Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning

D. Coppola[1] (speaker)

H. K. Lee[1], C. Guan[2]

[1] Bioinformatics Institute, A*STAR, Singapore

[2] School of Computer Science and Engineering, NTU, Singapore

Presented at the ISIC Skin Image Analysis Workshop @ CVPR 2020

Outline

- Introduction
- Methods
- Data
- Experiments
- Conclusions

Introduction

Melanoma identification

Rule-based procedures

- ABCD rule
- 7-point checklist method

7-point checklist method

Identification of 7 attributes; each carries a score (0, 1 or 2)

If the sum of the scores exceeds a certain threshold τ (typically 1 or 3), the lesion is deemed a melanoma

Introduction

Real-world medical application of DL is limited, despite good performance

Main barrier is the opaqueness of the models

Growing interest in developing methods to understand the mechanics of the models (XAI – Barredo Arrieta, 2020)

Introduction

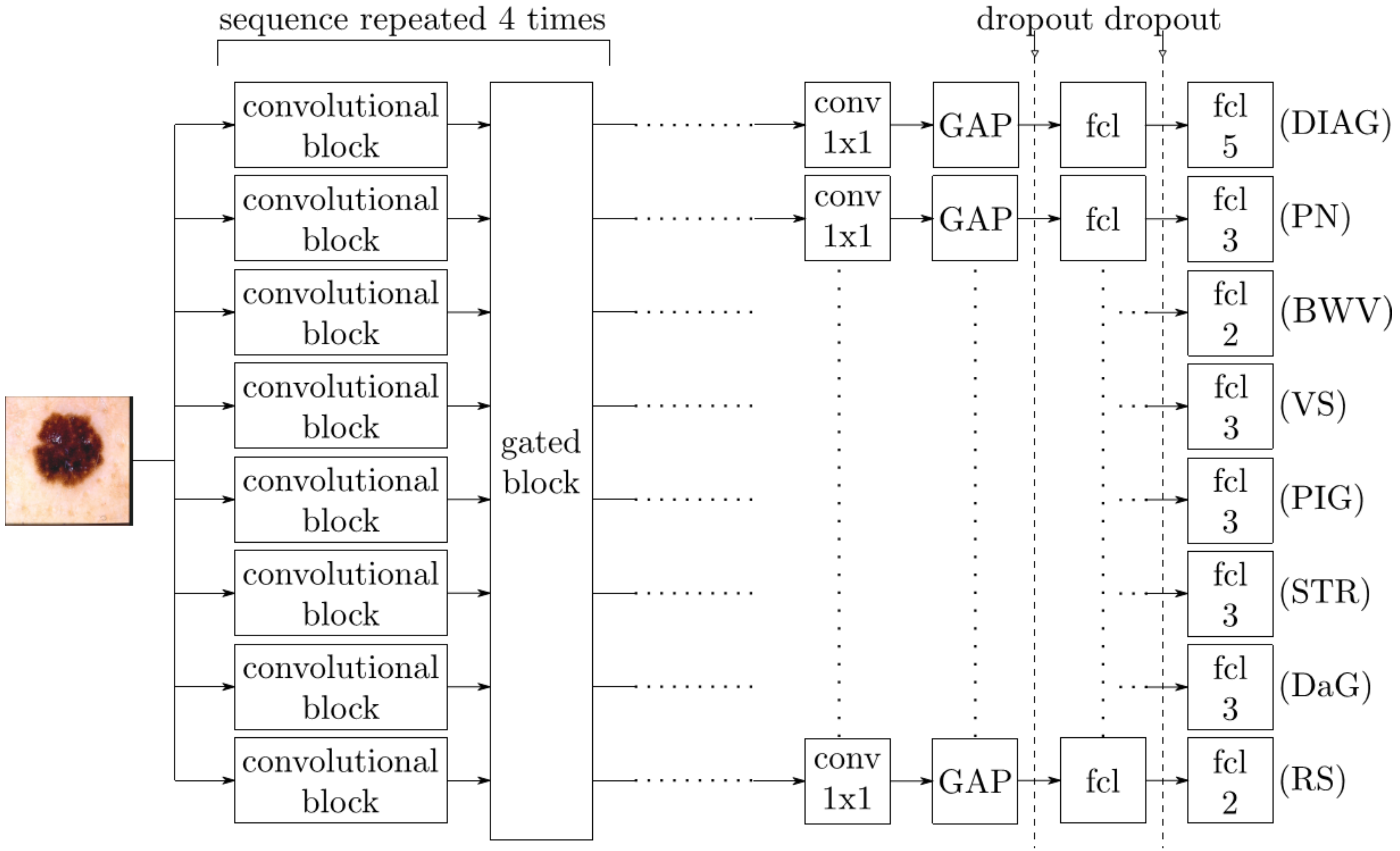
How to join rule-based methods with deep learning?

