Project Report (May 30, 2021)

Ankita Ghosh[1] **and Sahil Khose**[2]

[1]Research Assistant, ghoshankita0907@gmail.com , CSE, MIT Manipal
[2]Research Assistant, sahilkhose18@gmail.com, ICT, MIT Manipal

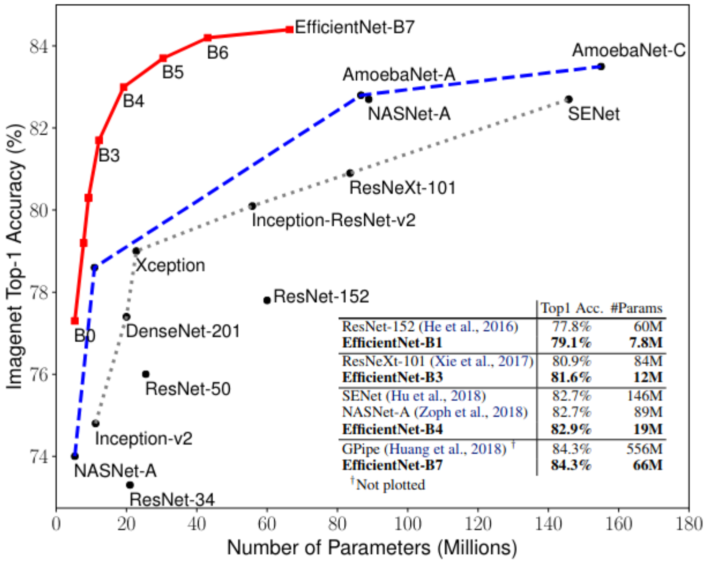
ABSTRACT

We discuss two very interesting histopathology deep learning papers today: 1. Deep learning based six-type classifier for lung cancer and mimics from histopathological whole slide images: a retrospective study *(Huan Yang et al)* and 2. Deep neural network models for computational histopathology: A survey *(Chetan L Srinidhi)*. We also mention the challenges we faced to retrieve the dataset and the steps we are planning to take to move forward.

Six-type cancer classifier

They developed the first deep learning-based six-type classifier for histopathological WSI classification of lung adenocarcinoma, lung squamous cell carcinoma, small cell lung carcinoma, pulmonary tuberculosis, organizing pneumonia and normal lung.

- They used EfficientNet-B5 based model and compared it with ResNet-50 architecture.
- EfficientNet-B5 seemed to outperform ResNet-50 on almost all of the tasks especially when the slides were collected from multiple sources which requires learning abstract features which help in distinguishing the different classes.



	Top1 Acc.	#Params
ResNet-152 (He et al., 2016)	77.8%	60M
EfficientNet-B1	**79.1%**	**7.8M**
ResNeXt-101 (Xie et al., 2017)	80.9%	84M
EfficientNet-B3	**81.6%**	**12M**
SENet (Hu et al., 2018)	82.7%	146M
NASNet-A (Zoph et al., 2018)	82.7%	89M
EfficientNet-B4	**82.9%**	**19M**
GPipe (Huang et al., 2018) †	84.3%	556M
EfficientNet-B7	**84.3%**	**66M**

†Not plotted

Figure 1. Source: EfficientNet(2020)

- As we can observe that with same number of parameters as ResNet-50 we get 6% better performance on the ImageNet dataset with the EfficientNet-B3 model and with more parameters we can get upto 8% more accuracy than the ResNet-50 architecture.
- Using the EfficientNet model group was our initial idea to deal with the dataset as mentioned in the previous technical reports. The publication of this paper a few months ago reinforces on the idea that we can improve on the domain with bigger and better models.

- The authors had access to only EfficientNet-B5 model implementation and model weights when the paper was implemented. Now we have access to B6-B8 models along with better architectures like NFNets, ViT, and released 2 weeks ago the EfficientNet V2 which outperforms all of the models.

