PET-guided Attention Network for Segmentation of Lung Tumors from PET/CT images

Varaha Karthik Pattisapu[1], Imant Daunhawer[1], Thomas Weikert[2], Alexander Sauter[2], Bram Stieltjes[2], and Julia E. Vogt[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
[2] Clinic of Radiology & Nuclear Medicine, University Hospital Basel, Switzerland

Abstract. PET/CT imaging is the gold standard for the diagnosis and staging of lung cancer. However, especially in healthcare systems with limited resources, costly PET/CT images are often not readily available. Conventional machine learning models either process CT or PET/CT images but not both. Models designed for PET/CT images are hence restricted by the number of PET images, such that they are unable to additionally leverage CT-only data. In this work, we apply the concept of visual soft attention to efficiently learn a model for lung cancer segmentation from only a small fraction of PET/CT scans and a larger pool of CT-only scans. We show that our model is capable of jointly processing PET/CT as well as CT-only images, which performs on par with the respective baselines whether or not PET images are available at test time. We then demonstrate that the model learns efficiently from only a few PET/CT scans in a setting where mostly CT-only data is available, unlike conventional models.

1 Introduction

Lung cancer is the second most frequently diagnosed cancer type and the leading cause of cancer-related deaths in men and women alike with high incidence and mortality rates [14]. For the staging of lung cancer, PET/CT imaging is widely used, because it provides complementary information: while the CT component visualizes anatomical properties, the PET component represents the metabolism. This gives additional information on tumor activity and is important for the detection of metastases. Despite its important role, combined PET/CT imaging is often unavailable, due to logistic and economic constraints.

Unfortunately, conventional machine learning models only cater to CT data or PET/CT data, but not both, which poses a significant problem, especially in resource-constrained populations. Prior work [3, 6] has highlighted this challenge, and several other approaches have been proposed to deal with it [4, 1, 19]. They attempt to learn effective joint representations of PET/CT modalities. However, such approaches still assume that a PET image is available for every CT image during training. This assumption greatly reduces the amount of effective training

data for a combination of CT-only and PET/CT data. Consequently, while such models might be efficient during inference, they fall short in not being able to learn effective joint representations for a combination of CT-only and PET/CT data. The problem is further compounded by the complexity of the data, which typically includes different types of malignant lesions (e.g., the main tumor, lymph nodes metastases, and distant metastases). As such, conventional models typically cannot make the best use of a combination of CT-only and PET/CT data, a typical scenario in resource-constrained environments.

To solve this problem, we apply the established concept of visual soft attention [12]. The attention mechanism allows us to input PET images when they are available. As such, the model benefits from the additional information contained in PET images but does not mandate them. The model is thus flexible to the availability of PET data. Consequently, it is possible to incorporate two separate models that are trained on unimodal (CT) or bimodal (PET/CT) data, respectively, into one single model. Additionally, since, we do not explicitly enforce the attention mechanism to learn a joint representation of PET/CT modalities, our model can be trained on a mix of CT-only and PET/CT images. Thus, our model has the potential to make efficient use of both CT-only and PET/CT data, unlike conventional models. We present the effectiveness of our model on a large dataset with the goal of segmenting tumorous regions. We acronym our model as PAG, which stands for PET-guided attention gate. To summarize, the three main contributions of the current work are:

i) We propose a novel approach for dealing with a combination of CT-only and PET/CT data based on a visual soft attention mechanism.
ii) Our model combines two discrete functions that deal with unimodal or bimodal data, respectively, in a single model.
iii) We demonstrate a realistic application of the model in scenarios when PET/CT images are scarce relative to CT-only images and show how the model makes efficient use of the combination of CT-only and PET/CT data.

2 Related work

Segmentation of anatomical structures such as tumors, lesions and lung nodules from PET/CT images is an active and dynamic area of research within medical imaging. [17] implemented the U-Net architecture [13] for the segmentation of nasopharyngeal tumors from dual-modality PET/CT images. [8] learned a probability map of tumorous regions from a CT image and then used a fuzzy variational model that incorporates the probability map as a prior and the PET images to obtain a posterior probability map of tumorous regions. [5] studied different fusion schemes of multi-modal images, all of which fuse the images at the pixel space. [18] refined the segmentation maps obtained separately from CT and PET images, using a graph-cut based co-segmentation model to refine the segmentation maps. [9] used belief functions to fuse the PET and CT images to obtain segmentation masks of the tumors.