How can we examine what a DL model is learning?

Introduction

Our proposal	MTL method that learns what to share between tasks through gates
	Gates allow inspection the relationships learned by the network
	Application to the 7-point checklist method (Argenziano, 1998)

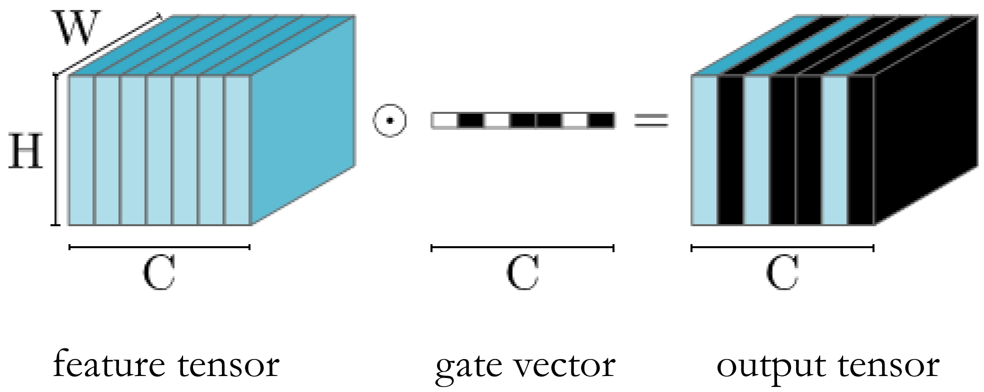
🛠 Methods – Overall System



Methods – Gates

Tasks should share features only when useful

A "gate" applied to a tensor of feature maps allows to selectively pick or suppress some features

