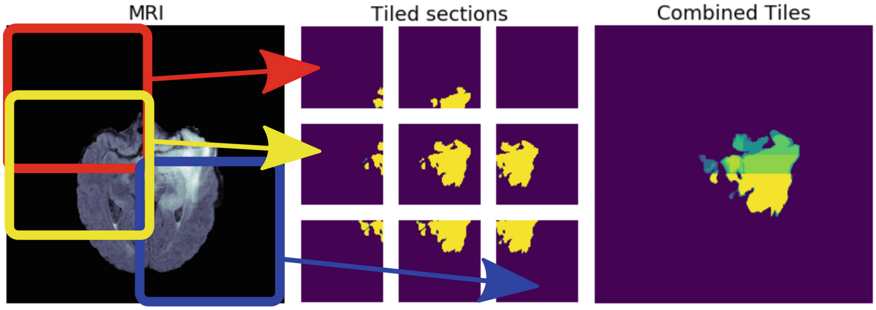

Fig. 2. An illustration of tiling. At training time random crops of the images/masks are used to build the model. At inference time, a series of overlapping crops from the whole image (depicted by the red, yellow, blue tiles) are input separately into the model and the multiple outputs are then reassembled and averaged to generate a prediction of the whole image. For the combined tile prediction (*Right*): yellow = high probability of tumor; blue = low probability of tumor; green = moderate probability of tumor (Color figure online)

3 Results

3.1 Single, Center Crop Tile

Figure 3 shows the result from using just a single, center 128×128 crop of the whole image to do prediction. For this case, a center crop removes the border from the MRI and retains most of the brain. The entire ground truth mask is contained within the cropped region. Nevertheless, the prediction using the entire image (including the border) gave a superior prediction (Dice 0.90 versus 0.69). This is of particular concern because it shows that a smaller patch itself– without any overlapping tiles– can produce inferior predictions to the whole image.

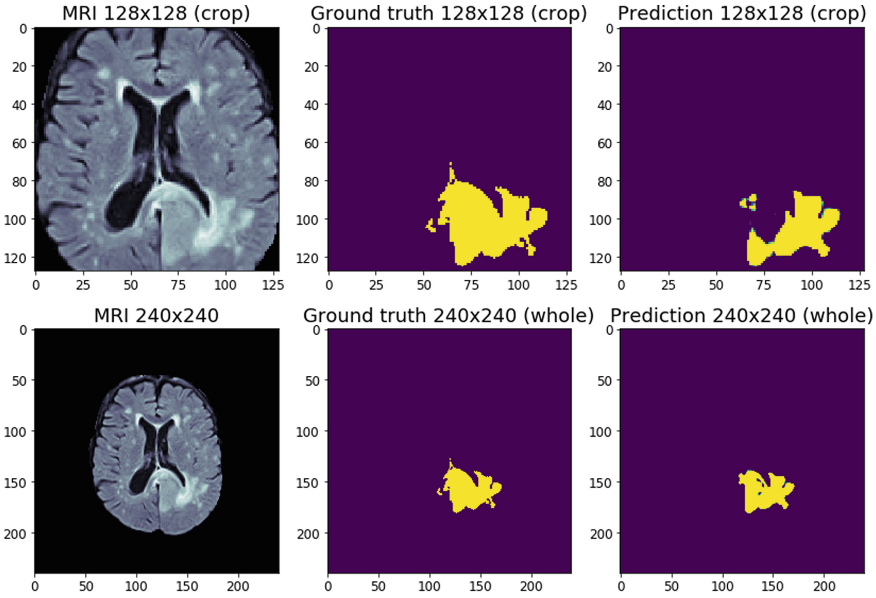
MRI 128x128 (crop)
Ground truth 128x128 (crop)
Prediction 128x128 (crop)
MRI 240x240
Ground truth 240x240 (whole)
Prediction 240x240 (whole)
0
20
40
60
80
100
120
25
50
75
125
150
200

Fig. 3. The prediction from the 128 × 128 center crop of the MRI image (*Top*) is poorer (Dice = 0.69) than one based on the full 240 × 240 MRI image (Dice = 0.90) (*Bottom*) even though the model was trained on 128 × 128 image patches.

