Supplementary Material - Studying and Exploiting the Relationship Between Model Accuracy and Explanation Quality

Yunzhe Jia[1] , Eibe Frank[1,2], Bernhard Pfahringer[1,2], Albert Bifet[1,3], and Nick Lim[1]

[1] AI Institute, University of Waikato, Hamilton, New Zealand
[2] Department of Computer Science, University of Waikato, Hamilton, New Zealand
[3] LTCI, Télécom Paris, IP Paris, France
{ajia,eibe,bernhard,abifet,nlim}@waikato.ac.nz

Appendix A - Results of explanation quality on test data

Table 1 gives the results of explanation quality on test data in the model selection scenario (Section 5.5).

α	AlexNet	DenseNet	Inception-V3	ResNet-18	ResNet-50	SqueezeNet	VGG-11
0	**0.5728**	**0.6039**	**0.4722**	**0.5935**	**0.6235**	**0.594**	**0.6215**
0.1	0.5727	0.6036	0.472	0.5933	0.6233	0.5916	0.6208
0.2	0.5716	0.6033	0.4711	0.5917	0.6226	0.5882	0.6195
0.3	0.5715	0.6024	0.4712	0.5917	0.6217	0.5874	0.6195
0.4	0.5647	0.6026	0.4712	0.5912	0.6191	0.5851	0.6195
0.5	0.5647	0.602	0.4712	0.5912	0.6191	0.5824	0.6195
0.6	0.5598	0.602	0.4712	0.5912	0.6191	0.5824	0.6195
0.7	0.5598	0.602	0.4712	0.5912	0.6191	0.5819	0.6195
0.8	0.5598	0.602	0.4712	0.5912	0.6191	0.5819	0.6195
0.9	0.5598	0.602	0.4712	0.5912	0.6191	0.5819	0.6195
1.0	0.5567	0.5999	0.4699	0.5875	0.614	0.5715	0.616

Table 1: Results - Explanation quality on test data of models selected using selection criterion with different α. Best metrics for each model are highlighted in bold.

Appendix B - Results of test accuracy of models selected with different percentages of expert explanations for other six neural networks

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	81.8%	81.8%	81.1%	82.3%	81.5%	82.3%	81.7%	82.0%	82.1%	80.8%
0.1	81.8%	81.8%	82.7%	82.2%	81.6%	82.3%	82.2%	82.0%	81.9%	81.4%
0.2	81.8%	**82.3**%	82.7%	82.2%	81.6%	82.3%	82.0%	82.0%	81.9%	81.5%
0.3	81.8%	82.1%	**82.8**%	82.2%	81.9%	**82.4**%	82.1%	82.2%	81.8%	81.6%
0.4	81.7%	82.1%	**82.8**%	82.3%	82.0%	82.3%	**82.4**%	82.2%	**82.3**%	**81.8**%
0.5	81.8%	82.1%	82.7%	**82.5**%	81.9%	82.0%	**82.4**%	82.2%	**82.3**%	**81.8**%
0.6	81.8%	82.1%	82.4%	82.2%	81.9%	82.1%	81.6%	**82.3**%	**82.3**%	81.7%
0.7	**82.3**%	82.0%	82.3%	81.5%	**82.1**%	82.1%	81.6%	81.9%	82.1%	81.3%
0.8	81.9%	81.9%	82.5%	81.4%	81.5%	82.1%	81.4%	81.7%	81.7%	81.3%
0.9	81.3%	81.9%	81.7%	81.3%	81.3%	81.5%	81.4%	81.7%	81.7%	81.3%
1.0	81.8%	81.8%	81.8%	81.8%	81.8%	81.8%	81.8%	81.8%	81.8%	**81.8**%

