Tailoring automated data augmentation to H&E-stained histopathology

Khrystyna Faryna, Jeroen van der Laak, Geert Litjens

Radboudumc
university medical center
Computational Pathology Group | Diagnostic Image Analysis Group

Domain shift in computational pathology

In computational pathology, data acquisition conditions vary among different labs, leading to variations in corresponding slides. On the Figure 1 you can observe slide patches originating from different labs. Such variation is not a problem for a human expert, however that is not the case for neural networks. CNNs are sensitive to domain shift: a model trained on only images from institution 1 is likely to fail on slides from institution 2 or 3.

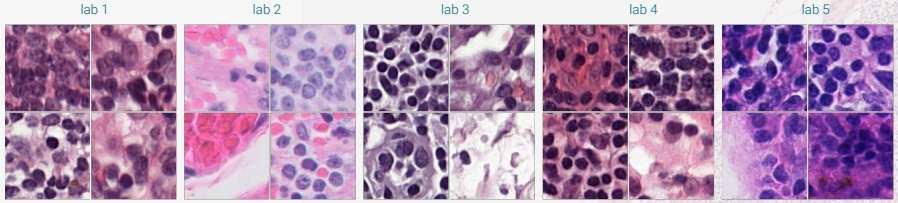


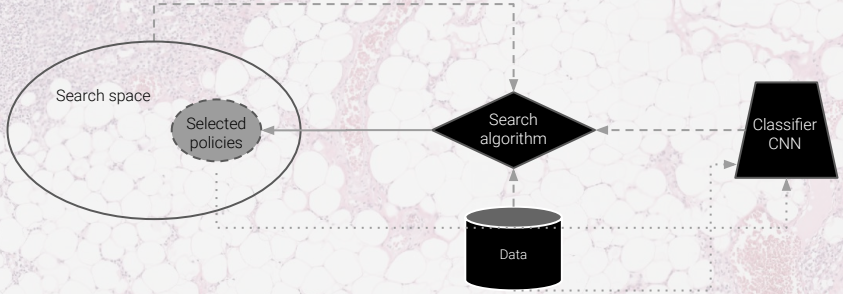
Fig. 1. Patches of slides originating from different labs. Images are taken from Camelyon17[1] dataset.

Improving robustness: data augmentation

We want to have models that generalize across domains. Data augmentation is one way to make CNNs robust to varying forms of domain shift. However, manual tuning of augmentation hyperparameters is extremely time consuming because of the large search space. Taking into account that normally you have a set of transforms, each of which has a set of magnitudes, for each of which there exists a set of probabilities, the search space can reach 10^{32}. At this point manual tuning becomes infeasible. In practice, researchers often do not perform augmentation parameter tuning completely or simply select transforms intuitively, which often can result in suboptimal performance.

Automated data augmentation: RandAugment

Recently, a number of automated augmentation methods appeared. These methods facilitate augmentation policy selection by either significantly reducing search space or applying efficient search algorithms. In this study, we utilize an automated augmentation method and adjust it to H&E stained histopathology. In particular, we opt for RandAugment, as it is computationally efficient and requires minimal modifications to the existing training pipeline.



In RandAugment the search space is reduced to only 2 hyperparameters: M and N. M is a single magnitude for all transformations. N, number of sequentially applied transforms. The transforms are always selected with uniform probability. Finally the optimal values of N and M are found through grid search.