3.2 Multiple, Overlapping Tiles

Figure 4 shows a false positive prediction. In this case, the prediction based on using the whole image is correct (*i.e.* no tumor, *Right*), but the prediction based on overlapping tiles shows a high confidence (yellow) of tumor present in the MRI (*Left*).

In Fig. 5 we experiment with smaller tile dimensions and greater overlap between tiles. In most cases, a larger tile size and a larger overlap between successive tiles improved the prediction, but did not completely resolve the false positives and negatives.

For the 8,215 images in the test dataset, we found that using the whole image gave a 0.045 average increase in the Dice coefficient when compared to the tiling approach (Dice = 0.874 for whole image versus 0.829 for tiled approach, Fig. 6). In 7,146 of the cases (87%) using the whole image gave a better than or equal to Dice metric than the tiled approach. In 617 of the cases (7.5%), using the whole image gave more than a 0.1 increase in the Dice coefficient. In 63 of the cases (0.7%) using the tiled approach gave more than a 0.1 increase in the Dice coefficient. However, many of these cases involved very small ground truth masks and poor predictions by both models which may have led to the result (Fig. 7).

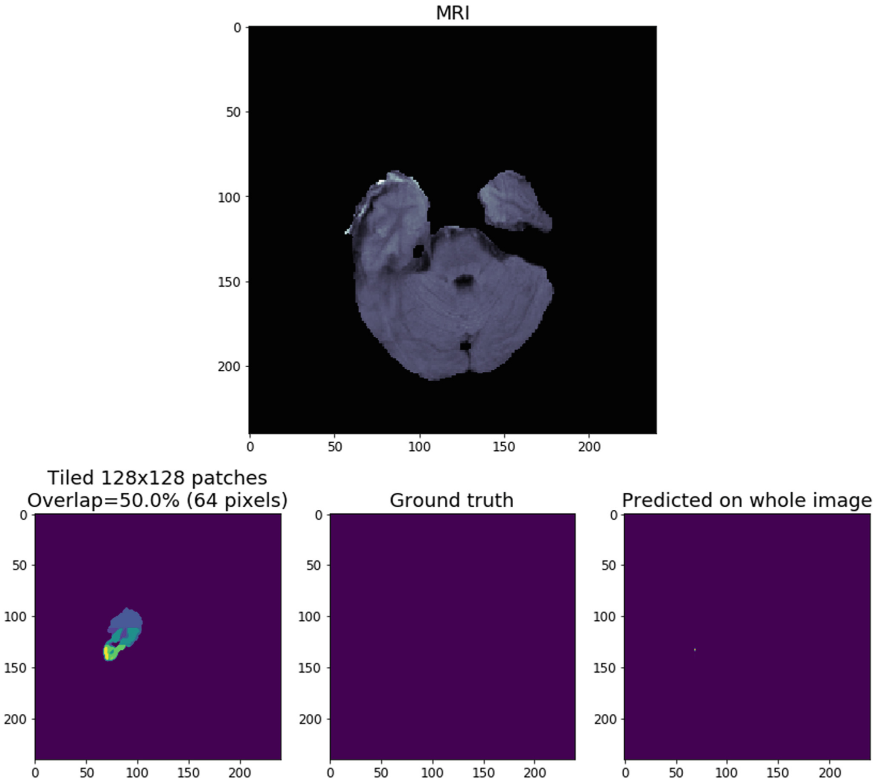

Fig. 4. 128×128 windows were slid over the whole image (*Top*) with a 50% window overlap in the horizontal and vertical directions. These 9 tiles were averaged to produce the tiled prediction (*Bottom left*). Although the tiled prediction does capture most of the true tumor, there are many false positive predictions in the tiled result. (*Bottom right*) The prediction using the whole image is far superior. (*Bottom center*) The radiologist's ground truth for the tumor. (Color figure online)

