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

D. Coppola¹(speaker)
H. K. Lee¹, C. Guan²

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
Introduction

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 $\tau$ (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?

15 Jun 2020

Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning
<table>
<thead>
<tr>
<th>Introduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our proposal</td>
</tr>
<tr>
<td>MTL method that learns what to share between tasks through gates</td>
</tr>
<tr>
<td>Gates allow inspection the relationships learned by the network</td>
</tr>
<tr>
<td>Application to the 7-point checklist method (Argenziano, 1998)</td>
</tr>
</tbody>
</table>
Methods – Overall System

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

Features $F^t$ obtained through conv layer for $T$ tasks

concatenate over feature axis

$F_{concat}$

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

$\alpha^1$

$\alpha^2$

$\alpha^T$

$F^1$

$F^2$

$F^T$

$F^1*$

$F^2*$

$F^T*$

$C$

$TC$

$TC$

$TC$

batch normalization

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)

\[ FL_s^t = \sum_j w_j^t y_{s,j} \left( 1 - \tilde{y}_{s,j}^t \right)^\beta \log(\tilde{y}_{s,j}^t) \]

This loss is applied to each sample for each task
<table>
<thead>
<tr>
<th>7pt-derm dataset</th>
<th>Data per patient</th>
<th>Labels for 8 tasks</th>
<th>Train-val-test split provided</th>
</tr>
</thead>
<tbody>
<tr>
<td>1011 patient samples</td>
<td>• metadata</td>
<td>• lesion diagnosis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• clinical image</td>
<td>• 7-point checklist attributes</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• dermoscopic image</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• labels</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Data

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

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

- **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

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

**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**

---

Kawahara et al., 2019

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

15 Jun 2020

Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning
Experiments – Application of the 7pt-checklist rule

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.
Experiments – Sharing Fraction

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

$$SF^t_i = \frac{1}{C} \sum_{c} \alpha^t_{i,c}$$

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

Interpreting mechanisms of prediction for skin cancer diagnosis using multi-task learning
Thank you for your attention 😊

Contacts

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