Khrystyna Faryna
khrystyna.faryna@gmail.com
KhrystynaFaryna
kfaryna

Tailoring automated data augmentation to H&E-stained histopathology

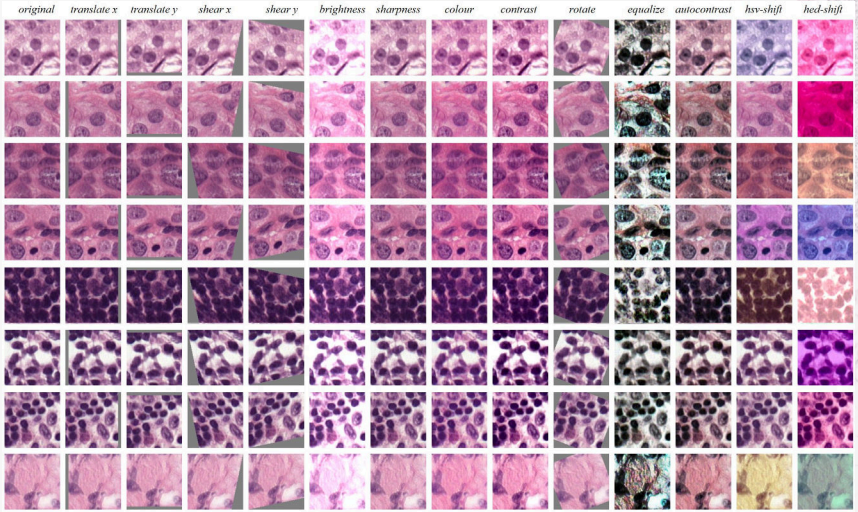
Khrystyna Faryna, Jeroen van der Laak, Geert Litjens

Radboudumc
university medical center
Computational Pathology Group | Diagnostic Image Analysis Group

H&E tailored RandAugment

In order to make this method beneficial to H&E stained histopathology we introduce several modifications. Firstly, instead of searching for an optimal constant value, we optimize for an upper bound of a uniform random value of a magnitude. Secondly we take out the irrelevant transforms such as invert, solarize and posterize. Thirdly we supply the list of transforms with histopathology relevant: hsv and hed transforms.

$$M \sim U(0, m_{opt}), \quad m_{opt} = \max_{\{m,n\}}(AUC_{val})$$



Experimental setup

DATA

In this study we use Camelyon17 challenge dataset for metastasis detection in breast lymph nodes. In this work, we assume that a model trained on data from institution a validated on centers b,c is capable of generalizing to data from other unseen institutions. Thus, we arrange the experiments in the following way: we always train the model only on data from rumc , the validation set contains a subset of rumc and two external institutions, while the test set consists of data from the remaining two centers. The datasets in validation and testing are subsequently permuted resulting in six possible unique combinations.

split #	1	2	3	4	5	6
validation	*lpe-umcu*	*rh-umcu*	*cwh-umcu*	*lpe-rh*	*rh-cwh*	*lpe-cwh*
test	*rh-cwh*	*lpe-cwh*	*lpe-rh*	*cwh-umcu*	*lpe-umcu*	*rh-umcu*

BASELINE

We compare against the current state-of-the-art for data augmentation in computational pathology: the extensive review of various augmentation strategies by TELLEZ et al., where the authors analyze the performance impact of both classical and domain-specific augmentation. We adopt a training pipeline and other hyperparameters from Tellez et. al, for fair comparison. We evaluate performance using AUC.

Results & Discussion

no augmentation				H&E tailored randaugment				Tellez et al., 2019			
cwh	rh	lpe	umcu	cwh	rh	lpe	umcu	cwh	rh	lpe	umcu
0.740±...	0.255±...	0.697±...	0.108±...	0.967±0.002	0.954±0.005	0.951±0.002	0.982±0.003	0.953±...	0.970±...	0.943±...	0.965±...

H&E tailored randaugment outperformed the manual baseline on ¾ sets. Automated frameworks offer a more structured and methodological approach to data augmentation. It is beneficial to combine the automated augmentation frameworks with domain-specific knowledge (HED, HSV shifts for H&E histopathology). Optimizing for an upper bound of a uniform random rather then for a constant value of magnitude might be beneficial when utilizing randaugment in medical imaging tasks.We present a method for a fast automatic selection of optimal data augmentation for H&E stained histopathology.

Khrystyna Faryna
khrystyna.faryna@gmail.com
KhrystynaFaryna
kfaryna