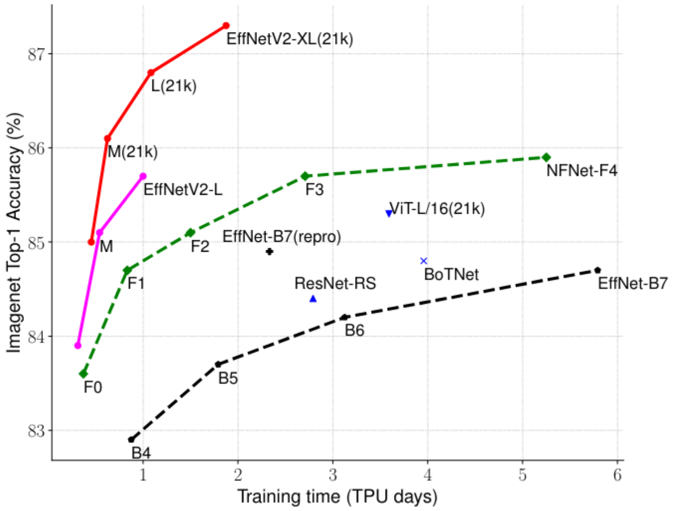
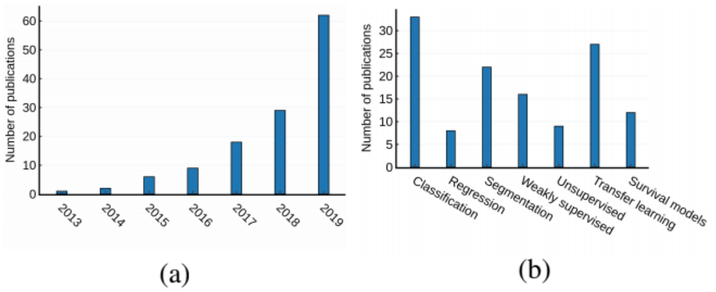


Figure 2. Source: EfficientNetV2(2021)

Survey on computational histopathology

This is a report which summarizes 130 histopathology papers.

- They observe an exponentially growing trend for the number of papers published in this area.



(a)

(b)

Fig. 1: (a) An overview of numbers of papers published from January 2013 to December 2019 in deep learning based computation histopathology surveyed in this paper. (b) A categorical breakdown of the number of papers published in each learning schemas.

Figure 3. Source: Srinidhi et al(2019)

- They compared the cancer types, staining, application, method and dataset of the abundance of papers they have collected.

- They discuss the field's progress based on the methodological aspect of different machine learning strategies such as supervised, unsupervised, transfer learning and various other sub-variants of these methods.

- They also provide an overview of deep learning based survival models that are applicable for disease-specific prognosis tasks.

- In most applications, standard architectures like VGGNet, InceptionNet, ResNet, MobileNet, DenseNet can be directly employed, and custom networks should only be used if it is impossible to transform the inputs into a suitable format for the given architecture, as the transformation may cause significant information loss that may affect the task performance.
- They recommend to use larger convolutional filters if the input size is large, skip connections in segmentation tasks, and batch normalization for faster convergence and to obtain better performance.
- Given the dataset with few images, training only the last decision layers while freezing the initial layers, using a non-linear decision layer, using regularization techniques like weight decay and dropout are recommended to avoid overfitting.

Challenges while retrieving the dataset

We followed the steps mentioned by Coudray in his code implementation to access the dataset. We installed the Data Transfer Tool Client and User Interface and queued the manifest file of the data after downloading from the GCD portal. We tried to download TCGA-LUSC and TCGA-LUAD.

Problems faced by us:

- The dataset is approximately of size 60GB. Despite keeping it for download for multiple hours and multiple times, we could barely download 5-10% of the dataset.
- For storing 60GB dataset, we tried external drive and set the path accordingly. However, during download it was redirecting to the home directory and saving huge amount of data on laptop is infeasible.

Solutions and further actions we have thought have of:

- We have raised an issue on the Coudray repository enquiring how to handle the dataset but we haven't received any reply yet.
- We are planning to find smaller datasets so that accessing the data and training the model becomes feasible.

Discussion

We have explored the Coudray repository and have a finer understanding of the workflow and the code implementation. Hence, our primary concern at the moment is to search and retrieve a dataset which suits our problem statement. As this problem statement has been approached by implementing some of the best models like EfficientNet, we are looking forward to more innovation in this project alongside implementing the best performing architectures.