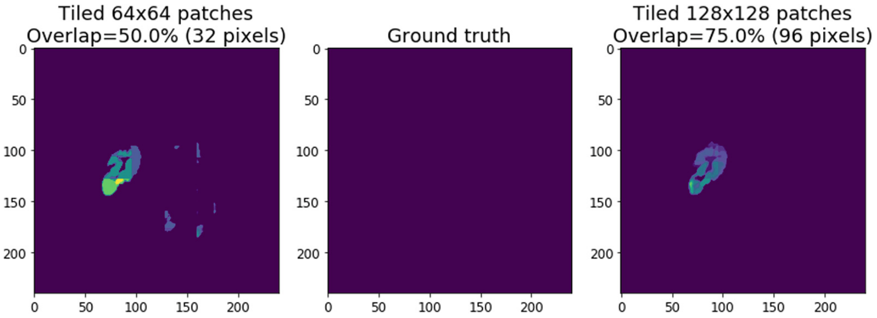

Fig. 5. Using a smaller patch size of 64×64 produces more false positives with the tiling method (*Left*). Using a greater overlap between patches does not completely reduce the false positive predictions (*Right*)

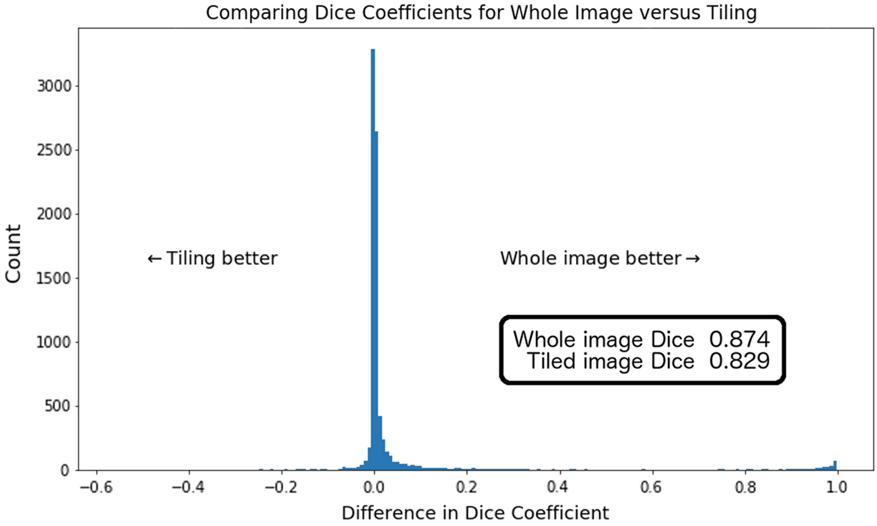

Fig. 6. The Dice metric between the prediction and the ground truth masks were compared for each image in the test dataset. The histogram shows the pairwise difference between using the whole image size and using the sliding tile method. The Dice metric using the whole image was on average 0.045 larger than using the tiling approach. Predictions based on the whole image were better or equal to tiling in 87% of the test dataset.

In Figs. 8, 9, and 10 we show three examples of where tiling leads to significant false negative predictions. In Fig. 8 tiling produces a low confidence prediction of the tumor mask (green pixels) whereas prediction on the whole image produces a high confidence prediction (yellow pixels). If a simple thresholding were used in this case, the tiling approach would result in a completely false negative prediction. In Fig. 9 we show that using the whole image was able to accurately predict even a very small tumor mask, but the tiling method on the same image again fails to detect the tumor with a significant confidence. In Fig. 10 we should that the tiling method produces a false negative prediction on the superior half of the tumor (green pixels).

4 Discussion

The overlapping tile approach is commonly used by researchers to apply fully convolutional models on large 2D and 3D images that would ordinarily not fit into available memory. Although this method works, we have demonstrated that clearly better predictions can be obtained by applying the convolutional model to the whole image.

