MULTI-TASK LEARNING IN HISTO-PATHOLOGY FOR WIDELY GENERALIZABLE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work we show preliminary results of deep multi-task learning in the area of computational pathology. We combine 11 tasks ranging from patch-wise oral cancer classification, one of the most prevalent cancer in the developing world, to multi-tissue nuclei instance segmentation and classification.

1 INTRODUCTION

The emerging area of computational pathology (CPath) is ripe ground for the application of deep learning methods to healthcare due to the sheer volume of raw pixel data in whole-slide images (WSIs) of cancerous tissue slides, generally of the order of $100K \times 80K$ pixels (Colling et al., 2019). However, despite the availability of raw pixel data in CPath, the ground truth for training AI models is sparse, expensive to obtain, and noisy. Diverse multi-centre data are usually limited to the developed world; yet, developing countries are where AI could find their most viable application amid global shortages of clinical experts (Coelho, 2012) and widespread incidence of oral cancer (Shrivastava et al., 2014) and lymphomas (Perry et al., 2016).

A major challenge facing wider adoption of AI, especially in a resource-strapped setting and in the absence of well-curated multi-centric high-quality datasets, is algorithm robustness due to the lack of feature generalisation (Zech et al., 2018). Some recent studies have pointed to limited clinical applicability of AI due to weak experimental design even with datasets obtained in the developed world (Liu et al., 2019; Zech et al., 2018). Additionally, AI research points to several vulnerabilities of deep learning models widely adopted in the aforementioned studies (Geirhos et al., 2018; Ghorbani et al., 2019). As such, particularly for the application within the developing countries, AI models must be robust and learn semantically meaningful features, in order to be able to generalise across a variety of tasks. It is thus reasonable to maximise the utility of available good quality datasets in the literature in order to test the feasibility of obtaining such a model. We propose to test deep multi-task learning (MTL) as a method to obtain general feature representation that would be applicable to new tasks or tasks with relatively small amounts of data (Raghu et al., 2019). We present preliminary results of a simple, yet promising approach to MTL in CPath and has potential to do well particularly on oral cancer, one of the most important cancers of relevance to the developing world.

2 MATERIALS & METHODS

Baxter (2000) theoretically showed that learning from multiple related tasks results in fast learning as measured by the number of training examples required per task; and that inductive biases learned on sufficiently many training tasks will likely generalise to novel tasks. Subramanian et al. (2018) demonstrated successful application of the aforementioned theory to natural language processing via deep MTL. Brenes (2019) demonstrated the potential of MTL in histology to obtain robust features using only two tasks. In this work, we collect 11 well-established tasks ranging from image classification to pixel-wise instance segmentation and classification (see Table 1).

To quantify the performance of segmentation tasks we use panoptic quality (PQ) Kirillov et al. (2019), and classification accuracy for image classification tasks. Baseline results are single-task results as quoted in literature. Epithelium segmentation PQ however was not available, we thus obtained it by training a single task segmentation model. All segmentation decoders were based

Table 1: Pixel-wise segmentation and image classification datasets, their sources, baseline results obtained from the literature and un-tuned MTL approach results.

Name (Source)	Task	Baseline	MTL
	Pixel-level Tasks	PQ	
Awan et al. (2017)	Gland Segmentation	0.76	0.60
Gamper et al. (2019)	Nuclei Inst. Segm. and Class.	0.38	0.18
Janowczyk & Madabhushi (2016)	Epithelium Segmentation	0.68	0.57
Fraz et al. (2019)	Vessel Segmentation	0.67	0.59
	Patch-level Classication Tasks	Accuracy	
Litjens et al. (2018)	Breast (2)	92.44	91.23
Kather et al. (2019)	Colon (9)	98.70	89.00
Alsubaie et al. (2018)	Lung (6)	91.00	78.56
Janowczyk & Madabhushi (2016)	Lymphoma (3)	96.58	58.22
Qureshi et al. (2008)	Meningioma (4)	82.10	92.19
Shaban et al. (2019)	Oral (3)	96.30	91.11
Köbel et al. (2010)	Ovary (5)	89.40	72.56

Algorithm 1: Multi-Task Training Setup

Input: A set of k histology tasks, their corresponding datasets $\mathbb{P}_1, ..., \mathbb{P}_k$ and a set of k task specific decoders $\mathbf{D}_1 ... \mathbf{D}_k$, a feature encoder $\mathbf{E}$ shared across all tasks. Let θ denote model parameters (encoder and decoders), and α be a probability vector $(p_1, ..., p_k)$ denoting the probability of sampling a task at a given iteration such that $\sum_{i=1}^{k} p_i = 1$. Let L and T denote the loss function and a maximum number of iterations, respectively.

