Solo or Ensemble?  
Choosing a CNN Architecture for Melanoma Classification

Fábio Perez¹, Sandra Avila², Eduardo Valle¹  
¹Recod Lab., DCA, FEEC, University of Campinas (UNICAMP)  
²Recod Lab., IC, University of Campinas (UNICAMP)
Convolutional Neural Networks

SotA for most computer vision problems, including skin lesion analysis

Used by all winner submissions in ISIC Challenges 2016, 2017, 2018
CNN Architectures

AlexNet
CNN Architectures

- AlexNet
- GoogLeNet
- VGG
- ZFNet
CNN Architectures

- AlexNet
- ZFNet
- VGG
- Inception
- GoogLeNet
- Inception-ResNet
- ResNet
- DualPathNet
- NASNet
- DenseNet
- MobileNet
- SE-Net
- Xception
- SqueezeNet
- ResNeXt
- PNASNet
- Inception-ResNet
CNN Architectures

ISIC Challenges

2016  ResNet

2017  ResNet, Inception

2018  ResNet, Inception, DenseNet, ResNeXt, PNASNet, DPN, SENet...
Transfer Learning

The **most critical factor** for model performance

**SotA** for most computer vision problems, including **skin lesion analysis**

Also used by all **ISIC Challenges winners**


Do better ImageNet models transfer better?

Short answer: Yes

For multiple natural datasets

Fine-tuning, fixed features, and random initialization

Do better ImageNet models transfer better?

Kornblith et al. (2018)  arxiv.org/abs/1805.08974
How to predict model performance?
Experimental Design

9 architectures
× 5 splits
× 3 replicates
= 135 experiments
Experimental Design

9 architectures \times 5 \text{ splits} \times 3 \text{ replicates} = \textbf{135} \text{ experiments}
Experimental Design

9 architectures × 5 splits × 3 replicates = 135 experiments

ISIC 2017

1750 train
500 validation
500 test
Explored factors

**Architectural**
- Acc@1 on ImageNet
- # of Parameters
- Date of Publication

**Training**
- AUC
- Accuracy
- Sensitivity
- Specificity
- Loss
- Validation
- Test
- # of Epochs
- Date of Publication
## Results

<table>
<thead>
<tr>
<th>16</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>54</td>
</tr>
<tr>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>38</td>
<td>52</td>
</tr>
<tr>
<td>25</td>
<td>12</td>
</tr>
<tr>
<td>39</td>
<td>55</td>
</tr>
<tr>
<td>28</td>
<td>40</td>
</tr>
<tr>
<td>38</td>
<td>55</td>
</tr>
<tr>
<td>17</td>
<td>20</td>
</tr>
<tr>
<td>46</td>
<td>82</td>
</tr>
<tr>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>44</td>
<td>82</td>
</tr>
<tr>
<td>54</td>
<td>82</td>
</tr>
<tr>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td>48</td>
<td>56</td>
</tr>
<tr>
<td>19</td>
<td>28</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>36</td>
<td>18</td>
</tr>
<tr>
<td>65</td>
<td>54</td>
</tr>
</tbody>
</table>
Results
Results (without MobileNetV2)
## Datasets

**Kornblith et al. (2018)**
- Multiple large datasets
- One factor: Acc@1
- Hyperparameter tuning

**Ours**
- ISIC 2017 (2750 images)
- Multiple factors
- “Best-practice” hyperparameters
<table>
<thead>
<tr>
<th><strong>Datasets</strong></th>
<th><strong>Kornblith et al. (2018)</strong></th>
<th><strong>Ours</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>➔ Multiple large datasets</td>
<td>➔ Multiple factors</td>
<td>➔ ISIC 2017 (2750 images)</td>
</tr>
<tr>
<td>➔ One factor: Acc@1</td>
<td>➔ “Best-practice” hyperparameters</td>
<td>➔ Five splits</td>
</tr>
<tr>
<td>➔ Hyperparameter tuning</td>
<td>➔</td>
<td>➔ Three replicates</td>
</tr>
<tr>
<td>➔ One split per dataset</td>
<td>➔ No replicates</td>
<td>➔</td>
</tr>
<tr>
<td>➔</td>
<td>➔</td>
<td>➔</td>
</tr>
</tbody>
</table>
Ensembles
Creating the Ensembles

9 architectures × 3 replicates = 27 models per split

For each split, ensemble 1, 2, ..., 27 models

Two strategies for adding models:
  in random order
  models with best validation AUC first
Results

The graph shows the AUC (Area Under the Curve) for different sorting methods across varying numbers of models. The x-axis represents the number of models, and the y-axis represents the AUC values. Different sorting methods are indicated by distinct lines, with 'validation' and 'random' marked specifically. The graph indicates that as the number of models increases, the AUC values tend to stabilize for both sorting methods.
Results (normalized)
For the SotA models, performance on ImageNet does not necessarily translate to performance on melanoma detection. Validation metrics correlate with test metrics much better than validation loss. Ensembles are needed for stable SotA performance; large ensembles work okay from simply picking at random from a pool of SotA individual models.
Acknowledgments

UNICAMP

reasoning for complex data

CNPq

FAPESP

eScience

NVIDIA

Azure
Thanks!