The named methods are based on bi-modal inputs i.e. both PET and CT modalities. Such models, typically assume that a complete set of all modalities used during training is available even during inference. Such methods do not have the capacity to incorporate for missing modalities. Accordingly, several other methods have been proposed that deal with missing modalities. [6] proposed the Hemis model to extract representations from multiple modalities—in their case, MR image sequences (such as DWI, T1-weighted, T2-weighted, FLAIR)—and fuse them in a latent space where arithmetic operations such as the first and second moments of the representations can be calculated. This composite representation can then be deconvolved accordingly. The authors tested the applicability of their model to MR image segmentation (on MSGC [15] and BRATS 2015 [10] datasets). They argue that instead of learning all combinations of functions, each dealing with a specific missing modality, one single model can be learned that deals with all such missing modalities. [3] proposed a generic multi-input, multi-output model, which is an improvement over the Hemis model [6] that is equivalently robust to missing modalities. The model, which is based on correlation networks [2], was proposed to tackle the challenge of learning shared representations from multi-modal images. Correlation networks [2] learn effective correlations among individual modality-specific representations in a coordinated representation space. Imposing correlations as such aid in learning a shared (or coordinated) representation space, especially for MR image modalities that are correlated among one another, in the sense that all the tumorous regions show specific distinctive properties from non-tumorous regions, varying only in their intensity patterns. [3] exploited this fact of MR images, by explicitly imposing correlations among representations extracted from individual modalities through the minimization of the Euclidean distance between modalities. However, it is essential to note that for the problem at hand, the PET and CT modality in PET/CT are not as well correlated as MR image modalities are. While tumorous regions show a distinctive glare from non-tumorous ones in a PET image, it is very much plausible that a similar glare can be observed in non-tumorous regions as well. Therefore, enforcing correlations, as done for MR images, may not be the best approach to learn representations of PET/CT scans.

Further, named methods assume that a complete set of modalities is available during training, which may not be a valid assumption. In particular, it is not always possible to compute correlations with incomplete PET/CT data, meaning that a CT scan is available, but no corresponding PET scan. In contrast, the proposed method treats PET representations as an *optional* context vector that is fused with the CT representations through an attention mechanism, which has the capability to amplify the signal in salient and discriminatory regions.

3 Methods

3.1 Objective

Let X^{CT} and X^{PET} represent the domain of CT and PET images respectively. Likewise, let Y represent the domain of segmented tumorous regions. Given is

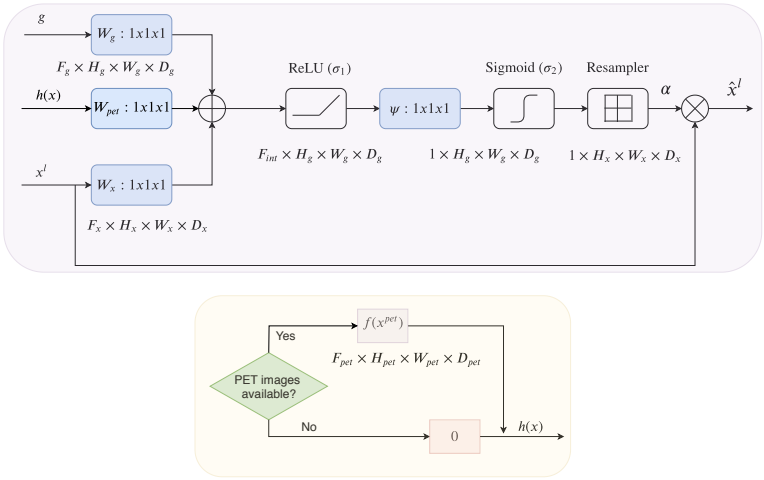

Fig. 1. Schematic of the proposed PET-guided attention gate. The input feature representation x^l is scaled by attention mask α, which is computed by the attention gate. Spatial and contextual information are captured by two gating signals: the encoded feature representation g and the composite function $h(x)$. The composite function is zero when the PET images are missing and the output of the function $f(x^{PET})$ when PET images are available. PET image representation and not the PET image itself is fed to the attention gate.

a dataset consisting of N PET/CT images $\{x_i^{CT}, x_i^{PET}\}_{i=1}^{N}$ and M CT images $\{x_i^{CT}\}_{i=1}^{M}$ ($x_i^{CT} \in X^{CT}$ and $x_i^{PET} \in X^{PET}$) for which corresponding PET images are missing. A typical scenario would be $0 \leq N < M$. For every CT image we have a segmentation mask of tumorous regions i.e. $\{x_j^{CT}, y_j\}_{j=1}^{M}$ where $x_j^{CT} \in X^{CT}$ and $y_j \in Y$.

We are interested in a composite function $H : (X^{CT}, s \cdot X^{PET}) \rightarrow Y$ where s equals 1 if PET images are available and 0 otherwise. The composite function H encompasses two functions: $F : (X^{CT}, X^{PET}) \rightarrow Y$ and $G : X^{CT} \rightarrow Y$. $\hat{y} = H(x^{CT}, s \cdot x^{PET})$ gives the probability map of tumorous regions. The proposed PAG model models the function $H : (X^{CT}, s \cdot X^{PET}) \rightarrow Y$.

3.2 Attention mechanism

Intuition. The attention gate proposed as part of the current model is built upon the one introduced by [12] for pancreas segmentation on CT images. They based their formulation upon a soft attention mechanism for image classification introduced by [7]. Their attention gate has two inputs: (a) a feature representation and (b) a gating signal. The gating signal filters the input feature representation to select salient regions. The attention gate learns attention masks by attending to parts of the input feature representation. It is enforced by allowing the input feature representation to be compatible with the input gating signal.

[12, 7] used the encoded feature representation (output of the encoder) as the gating signal.

The gating signal provides context for the input feature representation to learn salient attention masks. We propose to use PET image feature representation as an additional input along with the encoded feature representation, as shown in Figure 1. Since PET images can be thought of as a heatmap of tumorous regions in a given CT image, they can help to learn better and discriminatory attention masks by helping the attention masks to focus their attention on regions where PET images show a distinctive glare over their surroundings. Accordingly, the context provided by the encoded feature representation is only enhanced by the input of PET image features whenever they are available. Additionally, this formulation does not mandate the use of PET images features. PET image features can be fed to the model as and when available, making the model flexible to the non-availability of PET images.

Attention gate. Let g represent an encoded feature representation with $H_g \times W_g \times D_g$ spatial resolution and F_g filter channels respectively for an input CT image x^{CT}. Similarly let x^l represent a feature representation at an intermediate spatial resolution of the encoder branch (skip connection) with $H_x \times W_x \times D_x$ spatial resolution and F_l filter channels respectively for the same input CT image x^{CT}. Likewise, let x^{PET} be an input PET image corresponding to the input CT image x^{CT}.

The attention gate learns attention coefficients $\alpha_i^l \in [0, 1]$ for layer l and voxel position i that identify discriminatory image regions and discard those feature responses to preserve activations that are specific to the appropriate task at hand. The output of the attention gate is an element wise multiplication of the feature representation $x_i^l \in \mathbb{R}^{F_l}$ and attention coefficients α_i^l to obtain the filtered output $\hat{x}_i^l = x_i^l \odot \alpha_i^l$, where $\odot$ denotes the element-wise multiplication. We consider a single attention coefficient for the multi-dimensional vector x_i^l at voxel position i. Also note that $g_i \in \mathbb{R}^{F_g}$ for voxel position i.

Let $\theta_x \in \mathbb{R}^{F_l \times F_{int}}$ and $\theta_g \in \mathbb{R}^{F_g \times F_{int}}$ be linear transformations that are applied to the intermediate feature representation x^l and the encoded feature representation g respectively. Let $f(x^{PET})$ be a function that extracts PET image specific features before they are applied to the attention gate. Define a composite function

$$h(x) = \begin{cases} f(x^{PET}) \text{ when PET images are available} \\ 0 \text{ when PET images are unavailable} \end{cases} \tag{1}$$

Then the attention coefficients are given by

$$q_{att}^l = \psi^T(\sigma_1(\theta_x^T x_i^l + \theta_g^T g_i^l + h(x) + b_g)) + b_\psi \tag{2}$$

$$\alpha_i^l = \sigma_2(q_{att}^l(x_i^l, g_i; \Theta_{att})) \tag{3}$$

where $\sigma_1(x)$ and $\sigma_2(x)$ are ReLU and sigmoid activations respectively. The parameters of attention gate Θ_{att} are given by $\theta_x \in \mathbb{R}^{F_l \times F_{int}}$, $\theta_g \in \mathbb{R}^{F_g \times F_{int}}$, $\psi \in \mathbb{R}^{F_{int} \times 1}$ and bias terms $b_\psi \in \mathbb{R}$, $b_g \in \mathbb{R}^{F_{int}}$. The linear transformations are computed using channel-wise $1 \times 1 \times 1$ convolutions for the input tensors.

The composite function. The composite function defined by Equation 1 represents scenarios when the PET images are either available or missing. In the absence of PET images, the function takes on a value of zero, which boils down to having a simple attention gate akin to the one proposed by [12] on top of the encoder-decoder architecture. However, in the presence of PET images, the function is identical to the PET image feature extractor $f(x^{PET})$. Instead of passing the PET images directly as an input to the attention gate, we pass a higher dimensional feature representation extracted by the function $f(x^{PET})$, which supposedly encompasses a richer spatial and contextual information than the PET images themselves. This function could be any function approximator such as a neural network. A key insight of the proposed model is that, in contrast to previous modality fusion architectures [3, 6], there is no fusion of the respective modality-specific embeddings. Such a fusion of embeddings from different modalities can skew the intended embedding space while training the respective models with missing modalities such as in our case. Since there is no fusion of PET and CT embeddings in the proposed PAG model, we do not run the risk of learning skewed embeddings while training the model with a combination of PET/CT and CT images.

The model architecture is an encoder-decoder architecture similar to a U-Net architecture with three skip connections. The three skip connections are filtered through their respective attention gates, with each attention gate having its own set of parameters. More details about the model architecture can be found in the supplementary section.

4 Experiments

We consider four baselines to validate our approach. To make a fair comparison, the PAG model and all baselines use the same backbone architecture [11]. Unimodal and bimodal models process CT-only and PET/CT data respectively. The only difference between unimodal and bimodal models is that PET images are input to bimodal model as an additional channel along with CT images. On the other hand, unimodal+attn and bimodal+attn models are unimodal and bimodal models with the addition of a simple attention gate [12]. Similar to the unimodal and bimodal models, unimodal+attn and bimodal+attn models process CT-only and PET/CT images respectively. Unlike the two discrete unimodal and bimodal models (or unimodal+attn and bimodal+attn models), the PAG model is a single model that handles both unimodal and bimodal scenarios. PAG:ct denotes the PAG model with CT-only inputs during inference, whereas PAG:ct+pet denotes the model with PET/CT inputs respectively.

4.1 Ablation framework

The ablation study underscores the contribution of the proposed PAG model in contrast to the conventional models. To make things simpler to follow, consider a hypothetical scenario where we have 80 CT and 20 PET/CT scans respectively. So all in all, we have 100 CT scans, of which 20 CT scans have a corresponding PET series. While it is possible to train a unimodal model with 100 CT scans, a bimodal model can only be trained using 20 PET/CT scans. The rest of the 80 CT scans can not be used. On the other hand, since the PAG model is flexible to the availability of PET scans, it is possible to train the model on the 100 CT scans, including the 20 PET scans. Accordingly, the ablation study is designed to examine the performance of the model in scenarios such as these. Concretely, through this ablation study, we examine the performance of the baseline bimodal model and the proposed PAG model as the fraction of the total number of PET series that are made available for training is gradually reduced. With the decrease in the number of PET scans as such, the number of CT scans that can be used for training bimodal model also decreases. However, the PAG model can leverage upon the complete set of CT scans in conjunction with the restricted number of PET scans. It is important to note that since we keep the number of CT scans fixed, the corresponding number of annotated scans (ground truth segmentation masks) is also fixed.