Output: Trained encoder $\mathbf{E}$ and decoders $\mathbf{D}_1 ... \mathbf{D}_k$.

repeat

- $t+ = 1$
- Sample a task $i \sim \text{Cat}(k, \alpha)$
- Sample input, output pairs $\mathbf{x}, \mathbf{y} \sim \mathbb{P}_i$
- Encode inputs $h_{\mathbf{x}} \leftarrow \mathbf{E}_\theta(\mathbf{x})$
- Predict $\hat{\mathbf{y}} \leftarrow \mathbf{D}_{i,\theta}(\mathbf{h}_{\mathbf{x}})$
- Update $\theta \leftarrow \text{Adam}(\nabla_\theta L(\mathbf{y}, \hat{\mathbf{y}}))$

until $t = T$;

on Pyramid Scene Parsing network (Zhao et al., 2017), and classification decoders were simply a fully connected layer followed by softmax or sigmoid non-linearity. Each decoder processed 2048 dimensional features extracted using a trainable Resnet-50 encoder (He et al., 2016). Our MTL optimisation approach is described formally in the Algorithm 1 above. In summary, at every optimisation step we randomly pick task index, and use the corresponding task dataset to extract data, process input using the encoder and decode it using task specific decoder, followed by the evaluation of the loss function and gradient update using Adam (Kingma & Ba, 2014).

3 Results & Conclusion

Our preliminary results of deep multi-task training with default hyper-parameters and no-tuning are presented in Table 1 under column MTL. Results are encouraging, particularly given the variance of the loss during training as presented in Figure A1. Loss variance has been attributed to the direction of gradients and has been widely studied in lifelong learning and MTL (Lopez-Paz & Ranzato, 2017; Du et al., 2018; Chaudhry et al., 2018; Yu et al., 2020). Our preliminary results in Figure A3 and Figure A2 demonstrate small cosine distances. However, distribution of distances between gradient vectors becomes narrowly focused around zero as the dimensionality grows. The significance of small distances may increase with growing model size. In the future work we will further investigate MTL optimisation characteristics and match single task performance, as well as test the generalisation of obtained features.

REFERENCES

Najah Alsubaie, Muhammad Shaban, David Snead, Ali Khurram, and Nasir Rajpoot. A multi-resolution deep learning framework for lung adenocarcinoma growth pattern classification. In *Annual Conference on Medical Image Understanding and Analysis*, pp. 3–11. Springer, 2018.

Ruqayya Awan, Korsuk Sirinukunwattana, David Epstein, Samuel Jefferyes, Uvais Qidwai, Zia Aftab, Imaad Mujeeb, David Snead, and Nasir Rajpoot. Glandular morphometrics for objective grading of colorectal adenocarcinoma histology images. *Scientific reports*, 7(1):1–12, 2017.

Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12: 149–198, 2000.

David Brenes. Multi-task deep learning model for improved histopathology prediction from in-vivo microscopy images. *LXAI workshop at NeurIPS*, 2019.

Arslan Chaudhry, Marc'Aurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*, 2018.

Ken Russell Coelho. Challenges of the oral cancer burden in india. *Journal of cancer epidemiology*, 2012, 2012.

Richard Colling, Helen Pitman, Karin Oien, Nasir Rajpoot, Philip Macklin, David Snead, Tony Sackville, Clare Verrill, CM-Path AI in Histopathology Working Group, Velicia Bachtiar, et al. Artificial intelligence in digital pathology: A roadmap to routine use in clinical practice. *The Journal of pathology*, 2019.

Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

MM Fraz, SA Khurram, S Graham, M Shaban, M Hassan, A Loya, and NM Rajpoot. Fabnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. *Neural Computing and Applications*, pp. 1–14, 2019.

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. Pannuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In *European Congress on Digital Pathology*, pp. 11–19. Springer, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7, 2016.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1), 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9404–9413, 2019.