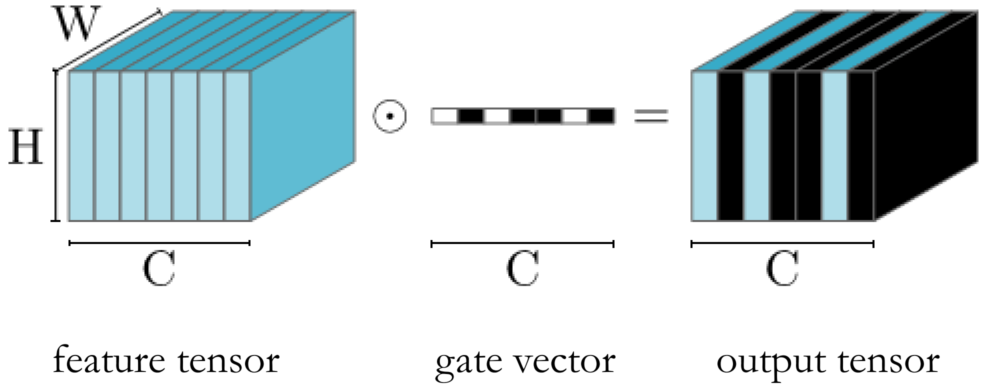


Methods – Gates

Ideally a gate would be binary

Not be learnable through gradient descent

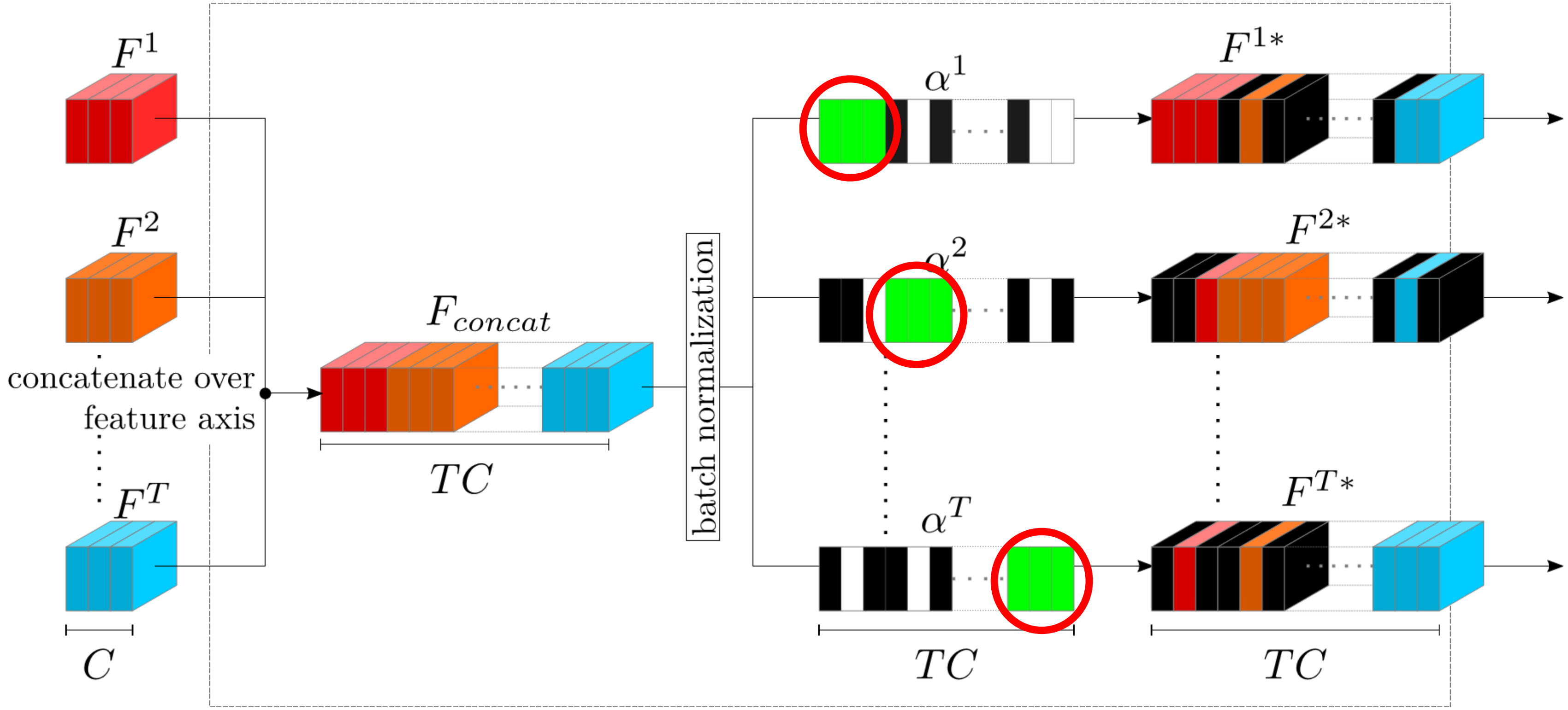
Modelled as vector of continuous values in [0, 1]



Methods – Gated Block

The gates are always “open” for the features corresponding to the task itself

Features F^t obtained through conv layer for T tasks



Features F^{t*} are input for next conv layer

Methods – Training matters

Implementation of sampling strategy from Kawahara et al. (2019)

Focal cross-entropy loss (Lin et al., 2017)

$$FL_s^t = \sum_{j}^{J^t} w_j^t y_{s,j}^t \left(1 - \widetilde{y_{s,j}^t}\right)^{\beta} \log(\widetilde{y_{s,j}^t})$$

This loss is applied to each sample for each task

t	Task index
s	Sample index
J^t	Labels for task t
j	Label index
w_j^t	Weight computed by sampling strategy
$y_{s,j}^t$	Ground truth label
$\widetilde{y_{s,j}^t}$	Predicted label
$\left(1 - \widetilde{y_{s,j}^t}\right)^{\beta}$	Focal cross-entropy coefficient ($\beta = 2$)

Data

7pt-derm dataset

1011 patient samples	Data per patient • *metadata* • *clinical image* • *dermoscopic image* • labels	Labels for 8 tasks • lesion diagnosis • 7-point checklist attributes	Train-val-test split provided

Experiments – Definition

♣ *Standard*

- basic architecture

♦ *Binary*

- DIAG has 5 unbalanced labels. What if they are grouped as “melanoma vs all”?

♠ *Gates-off*

- what happens if no sharing is permitted?