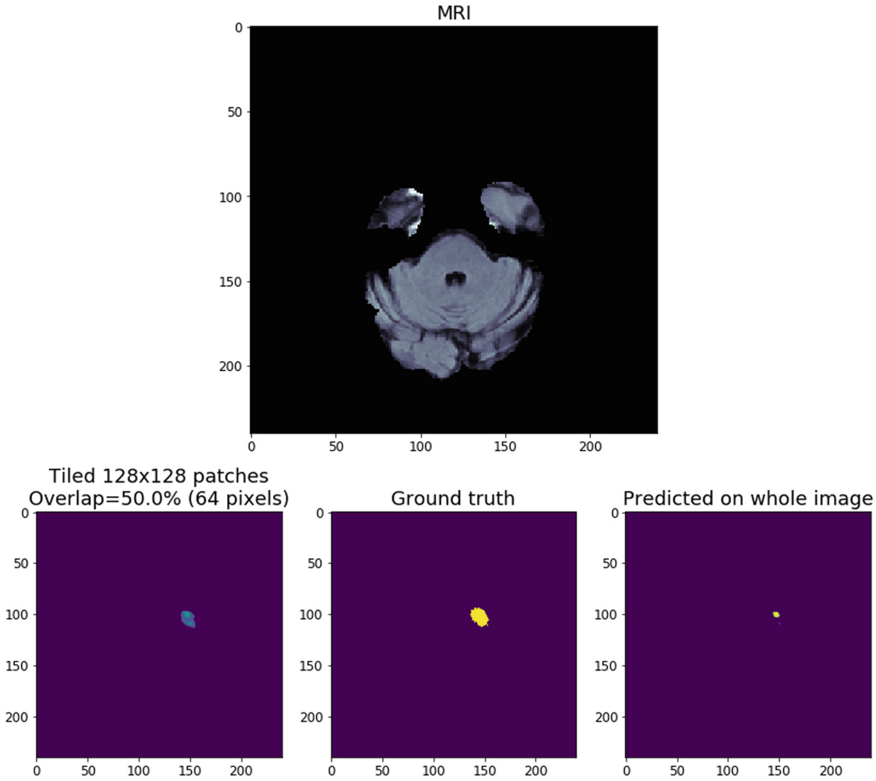

Fig. 7. One of the 0.7% of the cases where the tiled prediction gave a larger Dice coefficient (0.32) than the whole image prediction (0.21). In most of these cases, there were small ground truth masks and poor predictions generated by both methods.

Our results show that the tiling method often produces false positive and false negative predictions (Figs. 4, 8, 9, and 10). These false predictions can be reduced through a combination of increasing the tile dimensions and increasing the tile overlap. However, both of these corrections greatly increase the computational complexity of the model. For example, in our experiment, we used a tile dimension that was about one quarter size of the whole image (128×128 versus 240×240). With a 50% overlap between tiles, this cost 3 model predictions for each dimension– a total of 9 forward passes of the model in order to predict the whole image. If a 75% overlap were used instead, this would increase to 16 predictions (4×4). Therefore, the tiling approach can often lead to an $O(n^2)$ number of calculations for 2D images and a $O(n^3)$ for 3D images. These additional computations may easily negate any speed advantage attained by the tiling approach over the whole image approach– and at an increased cost in accuracy.