Martin Köbel, Steve E Kalloger, Patricia M Baker, Carol A Ewanowich, Jocelyne Arseneau, Viktor Zherebitskiy, Soran Abdulkarim, Samuel Leung, Máire A Duggan, Dan Fontaine, et al. Diagnosis of ovarian carcinoma cell type is highly reproducible: a transcanadian study. *The American journal of surgical pathology*, 34(7):984–993, 2010.

Geert Litjens, Peter Bandi, Babak Ehteshami Bejnordi, Oscar Geessink, Maschenka Balkenhol, Peter Bult, Altuna Halilovic, Meyke Hermsen, Rob van de Loo, Rob Vogels, et al. 1399 h&e-stained sentinel lymph node sections of breast cancer patients: the camelyon dataset. *GigaScience*, 7(6):giy065, 2018.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.

Anamarija M Perry, Jacques Diebold, Bharat N Nathwani, Kenneth A MacLennan, Hans K Müller-Hermelink, Martin Bast, Eugene Boilesen, James O Armitage, and Dennis D Weisenburger. Non-hodgkin lymphoma in the developing world: review of 4539 cases from the international non-hodgkin lymphoma classification project. *Haematologica*, 101(10):1244–1250, 2016.

Hammad Qureshi, Olcay Sertel, Nasir Rajpoot, Roland Wilson, and Metin Gurcan. Adaptive discriminant wavelet packet transform and local binary patterns for meningioma subtype classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 196–204. Springer, 2008.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, pp. 3342–3352, 2019.

Muhammad Shaban, Syed Ali Khurram, Muhammad Moazam Fraz, Najah Alsubaie, Iqra Masood, Sajid Mushtaq, Mariam Hassan, Asif Loya, and Nasir M Rajpoot. A novel digital score for abundance of tumour infiltrating lymphocytes predicts disease free survival in oral squamous cell carcinoma. *Scientific reports*, 9(1):1–13, 2019.

Saurabh R Shrivastava, Prateek S Shrivastava, and Jegadeesh Ramasamy. Oral cancer in developing countries: the time to act is upon us. *Iranian journal of cancer prevention*, 7(1):58, 2014.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *arXiv preprint arXiv:2001.06782*, 2020.

John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11), 2018.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

A APPENDIX

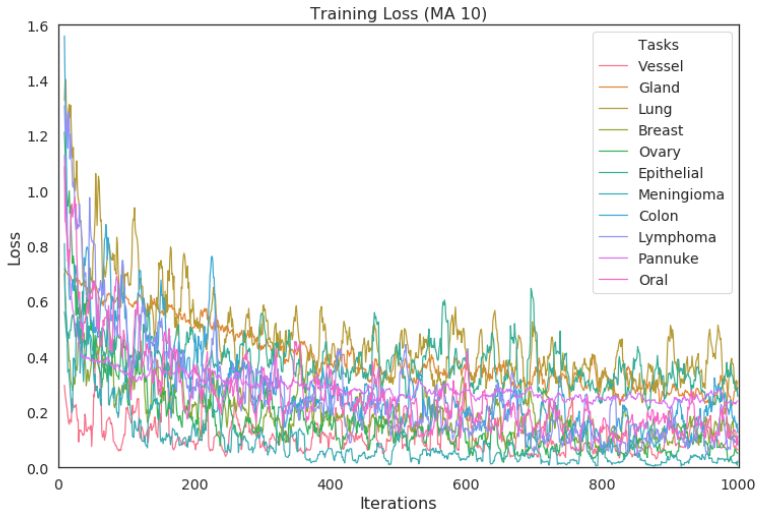

Figure 1: Loss over the first 1000 iterations, due to high variance of the loss it has been smoothed using rolling average with window of size 10.

