# 1 Introduction

Automated medical image segmentation has been extensively studied in the image analysis community due to the fact that manual, dense labelling of large amounts of medical images is a tedious and error-prone task. Accurate and reliable solutions are desired to increase clinical work flow efficiency and support decision making through fast and automatic extraction of quantitative measurements. With the advent of convolutional neural networks (CNNs), near-radiologist level performance can be achieved in automated medical image analysis tasks including cardiac MR segmentation [3] and cancerous lung nodule detection [17]. High representation power, fast inference, and filter sharing properties have made CNNs the de facto standard for image segmentation. Fully convolutional networks (FCNs) [18] and the U-Net [24] are two commonly used architectures. Despite their good representational power, these architectures rely on multi-stage cascaded CNNs when the target organs show large inter-patient variation in terms of shape and size. Cascaded frameworks extract a region of interest (ROI) and make dense predictions on that particular ROI. The application areas include cardiac MRI [14], cardiac CT [23], abdominal CT [26, 27] segmentation, and lung CT nodule detection [17]. However, this approach leads to excessive and redundant use of computational resources and model parameters; for instance, similar low-level features are repeatedly extracted by all models within the cascade. To address this general problem, we propose a simple and yet effective solution, namely attention gates (AGs). CNN models with AGs can be trained from scratch in a standard way similar to the training of a FCN model, and AGs automatically learn to focus on target 1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands. arXiv:1804.03999v3 [cs.CV] 20 May 2018 structures without additional supervision. At test time, these gates generate soft region proposals implicitly on-the-fly and highlight salient features useful for a specific task. Moreover, they do not introduce significant computational overhead and do not require a large number of model parameters as in the case of multi-model frameworks. In return, the proposed AGs improve model sensitivity and accuracy for dense label predictions by suppressing feature activations in irrelevant regions. In this way, the necessity of using an external organ localization model can be eliminated while maintaining the high prediction accuracy. Similar attention mechanisms have been proposed for natural image classification [11] and captioning [1] to perform adaptive feature pooling, where model predictions are conditioned only on a subset of selected image regions. In this paper, we generalize this design and propose image-grid based gating that allows attention coefficients to be specific to local regions. Moreover, our approach can be used for attention-based dense predictions. We demonstrate the implementation of AG in a standard U-Net architecture (Attention U-Net) and apply it to medical images. We choose the challenging CT pancreas segmentation problem to provide experimental evidence for our proposed contributions. This problem constitutes a difficult task due to low tissue contrast and large variability in organ shape and size. We evaluate our implementation on two commonly used benchmarks: TCIA Pancreas CT-82 [25] and multi-class abdominal CT-150. The results show that AGs consistenly improve prediction accuracy across different datasets and training sizes while achieving state-of-the-art performance without requiring multiple CNN models.

* 1. Related Work CT Pancreas Segmentation:

Early work on pancreas segmentation from abdominal CT used statistical shape models [5, 28] or multi-atlas techniques [22, 34]. In particular, atlas approaches benefit from implicit shape constraints enforced by propagation of manual annotations. However, in public benchmarks such as the TCIA dataset [25], Dice similarity coefficients (DSC) for atlas-based frameworks ranges from 69.6% to 73.9% [22, 34]. In [39] a classification-based framework is proposed to remove the dependency of atlas to image registration. Recently, cascaded multi-stage CNN models [26, 27, 38] have been proposed to address the problem. Here, an initial coarse-level model (e.g. U-Net or Regression Forest) is used to obtain a ROI and then a cropped ROI is used for segmentation refinement by a second model. Similarly, combinations of 2D-FCN and recurrent neural network (RNN) models are utilised in [4] to exploit dependencies between adjacent axial slices. These approaches achieve state-of-the-art performance in the TCIA benchmark (81.2% − 82.4% DSC). Without using a cascaded framework, the performance drops between 2.0% and 4.4%. Recent work [37] proposed an iterative two-stage model that recursively updates local and global predictions, and both models are trained end-to-end. Besides standard FCNs, dense connections [6] and sparse convolutions [8, 9] have been applied to the CT pancreas segmentation problem. Dense connections and sparse kernels reduce computational complexity by requiring less number of non-zero parameters. Attention Gates: AGs are commonly used in natural image analysis, knowledge graphs, and language processing (NLP) for image captioning [1], machine translation [2, 30], and classification [11, 31, 32] tasks. Initial work has explored attention-maps by interpreting gradient of output class scores with respect to the input image. Trainable attention, on the other hand, is enforced by design and categorised as hard- and soft-attention. Hard attention [21], e.g. iterative region proposal and cropping, is often non-differentiable and relies on reinforcement learning for parameter updates, which makes model training more difficult. Recursive hard-attention is used in [36] to detect anomalies in chest X-ray scans. Contrarily, soft attention is probabilistic and utilises standard back-propagation without need for Monte Carlo sampling. For instance, additive soft attention is used in sentence-to-sentence translation [2, 29] and more recently applied to image classification [11, 32]. In [10], channel-wise attention is used to highlight important feature dimensions, which was the top-performer in the ILSVRC 2017 image classification challenge. Self-attention techniques [11, 33] have been proposed to remove the dependency on external gating information. For instance, non-local self attention is used in [33] to capture long range dependencies. In [11, 32] self-attention is used to perform class-specific pooling, which results in more accurate and robust image classification performance.

* 1. Contributions In this paper:

We propose a novel self-attention gating module that can be utilised in CNN based standard image analysis models for dense label predictions. Moreover, we explore the benefit of AGs to medical image analysis, in particular, in the context of image segmentation.

[https://www.google.com/](#_top)