Table 2: Results - Test accuracy of Alexnet models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	74.1%	73.7%	**75.2**%	73.5%	74.9%	73.9%	74.0%	**75.1**%	72.6%	67.2%
0.1	74.1%	74.1%	74.8%	**74.8**%	**75.0**%	73.9%	74.0%	74.9%	74.3%	72.9%
0.2	**75.2**%	74.1%	74.8%	74.3%	73.7%	**75.1**%	74.0%	**75.1**%	74.3%	74.0%
0.3	**75.2**%	74.2%	74.8%	74.3%	74.9%	**75.1**%	74.1%	74.7%	74.3%	**74.2**%
0.4	**75.2**%	74.3%	75.0%	74.3%	74.5%	74.3%	**74.6**%	74.5%	74.2%	**74.2**%
0.5	**75.2**%	74.0%	75.1%	74.6%	74.5%	74.3%	**74.6**%	74.5%	74.0%	**74.2**%
0.6	**75.2**%	74.0%	74.8%	74.6%	74.4%	73.8%	74.5%	74.5%	74.1%	**74.2**%
0.7	75.0%	74.3%	74.5%	74.6%	74.4%	74.1%	**74.6**%	74.5%	**74.6**%	**74.2**%
0.8	74.7%	74.3%	74.5%	74.6%	74.4%	74.2%	**74.6**%	74.6%	**74.6**%	**74.2**%
0.9	74.6%	**74.5**%	74.5%	74.7%	74.5%	74.2%	**74.6**%	74.6%	**74.6**%	**74.2**%
1.0	74.0%	74.0%	74.0%	74.0%	74.0%	74.0%	74.0%	74.0%	74.0%	74.0%

Table 3: Results - Test accuracy of Inception-V3 models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	81.5%	81.8%	81.5%	81.8%	81.4%	81.3%	81.5%	81.6%	81.7%	**81.7**%
0.1	81.6%	81.8%	81.5%	81.8%	81.5%	81.3%	81.5%	81.8%	**81.9**%	**81.7**%
0.2	81.7%	81.8%	81.5%	81.8%	81.6%	81.3%	**81.8**%	81.8%	**81.9**%	**81.7**%
0.3	81.7%	81.8%	81.5%	81.8%	81.6%	81.3%	**81.8**%	**81.9**%	**81.9**%	**81.7**%
0.4	81.7%	81.8%	81.6%	81.8%	**81.7**%	81.3%	**81.8**%	81.8%	**81.9**%	**81.7**%
0.5	81.7%	**81.9**%	81.6%	81.8%	81.6%	**81.6**%	**81.8**%	81.8%	**81.9**%	**81.7**%
0.6	81.7%	**81.9**%	81.6%	**81.9**%	81.6%	**81.6**%	81.7%	81.7%	81.8%	**81.7**%
0.7	**81.8**%	**81.9**%	**81.8**%	81.6%	81.6%	**81.6**%	81.7%	81.7%	81.7%	**81.7**%
0.8	81.7%	**81.9**%	81.6%	81.7%	81.6%	**81.6**%	81.7%	81.7%	81.7%	**81.7**%
0.9	81.6%	81.8%	81.5%	81.7%	81.6%	**81.6**%	81.7%	81.7%	81.7%	**81.7**%
1.0	81.4%	81.4%	81.4%	81.4%	81.4%	81.4%	81.4%	81.4%	81.4%	81.4%

Table 4: Results - Test accuracy of Resnet-18 models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	80.4%	81.0%	81.5%	81.2%	79.2%	**81.5**%	81.4%	81.3%	81.3%	80.2%
0.1	80.4%	81.0%	81.5%	81.2%	81.3%	**81.5**%	81.4%	**81.5**%	**81.5**%	80.3%
0.2	81.4%	81.0%	81.5%	81.5%	81.3%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
0.3	81.4%	81.0%	81.5%	81.5%	81.3%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
0.4	81.4%	81.0%	81.5%	81.5%	**81.5**%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
0.5	81.4%	**81.2**%	81.5%	81.5%	81.4%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
0.6	81.4%	**81.2**%	81.5%	81.5%	81.4%	**81.5**%	**81.6**%	**81.5**%	**81.5**%	**81.6**%
0.7	81.4%	**81.2**%	**81.6**%	**81.6**%	81.4%	**81.5**%	**81.6**%	**81.5**%	**81.5**%	**81.6**%
0.8	81.3%	**81.2**%	81.5%	81.5%	**81.5**%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
0.9	**81.5**%	**81.2**%	81.5%	81.5%	**81.5**%	**81.5**%	81.5%	**81.5**%	**81.5**%	**81.6**%
1.0	81.0%	81.0%	81.0%	81.0%	81.0%	81.0%	81.0%	81.0%	81.0%	81.0%