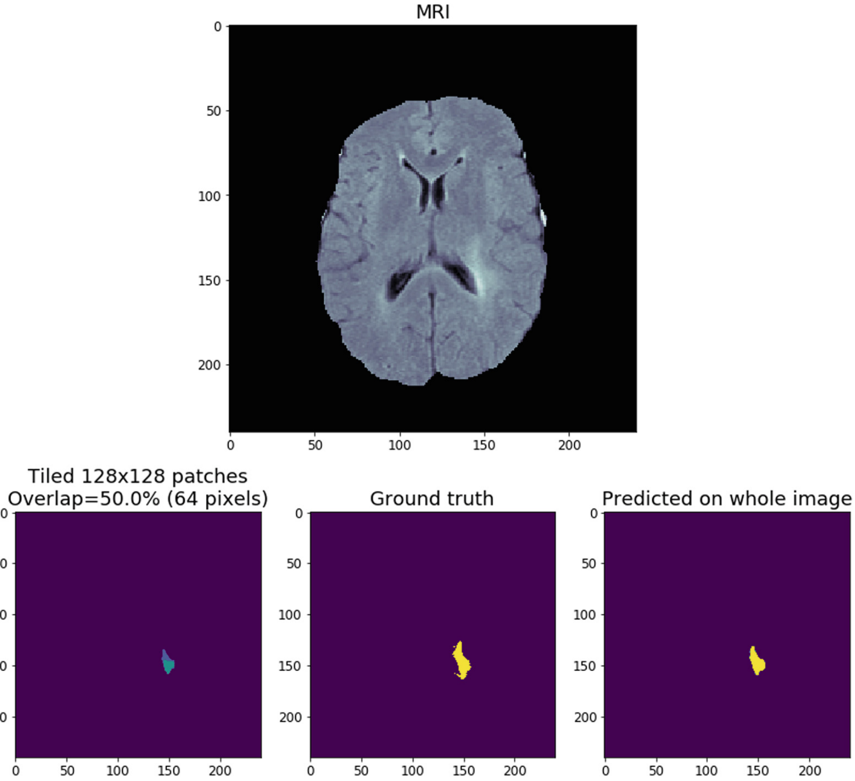

Fig. 8. The tiling approach produced a large false negative prediction (*Left*). If the prediction had been thresholded (*i.e.* $<0.5 = 0$; $\geq 0.5 = 1$) then tiling would have completely failed to predict the tumor within this slice (*Left*). (Color figure online)

Table 1. Memory and accuracy variations with training and inference methods

	Inf: FULL	Inf: TILED
Train: FULL	Inf Mem Req: High Inf Accuracy: High	Inf Mem Req: Low Inf Accuracy: Low
Train: TILED	Inf Mem Req: High Inf Accuracy: High?	Inf Mem Req: Low Inf Accuracy: High?

We hypothesize that the difference in predictions between the whole image and a tile may be due to the combination of max pooling and non-linear activation functions of the network. Note that this difference occurs even without averaging overlapping tile predictions (cf. Fig. 3). The larger field of view provided by the whole image approach may give a richer set of features at both training and inference time necessary to "push" the information past the ReLU activations and into the larger receptive fields created by the max pooling layer.

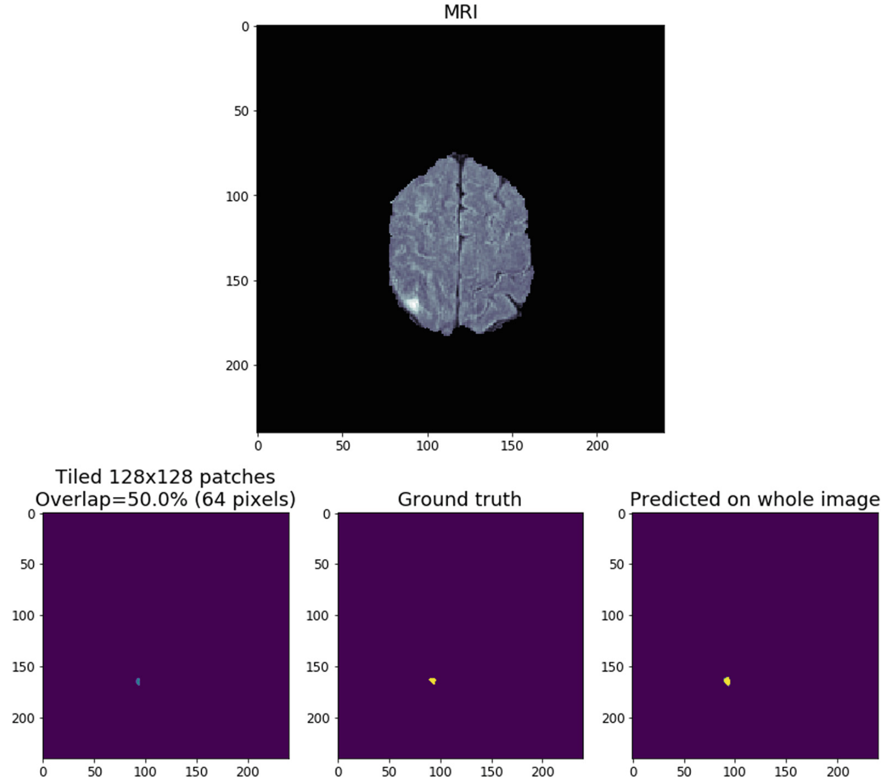

Fig. 9. Prediction on the whole image (*Right*) typically outperforms prediction using the tiling approach (*Left*) even when the ground truth masks are very small.