In other words define the ratio $r = n_{pet}/N_{pet}$ where n_{pet} are the number of PET scans available for training and N_{pet} the total number of PET scans in the given dataset. We then decrease the ratio r gradually from 1 to 0 (N_{pet} is fixed). It is expected that with decreasing ratio r, the performance of the bimodal model decreases noticeably. However, we expect the decrease in the performance of the novel PAG model to be less pronounced. At all times, even in the limit of zero PET scans, it should perform at least as good as a unimodal model that is trained on the complete set of CT scans. In the following, we provide details about the dataset and implementation details for the experiments, before we continue with the presentation and discussion of the results.

Evaluation data. We evaluate our approach on a dataset of 397 PET/CT scans of patients suffering from lung cancer, collected and labeled by the radiology department of the University of Basel, Switzerland. PET/CT images provide complementary information on the regions of interest compared to CT-only data. PET images can be thought of as a heatmap for the corresponding CT images where the tumorous regions show a marked contrast or a distinctive glare between their surroundings. An example of such a pair of CT images and a PET/CT image (PET image superimposed on CT image) is shown in Figure 2. Note that the tumorous region, which is bound by a red bounding box in the CT image, has a marked contrast over its surroundings in the PET/CT image. This is because of the greater 18F-FDG uptake by the malignant tumors due to higher metabolic activity, which can be detected from PET images.

The dataset contains a rich diversity of primary tumors, lymph node metastases, and other metastases that were independently segmented by two expert

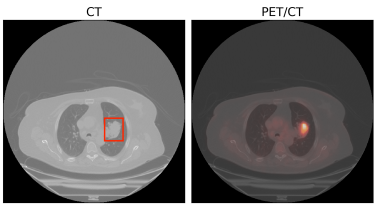

Fig. 2. (Left) An example of a malignant tumor in the right lung. The tumor is surrounded by the bounding box in red. (Right) PET/CT image for the same region. There is a distinctive glare in the region for the corresponding tumorous region.

radiologists. Therefore, the dataset provides a rich data source that is an order of magnitude larger than existing public PET/CT datasets with labelled segmentation maps. More details about the dataset are provided in the supplementary information.

Evaluation criteria. We use dice coefficient as our metric to evaluate the proposed model. Dice coefficient is one of the most widely used metric to evaluate segmentation algorithms. It measures the degree of overlap between the ground truth and predicted segmentation masks factored by the number of true positives and false positives. It falls within a range of [0,1] with 0 signifying absolutely no intersection between the two sets while 1 signifying a perfect intersection with no false positives or false negatives, meaning both sets are alike. A correctly predicted segmentation mask has a dice coefficient of 1, whereas a segmentation mask that predicts zeros for all the voxels has a dice coefficient of 0. Therefore we would expect the dice coefficient of a segmentation algorithm to lie in the range of $[0, 1]$ and the higher the dice coefficient, the closer is the predicted segmentation mask to the ground truth segmentation mask.

Training details. We developed all our models using the PyTorch framework.[1] Each of the models occupies approximately 12GB of GPU memory for model parameters, forward and backward pass. So with a batch size of 2, the memory requirement is approximately doubled i.e., 24GB. All models were trained on a server of 8 NVIDIA Tesla V-100-SXM2 32 GB GPUs. We chose a weighted combination of Sorenson-Dice loss and binary cross-entropy loss as our loss function, a default choice for segmentation tasks. All the models were trained using Adam optimizer (default parameters) and group normalization [16]. Initially, the learning rate α was set to 0.0001; the learning rate was then gradually decayed after every training epoch. The model parameters were regularised using L2 regularisation with regularisation parameter β set to 10^{-5}. Augmented data was included for training at every training epoch but with a probability $p_{data-aug} = 0.25$. All models were trained for 75 epochs.

[1] https://github.com/pvk95/PAG

While training the PAG model, it is critical that PET images are randomly excluded at every training step with a non zero probability p. We do this to ensure that the PAG model does not overfit to either of the scenarios when PET images are available or not. We set this probability value $p = 0.5$. (See appendix for a list of hyper-parameters).

From the dataset consisting of 397 PET/CT labeled images, 77 PET/CT images and their corresponding labels were randomly selected and set apart as our test dataset. The remaining 320 samples were used for training and validation. All the baseline and PAG models have been evaluated using four-fold cross-validation experiments. The training and validation dataset is randomly split into four folds. One of the folds was kept out for validation. The remaining three folds were used for training the models. Each of the models was then retrained on the entire 320 PET/CT images before testing the models on the test dataset. When constraining the number of PET images in the ablation study, we randomly sampled the appropriate number of PET images from the samples that were initially earmarked for training and then trained accordingly. It is noteworthy that the respective validation folds across all the models and all the ratios r in the ablation study remain the same. More details about the training are provided in the supplementary information.

4.2 How well does the model incorporate the two scenarios: CT only images and PET/CT images?

Figure 3 shows the performance of the individual baseline models and the PAG model when a PET image is available for every CT image while training the models. We thus do not place any restriction on the availability of PET images. We do this primarily to validate whether our model is able to handle the combination of CT and PET/CT images well. We observe that the PAG:ct+pet model performs on par with bimodal and bimodal+attn models. Similarly, PAG:ct performs on par with unimodal and unimodal+attn models.

We incorporated the attention mechanism of [12] to the unimodal and bimodal models, and denote the resulting models with unimodal+attn and bimodal+attn. We expected them to outperform their non-attention counterparts (i.e., the unimodal and bimodal models). However, this is not the case, considering Figure 3. The reason for this behaviour could be the complexity of our dataset. The attention gate [12] of the unimodal+attn and bimodal+attn models was originally tested on two publicly available datasets for pancreas image segmentation. The pancreas has a definite shape, structure, and morphology. They are found in a single location within the body. However, the tumors of the current dataset exhibit varying shapes, structures, morphologies, and even locations within the body. This could explain why we do not observe a significant performance gain on our dataset, by adding their attention gate to the unimodal and bimodal models. However, this does not imply that the attention mechanism is not at play here, but that the attention masks are not informative enough. However, it becomes clear from Figure 3 that accommodating PET images as part of the proposed attention gate significantly improves performance

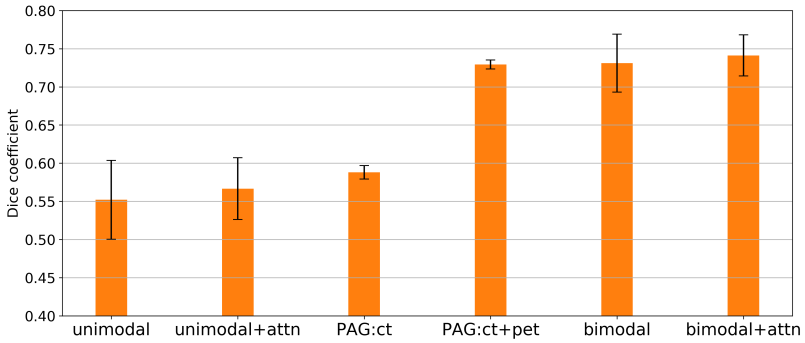

Fig. 3. The Figure shows the performance of the baseline models and the PAG model on the test data set. PAG:ct and PAG:ct+pet are both based on the PAG model. PAG:ct+pet is PAG model when PET images are input to the model in addition to the CT images. Conversely, PAG:ct is PAG model when only CT images are input to the model, but no PET images are used as input to the model. PAG:ct performs at par with the unimodal and unimodal+attn models. Similarly PAG:ct+pet model performs at par with the bimodal and bimodal+attn models.

of the models, considering the better performance of PAG:ct+pet model (dice coefficient=0.73) over PAG:ct model (dice coefficient=0.58). This is not just a consequence of the addition of PET images to the PAG:ct+pet model but because of the addition of PET images to the PAG:ct+pet model in association with the proposed attention gate, the very means of how PET images are fed to the model.

Consequently, we conclude that when PET images are available during inference, PAG:ct+pet performs on par with bimodal and bimodal+attn models, and when they are not available, PAG:ct performs on par with unimodal and unimodal+attn models. This supports the claim that the PAG model successfully encompasses the two discrete models: unimodal and bimodal models. Further, the addition of PET images through the proposed attention gate makes a significant impact on the performance of the PAG model. This validates that the attention gate effectively integrates information from PET images, whenever they are available.

4.3 How well does the model handle a combination of CT and PET/CT images?

Figure 4 shows the result of the ablation study, as described earlier in section 4.1. The performance of the PAG:ct+pet model and the bimodal models is evaluated as the ratio $r = n_{pet}/N_{pet}$ is gradually reduced. The ratio points considered are [1.0, 0.5, 0.3, 0.15, 0.1, 0.05, 0.03]. The majority of examined data points are close to zero, in order to compare and contrast the significance of the PAG model when PET images are very scarce. The unimodal model is illustrated

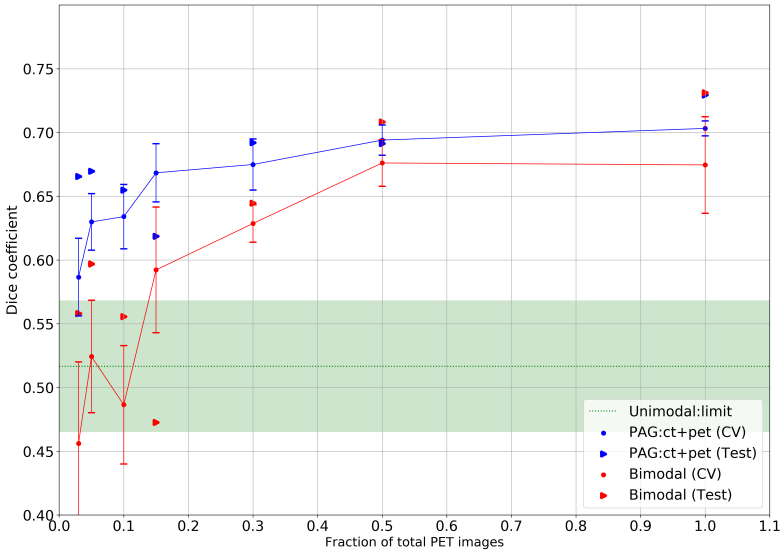

Fig. 4. The Figure shows the dice coefficient for the PAG model and the bimodal model when the fraction of total PET images that are made available for training the models is restricted. Results are shown for the validation (CV) and test (Test) data sets. The green band is the mean and standard deviation of the unimodal model trained on CT images. The degradation in performance of the bimodal model is much more drastic than the PAG:ct+pet model. Note that PAG:ct+pet model always maintains the edge over unimodal model because either of the models were trained on the same number of CT images, with additional PET images for PAG:ct+pet model.

as well with its mean (green dotted line) and standard deviation (green band around the dotted line). It can be seen that for all the values of ratio r, the dice coefficient of PAG:ct+pet is greater than the bimodal model. Consider, for instance a point at $r = 0.15$. This point represents a scenario where one has 36 PET/CT images and 204 CT-only images or 240 CT images in total. The bimodal model was trained on the small set of 36 PET/CT images while the PAG model was trained on 204 CT images and 36 PET/CT images. This shows that the extra 204 CT-only images which would otherwise have been discarded while training the bimodal model could be used for training the PAG model. Clearly, the extra 204 CT-only images make a difference in boosting the dice coefficient of the model. This performance gain becomes more and more extreme as the ratio r approaches values closer to zero.

There is another facet to the PAG:ct+pet model. Irrespective of the ratio r, PAG:ct+pet was trained on the same number of CT images. This implies that even in the limit of zero PET images, the performance of PAG:ct+pet should not degrade below the performance of unimodal model which can be clearly observed for points closer to zero ([0.03, 0.05, 0.1]). For example, consider a point $r = 0.03$. This point represents a data set with 7 PET/CT images and 233

CT-only images or 240 CT images in total. The unimodal model was trained on 240 CT images while the PAG model was trained on 240 CT images including 7 PET images. Clearly, the extra number of 7 PET images yielded in significant performance gains (dice coefficient = 0.66) over the unimodal model (dice coefficient = 0.56). Naturally, the improvement in performance becomes more and more obvious with increasing ratios r.