Table 5: Results - Test accuracy of Resnet-50 models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	79.9%	80.3%	81.3%	78.9%	**80.8**%	81.0%	81.0%	78.5%	79.6%	77.6%
0.1	80.4%	80.9%	82.2%	79.8%	**80.8**%	81.0%	81.0%	81.0%	**81.7**%	79.6%
0.2	80.4%	80.9%	82.2%	80.2%	**80.8**%	81.0%	81.1%	**81.6**%	81.4%	80.4%
0.3	80.8%	81.2%	82.2%	80.7%	80.5%	81.1%	81.1%	81.4%	81.5%	81.0%
0.4	80.8%	**82.0**%	**82.3**%	**81.6**%	80.5%	**81.2**%	81.1%	81.1%	81.4%	**81.2**%
0.5	81.4%	81.6%	**82.3**%	81.1%	**80.8**%	**81.2**%	81.0%	81.2%	81.1%	**81.2**%
0.6	81.4%	81.7%	**82.3**%	81.1%	80.4%	80.6%	**81.6**%	81.1%	81.5%	**81.2**%
0.7	**81.5**%	81.7%	81.6%	81.1%	80.6%	80.6%	81.2%	80.5%	80.9%	**81.2**%
0.8	81.1%	81.1%	81.4%	81.4%	80.7%	80.6%	81.2%	80.5%	81.0%	**81.2**%
0.9	81.2%	81.0%	80.3%	81.2%	80.7%	80.6%	81.2%	80.5%	81.0%	**81.2**%
1.0	79.8%	79.8%	79.8%	79.8%	79.8%	79.8%	79.8%	79.8%	79.8%	79.8%

Table 6: Results - Test accuracy of SqueezeNet models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.

	Level of expert explanation availability									
α	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
0	83.1%	**84.0**%	83.7%	83.8%	83.8%	**83.9**%	**84.3**%	**83.8**%	**84.1**%	84.0%
0.1	83.1%	**84.0**%	83.8%	83.8%	83.4%	**83.9**%	**84.3**%	**83.8**%	**84.1**%	**84.1**%
0.2	83.6%	**84.0**%	83.8%	83.8%	83.6%	**83.9**%	84.2%	83.7%	**84.1**%	83.9%
0.3	83.6%	**84.0**%	83.8%	**84.0**%	83.8%	83.8%	84.2%	83.7%	84.0%	83.9%
0.4	83.6%	**84.0**%	83.8%	83.9%	**84.1**%	83.8%	84.1%	83.7%	**84.1**%	83.9%
0.5	83.6%	**84.0**%	**84.0**%	83.9%	**84.1**%	83.8%	84.1%	83.6%	83.9%	83.9%
0.6	83.6%	**84.0**%	**84.0**%	83.9%	**84.1**%	83.8%	84.0%	83.6%	83.9%	83.9%
0.7	**84.3**%	83.8%	**84.0**%	83.9%	**84.1**%	83.8%	84.0%	83.6%	83.9%	83.9%
0.8	84.2%	83.7%	83.8%	83.9%	**84.1**%	83.8%	84.0%	83.6%	83.9%	83.9%
0.9	84.0%	83.7%	83.8%	83.7%	83.8%	83.8%	84.0%	83.6%	83.9%	83.9%
1.0	83.3%	83.3%	83.3%	83.3%	83.3%	83.3%	83.3%	83.3%	83.3%	83.3%

Table 7: Results - Test accuracy of VGG-11 models selected with different percentages of expert explanations. Best results for each model are highlighted in bold.