Model is always trained from scratch

Experiments – Performance

experiment	metric	Diagnosis (DIAG)	Avg. 7pt-checklist attributes
♣ standard	accuracy	45.8	61.3
	recall	45.5	57.7
	precision	40.3	55.2
♦ gates-off	accuracy	44.3	51.4
	recall	38.5	55.6
	precision	35.3	51.7
♠ binary	accuracy	77.2**	61.3
	recall	71.0 **	58.3
	precision	70.3 **	55.6
Kawahara et al., 2019	accuracy	74.2	73.6
	recall	60.4	64.7
	precision	69.6	65.4

Standard has best performance among experiments with similar setup

Closing the gates shows slight drop in performance

Binary has easier DIAG classification but otherwise comparable performance

Experiments – Performance

experiment	metric	Diagnosis (DIAG)	Avg. 7pt-checklist attributes
♣ standard	accuracy	45.8	61.3
	recall	45.5	57.7
	precision	40.3	55.2
♦ gates-off	accuracy	44.3	51.4
	recall	38.5	55.6
	precision	35.3	51.7
♠ binary	accuracy	77.2**	61.3
	recall	71.0 **	58.3
	precision	70.3 **	55.6
Kawahara et al., 2019	accuracy	74.2	73.6
	recall	60.4	64.7
	precision	69.6	65.4

Method by Kawahara et al. (2019) has better overall performance

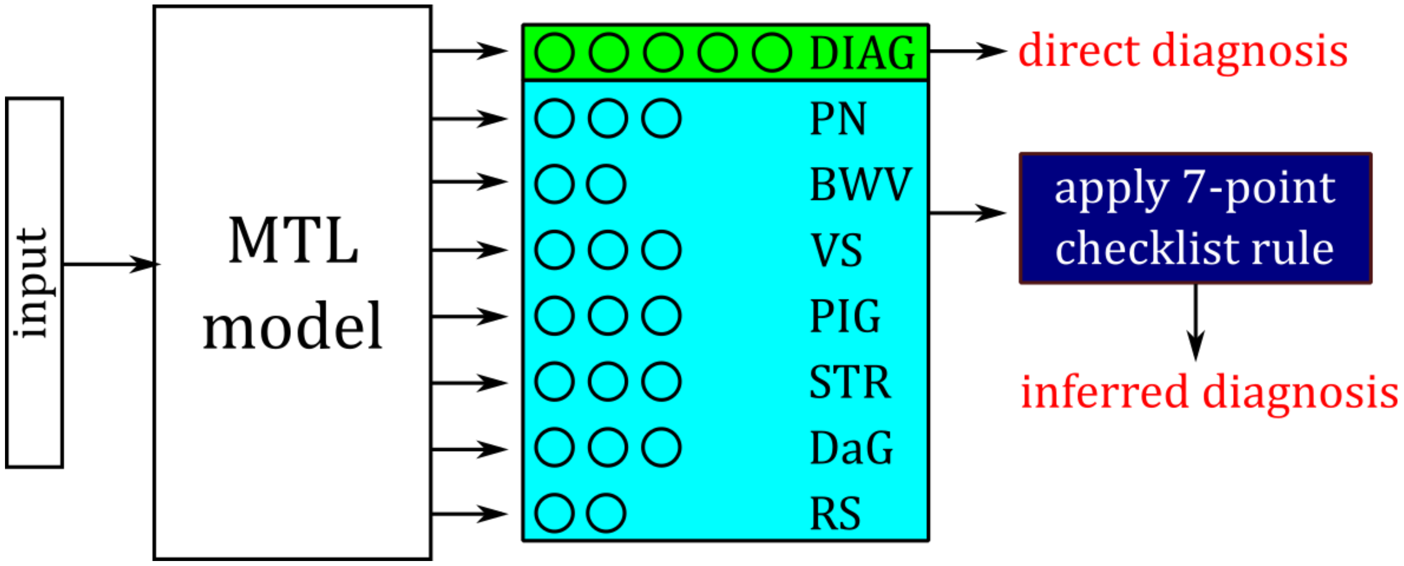
Possible reasons

Use of additional data (metadata, clinical images) in the pipeline	Starts from pre-trained network on ImageNet

Experiments – Application of the 7pt-checklist rule

The 7-point checklist rule can be applied on the predicted attributes as an additional way of determining the diagnosis (only as "melanoma vs all)

- *Direct diagnosis*: the model's prediction of the DIAG task
- *Inferred diagnosis*: the diagnosis obtained by applying the 7-point checklist method on the predicted attributes

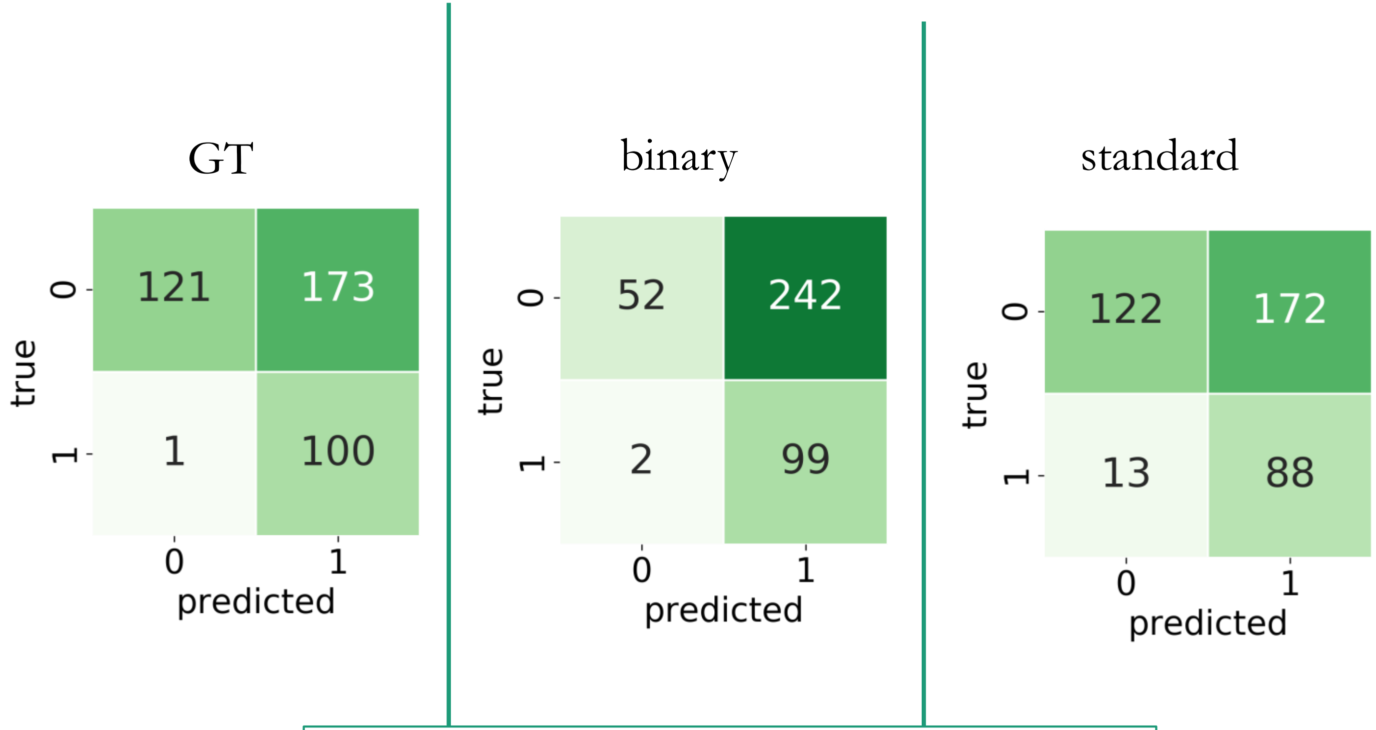


Experiments – Application of the 7pt-checklist rule

GT: application of the 7-point checklist rule on the ground truth labels

Using the 7pt rule, *binary* and *standard* have similar performance to GT when inferring melanoma

A low threshold ($\tau = 1$) provides high sensitivity to melanoma but many false positives



1: melanoma; 0: otherwise

Experiments – Sharing Fraction

Defined as the average value of the gates between task t (taking the features) and i (giving the features)

$$\mathrm{SF}_i^t = \frac{1}{C}\sum_{c}^{C} \alpha_{i,c}^t$$

Indicates the amount of sharing between two tasks at a given *gated block*.

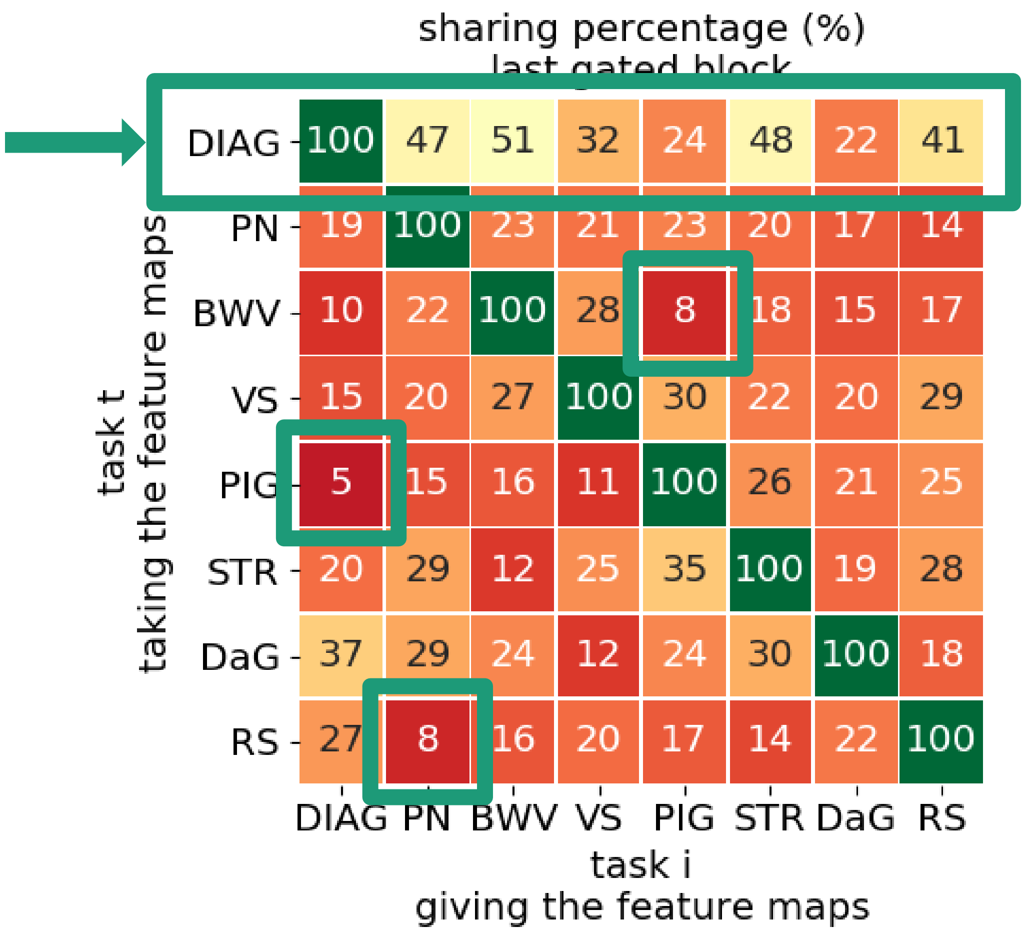
Experiments – Sharing Fraction

Looking at the SF at the last gated block for experiment *standard*

DIAG is the task that has more sharing with the other task

- High values with the major criteria (PN, BWV, VS)

In the other rows, some values are close to 0, the model is learning to be selective



Conclusions – Summary

New framework for MTL

- Based on gates that learn what to features to share among tasks
- 7-point checklist fits MTL model design

Gates allow to inspect the learned relationships between tasks

- Give insights on the mechanisms of the model
- Strategy shows selectivity in choosing which features to share

Conclusions – Future directions

Performance matters

- Experiment with different task-specific architectures
- Include the metadata in the pipeline

Qualitative insights

- Explore advanced metric to evaluate the sharing between tasks
- Discuss findings with practitioners

Thank you for your attention ☺

Contacts

Davide Coppola (davidec@bii.a-star.edu.sg)

References

- Argenziano, G. et al., 1998. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis. Arch Dermatol 134, 1563–1570. https://doi.org/10.1001/archderm.134.12.1563
- Barredo Arrieta, A. et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
- Kawahara, J. et al., 2019. Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. IEEE Journal of Biomedical and Health Informatics 23, 538–546. https://doi.org/10.1109/JBHI.2018.2824327
- Lin, T.-Y. et al., 2017. Focal Loss for Dense Object Detection. arXiv:1708.02002 [cs].