Hence, in the limit of zero PET images, the PAG model is able to successfully leverage upon the extra number of CT-only images. This behaviour is reflected in the higher dice coefficient of PAG:ct+pet model over the bimodal model. In the scenario when the PAG model is trained with CT-only images, the performance boundary would be the unimodal model. Consequently, just with the addition of a few PET images to the PAG model, we observe significant performance gains, considering higher dice coefficient of PAG:ct+pet model over unimodal model. This supports our claim that the model makes efficient use of the combination of CT-only and PET/CT data.

5 Discussion and Conclusion

Although PET/CT imaging is the gold standard for the staging of lung cancer, due to logistic and economic constraints, PET images are often unavailable. This problem is especially prominent in resource-constrained healthcare systems. While conventional methods are unable to handle a combination of CT-only and PET/CT data, we tackled this challenge by adapting an established visual soft attention mechanism to the problem at hand. We demonstrated that our proposed approach performs on par with unimodal and bimodal baselines. We further present that our model is especially useful when the number of PET images is small in comparison to the number of CT images, which is relevant in resource-constrained environments.

It is noteworthy, irrespective of the number of PET/CT images that are available, the model always requires the same number of segmentation masks as the number of total number of CT images. This could be a limitation considering the manual effort in procuring the segmentation masks. In future work, we would like to explore the possibility of reducing the number of segmentation masks by generative models. Thereby, we could extend the resource efficiency of the algorithm to leverage a reduced number of segmented images.

Another interesting direction for future research would be to extend the proposed PAG model to other imaging modalities such as MRIs, as our formulation is not limited to a single additional modality. It would be interesting to investigate further the behaviour of the proposed attention gate with additional modalities.

Acknowledgements

ID is supported by the SNSF grant #200021_188466.

References

1. Cai, L., Wang, Z., Gao, H., Shen, D., Ji, S.: Deep adversarial learning for multi-modality missing data completion. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1158–1166 (2018)
2. Chandar, S., Khapra, M.M., Larochelle, H., Ravindran, B.: Correlational neural networks. Neural computation **28**(2), 257–285 (2016)
3. Chartsias, A., Joyce, T., Giuffrida, M.V., Tsaftaris, S.A.: Multimodal mr synthesis via modality-invariant latent representation. IEEE transactions on medical imaging **37**(3), 803–814 (2017)
4. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019)
5. Guo, Z., Li, X., Huang, H., Guo, N., Li, Q.: Medical image segmentation based on multi-modal convolutional neural network: study on image fusion schemes. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 903–907. IEEE (2018)
6. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 469–477. Springer (2016)
7. Jetley, S., Lord, N.A., Lee, N., Torr, P.H.: Learn to pay attention. arXiv pp. arXiv–1804 (2018)
8. Li, L., Zhao, X., Lu, W., Tan, S.: Deep learning for variational multimodality tumor segmentation in pet/ct. Neurocomputing (2019)
9. Lian, C., Ruan, S., Denoeux, T., Li, H., Vera, P.: Joint tumor segmentation in pet-ct images using co-clustering and fusion based on belief functions. IEEE Transactions on Image Processing **28**(2), 755–766 (2018)
10. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
11. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018)
12. Oktay, O., Schlemper, J., Le Folgoc, L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas (2018)
13. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
14. Society, A.C.: About lung cancer (March 2020), https://www.cancer.org/cancer/lung-cancer/about.html
15. Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S.: 3d segmentation in the clinic: A grand challenge ii: Ms lesion segmentation. Midas Journal **2008**, 1–6 (2008)
16. Wu, Y., He, K.: Group normalization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 3–19 (2018)
17. Zhao, L., Lu, Z., Jiang, J., Zhou, Y., Wu, Y., Feng, Q.: Automatic nasopharyngeal carcinoma segmentation using fully convolutional networks with auxiliary paths on dual-modality pet-ct images. Journal of digital imaging **32**(3), 462–470 (2019)

18. Zhong, Z., Kim, Y., Zhou, L., Plichta, K., Allen, B., Buatti, J., Wu, X.: 3d fully convolutional networks for co-segmentation of tumors on pet-ct images. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 228–231. IEEE (2018)
19. Zhou, T., Canu, S., Vera, P., Ruan, S.: Brain tumor segmentation with missing modalities via latent multi-source correlation representation. arXiv pp. arXiv–2003 (2020)