AutoPET3 Challenge. When do we need models that generalize and a mixture of experts who specialize?

Maxim Shatskiy

shatskiy.maxim@gmail.com

Abstract. This document describes solution to AutoPET3 Challenge. We show how an ensemble of Unet++ models with EfficientNet-B7 backbones trained separately on FDG and PSMA data can perform well in this competition. Can a single model beat two specialized models? We see what results of this competition will bring.

Keywords: PET/CT · Unet++ · EfficientNet-B7 · FDG, PSMA · generalization · specialization · multitracer

1 Introduction

These are notes for submission for AutoPET3 challenge. Refer to the competition website `https://autopet-iii.grand-challenge.org/`.

This was the first AutoPET competition for us and when working on this competition we tried to understand:

- Can a single model be trained to perform well on both FDG and PSMA tracers?
- How does the performance of a single model compare against specialized models? Is it more or less difficult to train a single model vs. separate specialized models?
- Even if we can, do we need a single model that can predict for both FDG and PSMA?

1.1 Data

1.2 Data prepossessing

It is often helpful to use a heavy augmentation strategy. However, the idea here was to use minimal, but sufficient preprocessing, which could speed up experimentation. A patch of size (160,160,160) for FDG and (128,128,128) for PSMA was randomly extracted with the probability of including tumor p=1, then rotation with a random angle between -45/45 along a randomly picked axis was applied with p=0.5, then mirroring was done with p=0.5 and finally, normalization was applied as follows. For CT images:

$$img = np.clip(img, -1124, 1124)/1124,$$

and for PET images:

$$img = (img - 7.063827929027176)/(7.960414805306728).$$

1.3 Model

We hypothesize that there is no added advantage in creating a model that generalizes across FDG and PSMA tracers because it makes model training more complicated and does not necessarily improve the quality of prediction upon using two individual models (maybe the results of this competition allow us to reject this hypothesis?). In real life, the type of the tracer will be known before segmentation is done. Therefore, we think it makes sense to make models that generalize across sites, with the same tracer, and across different spacing, with the same tracer, but not across tracers, despite goes against the premise of the competition itself, which states *"... autoPET III expands the scope to address the critical need for models to generalize across multiple tracers and centers."*.

Based on this idea we created a model, which at first classifies if the tracer is PSMA or FDG, since this information is not available during inference time. In a real-life scenario, this model is not required since the type of tracer is known. However, in our case we had to use such a model, let's refer to it as a *router* model. As a *router*, we took a multi-class Unet++, which was trained for one of the attempts to combine PSMA and FDG into a single model. This model predicts 3 classes: 0 - background, 1-FDG pixel, 2- PSMA pixel. Then we count several pixels of a particular class and by majority voting decide whether the current sample belongs to PSMA or FDG. However, with this approach, we encountered problems that in certain cases there is neither FDG nor PSMA pixels detected. This could be indeed a sample without foreground or classification error. If the number of pixels in the foreground is less than 2, then we consider that there are neither FDG nor PSMA results available and set all pixels to 0. Instead of such *router* model that we used, it would be more appropriate to use a 3D classification model, which could distinguish between 4 classes: FDG with tumor, FDG without tumor, PSMA with tumor, PSMA without tumor. However, since we already had a multi-class Unet++ model and under time constraints, we decided to use the existing model and not train a separate classification model. *This is a weak point of our approach, which will likely lead to a lower score on a test set due to the borderline cases described above, which will be misclassified and inappropriate segmentation models will be used.*

The *router* model was trained on the common spacing of FDG tracers, which means PSMA samples were resampled.

After *router* decides the type of the tracer, we give it to corresponding specialized *expert* models. Here for both FDG and PSMA similar ensembles of 2 models were created. Each model in the ensemble is Unet++ with EfficientNet-B7 encoder. One model in each ensemble was trained only on data with tumors, we observed that such a model works better for smaller tumors. The final prediction in each ensemble is done as voting if at least one model predicted foreground/cancer. We tried different ResNet encoders, as well as EfficientNet B8,

but we did not get any significantly better results. Potentially if the time limit on inference was not set, then a bigger ensemble with different encoders could have provided a better overall result. The best parameters were first selected on a single fold. A rigorous cross-validation procedure was not followed to save time, but simply the final model was trained on all data. For the FDG model a Generalized Dice-Focal Loss was used, for PSMA Generalized Dice and CE losses were combined.

1.4 Postprocessing

No postprocessing was used. The time limit on inference did not allow us to experiment with TTA.

1.5 Results

We have not used a cross-validation strategy to rigorously assess performance and choose the best model. After training using fold 0 and choosing configuration, the models were trained using all data as described above and then inference was done for the entire dataset. *Therefore it might be difficult to compare performance with other models. We also observed from the reports from the previous competitions that there is no consistent way of reporting such intermediate results of the competitions. Therefore, for future competitions, it would be good that organizers required to report intermediate results in a particular, unified, way such that they could be compared not only on the final test set, which will not be available after competitions, but also on some publicly available data, for example, to ask all participants to report 5-fold CV results or at least results on a single fold.*

Table 1. Results on entire dataset

Data	dice_score	fp_volume	fn_volume
Summary FDG	0.804	9.088	2.242
Summary PSMA	0.742	11.157	5.143
Summary	0.772	9.855	3.743

Other considerations/observations:

- final results are also sensitive to sw_batch_size and sw_overlap parameters. Due to time and resource constrained, we have not done extensive ablation studies on these or other hyperparameters;
- using classes from totalsegmenter as additional targets. In the previous competitions, it was reported that this approach improves performance. Even though this approach looks reasonable, it is a relatively laborious task to pre-segment images and then map from totalsegmenter classes. Additionally, training takes significantly longer;

- training multi-class models did not give any good results for us;
- training PSMA on resampled spacing. This was not so clear if and to which sampling to resample. Some reviews of approaches in previous competitions and auto-segment packages did not give conclusive results. We found that some of the samples are classified better at certain spacing than others. Probably this is one of the most important aspects of prepossessing pipeline and we have not found a definitive answer to it. Automatic segmentation (autoseg3d and nnUnet) packages come up with different suggestions on how to make this resampling. Therefore we decided to stick with the simplest solution keeping the original resolution. Likely ensemble of models trained on different spacing could provide better results at the expense of training and inference time.

1.6 Some thoughts on competitions in medical domain

Competitions in the medical domain were pretty isolated from the overall community, with some exceptions and in my opinion, were lagging on the methods. Most of the competitions are published on Grand Challenge website and only recently there are some competitions on probably the most popular competition's website kaggle.com. One can observe a difference, not a single competition on kaggle.com was won using nnUnet, whereas many solutions at Grand Challenge use it, whereas people at kaggle.com use other packages and solutions not focusing on Monai or nnUnet. Also, kaggle.com bolsters sharing ideas during and after the competition, helping people to start in new domains and test their ideas faster. Overall, I have an impression that medical competition would benefit by moving from isolated platforms and limited community praying to nnUnet to a broader community.

1.7 Conclusion and future work

In conclusion, we showed that with minimal augmentation and without generalization over FDG and PSMA tracers it is possible to create a competitive model. We hope that the results of this competition and our contribution to it give some answers to the questions stated in the introduction. Additionally, recently it was shown that foundational models, like DINOv2 have good performance in the medical domain with relatively small adapters, therefore one could try to create adapters for fine-tuning foundational models to FDG and PSMA data and see how data efficient these solutions are against traditional, mostly Unet based solutions. Does the adapter perform better or worse if we finetune it with both FDG and PSMA?

1.8 Github Link

Thanks to the organizers for creating a great data-centric starting package based on PyTorch Lightning and Monai. We think this approach to benchmarking

and developing the models is better in many aspects than other not new (nn), but packaged solutions, that gained popularity in the medical domain. Link to inference: `https://github.com/maxshatskiy/autopet3_inf`