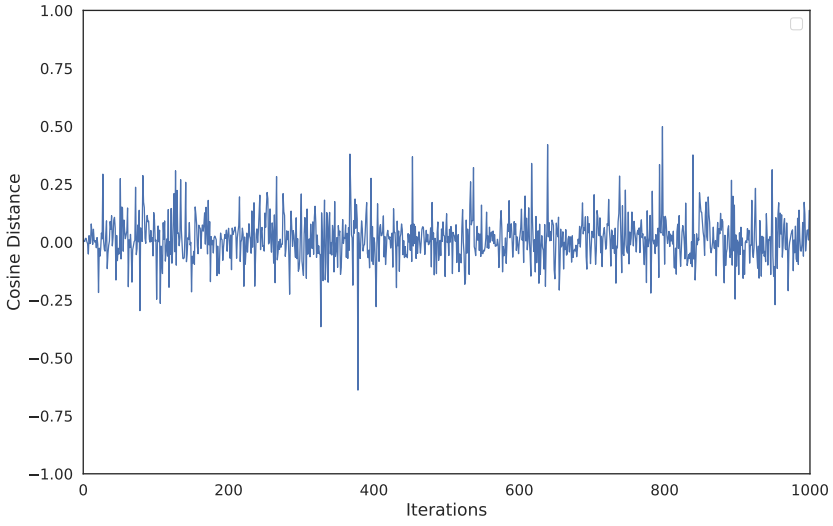

Figure 2: Cosine distance between encoder gradients for task sampled at iteration t and task at iteration $t-1$.

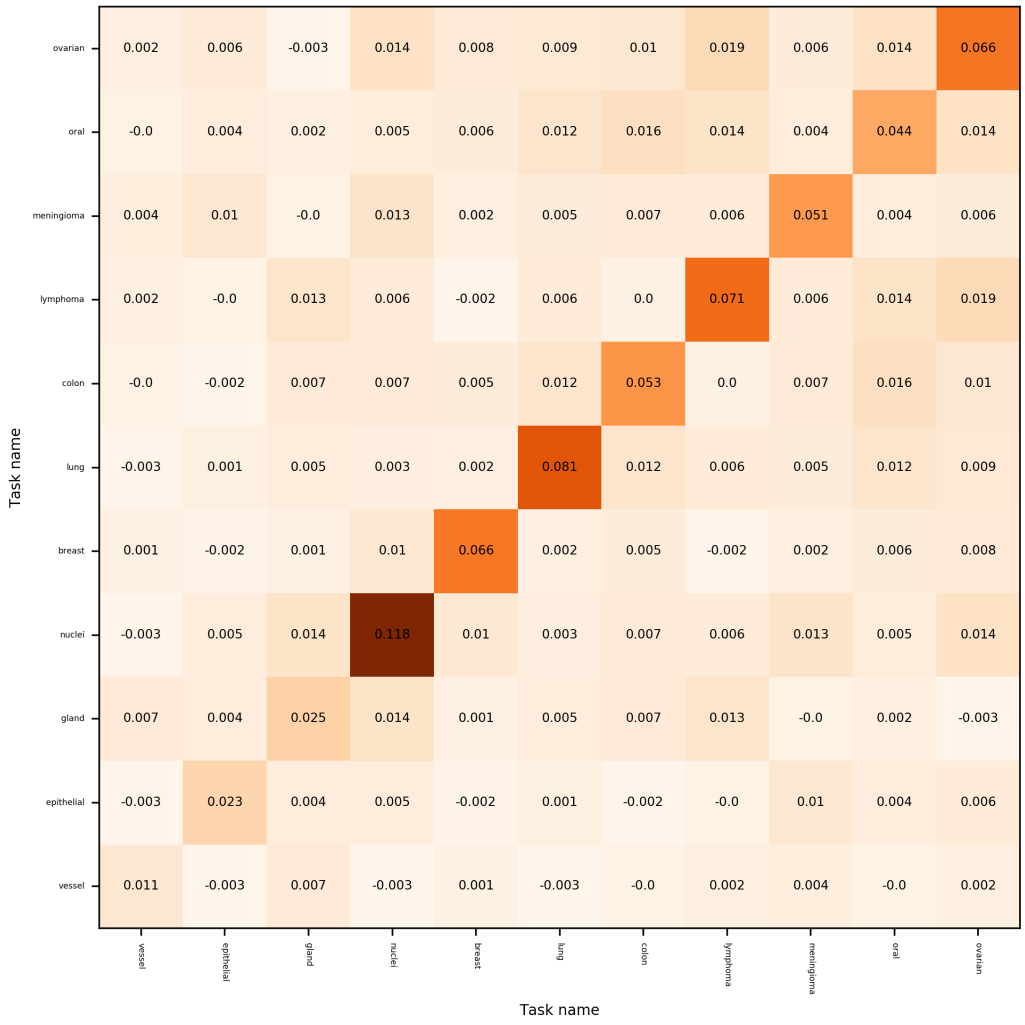

Figure 3: Rolling mean cosine distance between task gradients of the encoder.