For example, a tumor might be too small to reliably detect in a narrowly defined field of view, but– over a larger field of view– the mass effect it produces on the surrounding tissue may provide enough additional feature information to increase the model's confidence. We liken this to a small pebble in a pond. The pebble may be too small to see, but its ripple effect on the surrounding water may still allow it to be detected.

We believe that researchers are artificially simplifying their models in order to fit into the memory limitations of current hardware. We suggest that instead researchers should be working with hardware and software manufacturers to more easily allow for greater memory capacity (Table 1). We believe that such efforts would help these models to be wider, deeper, and allow for larger inputs so that they can move to a new level in state of the art accuracy.

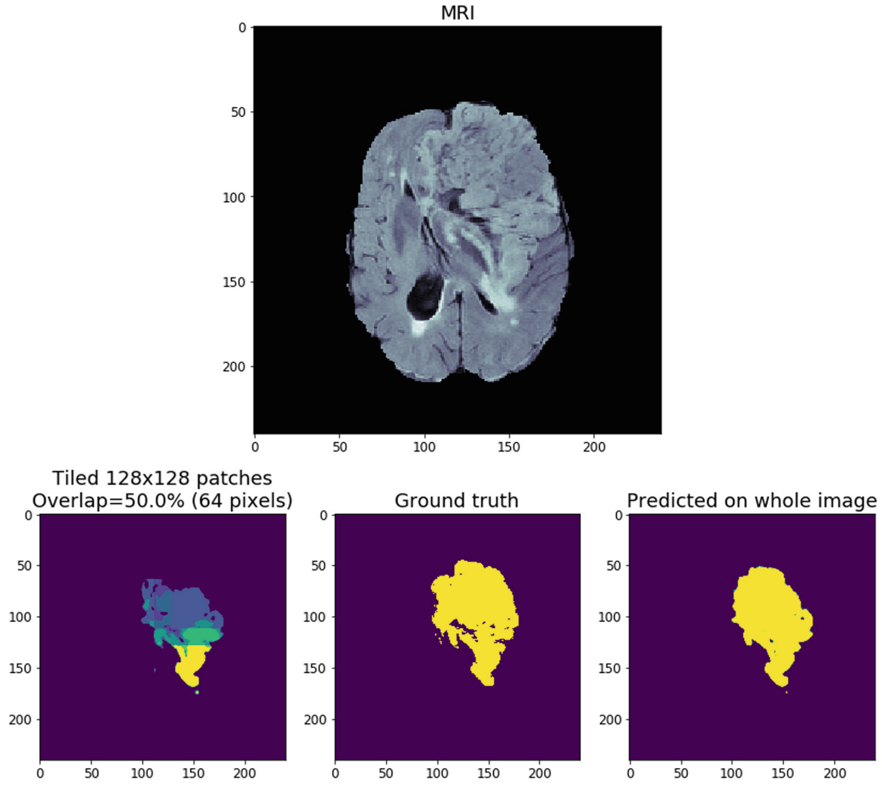

Fig. 10. The tiling methods the tiling method produces a false negative prediction on the superior half of the tumor (*Left*) but the whole image prediction correctly predicts the entire mass (*Right*). (Color figure online)

References

1. Greenspan, H., Van Ginneken, B., Summers, R.M.: Deep learning in medical imaging: overview and future promise of an exciting new technique. IEEE Trans. Med. Imaging **35**(5), 1153–1159 (2016)
2. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017). https://doi.org/10.1038/nature21056
3. Tianqi, C., Bing, X., Chiyuan, Z., Guestrin, C.: Training deep nets with sublinear memory cost. arXiv:1604.06174v2 [cs.LG] 22 April 2016
4. Pinckaers, J.H.F.M., Litjens, G.J.S.: Training convolutional neural networks with megapixel images. arXiv:1804.05712v1 [cs.CV] 16 April 2018
5. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation, 18 May 2015. https://arxiv.org/abs/1505.04597
6. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694

7. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nature Sci. Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
8. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
9. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF