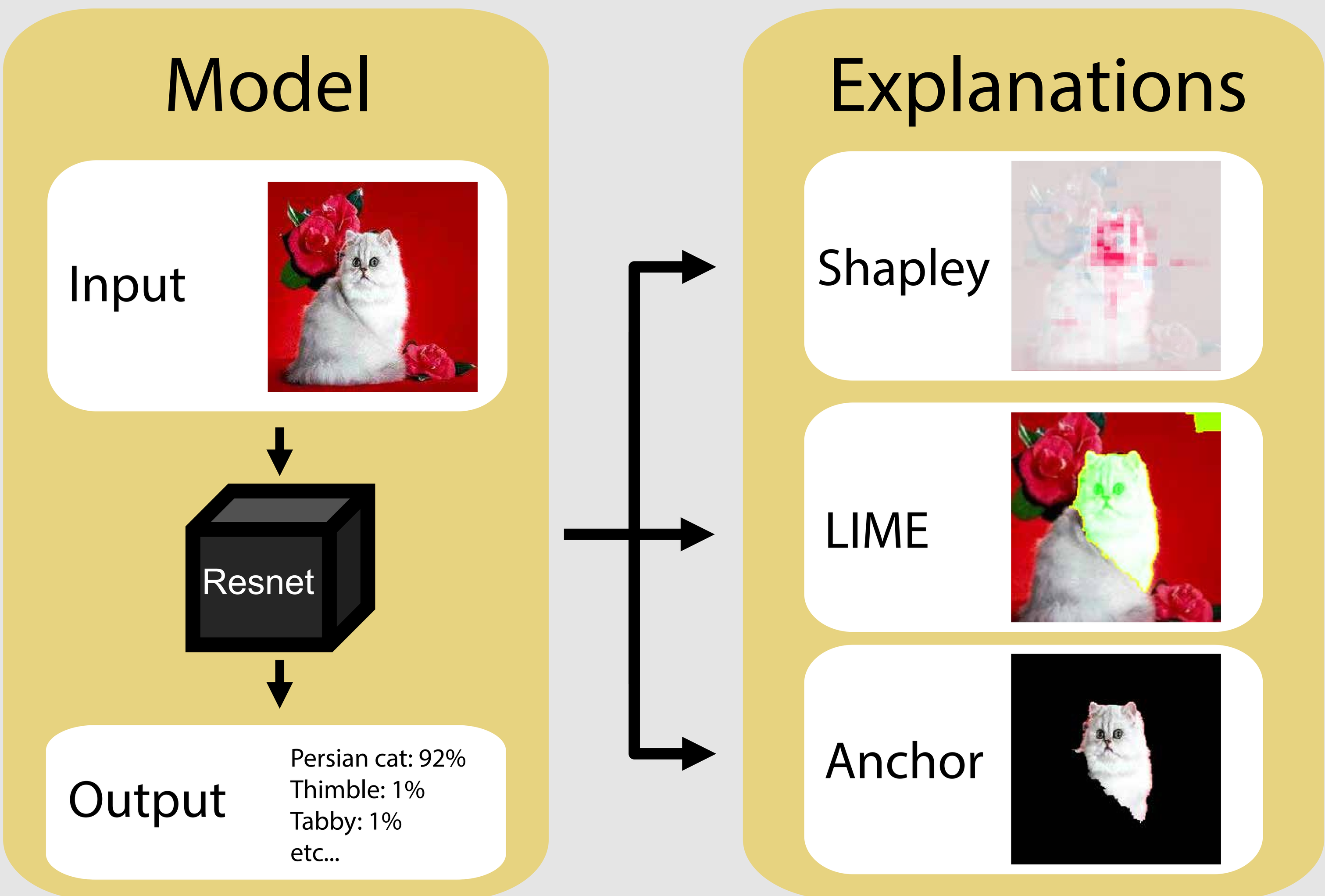
Introduction

Machine learning models are more capable than ever before of performing complicated tasks like pattern recognition. But with their growth has come an increased need for transparency.

Transparency is important because these models are being used in sensitive areas, and we want to know if our models are biased or are learning the wrong things. In the EU, users are even guaranteed an explanation of any algorithm that significantly affects them. But for image models, what does an "explanation" even mean?



Shapley

Shapley values, originating from cooperative game theory, treat each group of pixels like a player and class confidences like a game. Each "player's" contribution is measured by the confidence of the model with random subsets of the other pixels excluded. Those that were useful in determining the model's class output are colored red, while those that counted against are blue.

LIME

LIME separates the image into a few constituent "superpixel" groups, with similar colors being grouped together. Then, by toggling each one off in turn, it ascertains which ones make the model's confidence change fastest. Sections that the model is confused without are marked in green, while those that the model is surer without (if any) are marked in red.

Anchoring

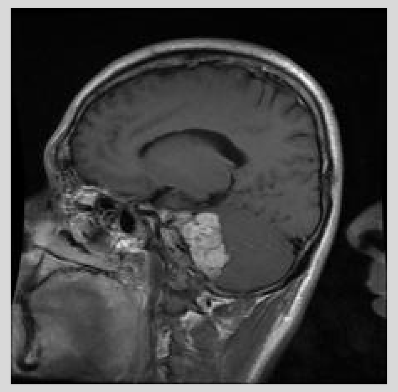
Anchoring, similarly, tries to find which superpixels are the biggest determiners of the model's confidence. Upon finding them, it blacks out the least important superpixels until the model's remaining confidence in the prediction drops to a preset threshold.

Explainable AI: Breaking Down the Black Box

Josh Moore, Adrian Boskovic, Chris Melville,
Lev Shuster, Sam Johnson-Lacoss, Thomas Pree

This is a German shorthaired pointer

(True)

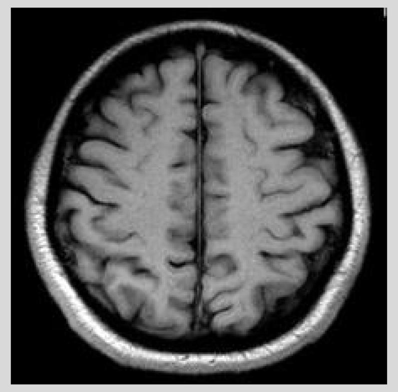


This person has a meningioma tumor

(True)

This is a horned viper

(False)



This person has a meningioma tumor

(False)

This is a polar bear

(False)

This is a tennis ball

(Arguably)

References & Acknowledgements

Our group was led and mentored by Anna Rafferty. In addition, we drew heavy inspiration from Ribearo et al. for the design of our user study.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778

Besse, Philippe & Castets-Renard, Céline & Garivier, Aurélien & Loubes, Jean-Michel. (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version). 10.13140/RG.2.2.22973.31207.

Further Citations

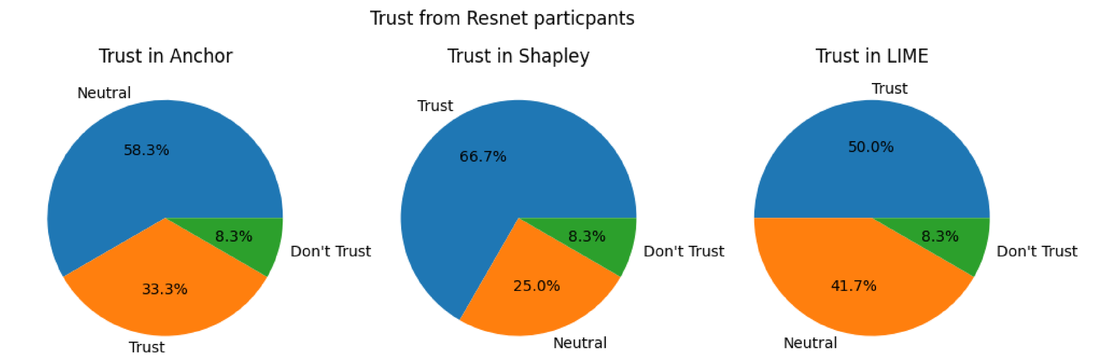
One classic example of the power of explainable AI comes from a model that was trained to differentiate between wolves and huskies. The model had high accuracy, until it was tested in the field. As it turned out, all the images of wolves in the training data had snow in them as well, so the model had only learned to recognize snow. (Besse et al.)

User Study

We asked 12 Carleton students to review curated pictures of cats and dogs, along with Resnet's prediction about the subject of the image and our three explanations of the predictions. For each explanation, we asked how it changed their understanding of the model's thought process, and which visualizations they preferred. Overall, we aimed to determine which of these techniques – applied "out of the box" to image data – were most useful at building understanding of the model.

Findings

Shapley was the favorite technique, but not by a consistent margin.
"If Shapley always had the edge in accuracy I would choose that, but anchoring and lime have better display"



There were a few classes of problems that the techniques couldn't explain. (Regarding the third image from the to) "I can see why the algorithm would think its a snake, but can't see why it doesn't think its a cat"

Participants' CS experience levels did not strongly correlate with preferred techniques, but those with more experience were much less accepting of the discrepancies between different techniques.
"If I was given any single one, I'd say they help, but they confuse in total"

Conclusions

- These techniques can be applied, with little customization, to a range of AI models and on both numerical/image data.
- However, without a preexisting understanding of the data used by the model, these techniques may only confirm biases or confuse.
- Confirmation bias lessens the ability of these techniques to report model failures.
- The explanations are unable to compare confidences between different classes, explore global data, or find where the machine learned its judgements.
- Despite the limitations, they do provide more insight than the black box.

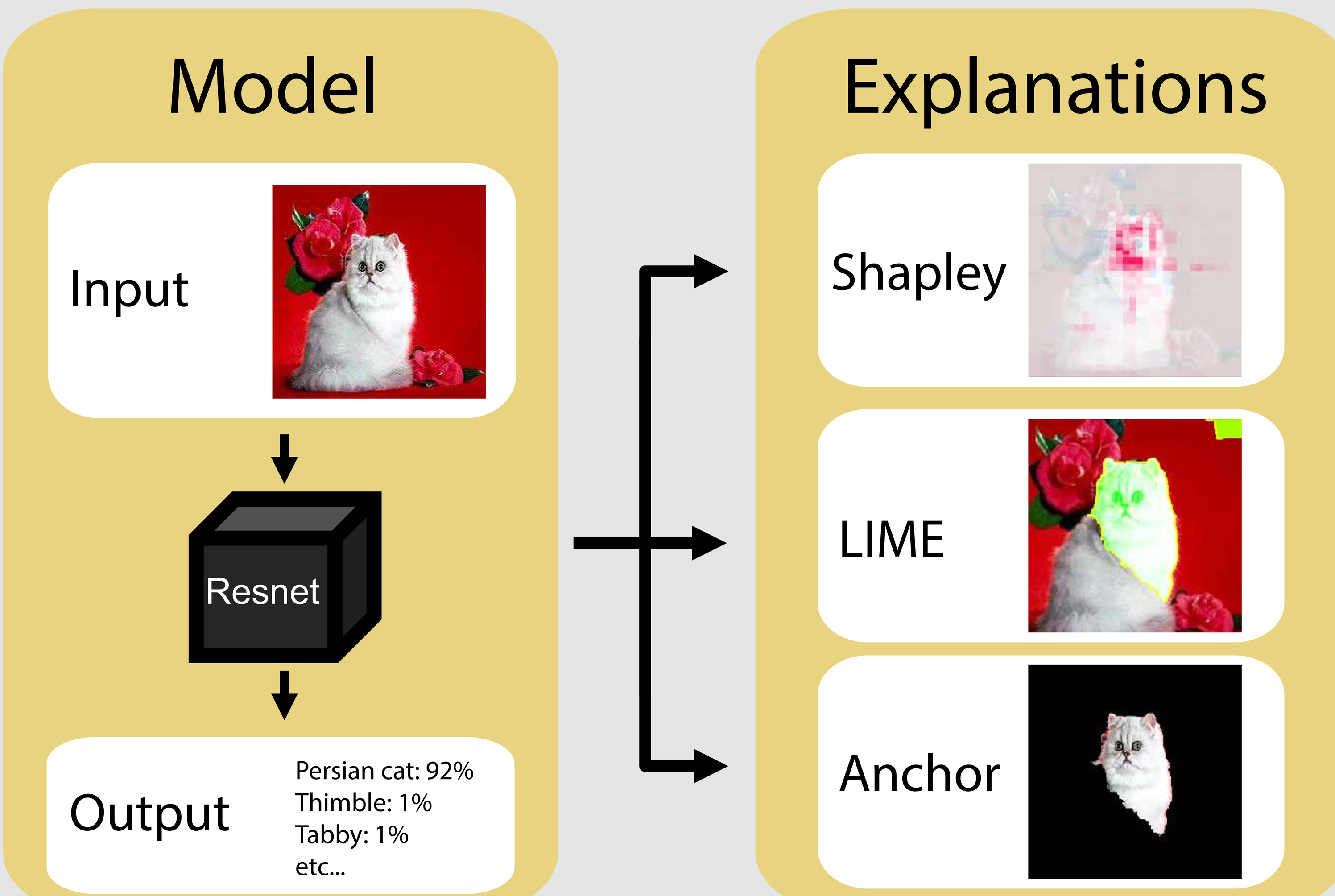
Explainable AI: Breaking Down the Black Box

Josh Moore, Adrian Boskovic, Chris Melville,
Lev Shuster, Sam Johnson-Lacoss, Thomas Pree

Introduction

Machine learning models are more capable than ever before of performing complicated tasks like pattern recognition. But with their growth has come an increased need for transparency.

Transparency is important because these models are being used in sensitive areas, and we want to know if our models are biased or are learning the wrong things. In the EU, users are even guaranteed an explanation of any algorithm that significantly affects them. But for image models, what does an "explanation" even mean?



Shapley

Shapley values, originating from cooperative game theory, treat each group of pixels like a player and class confidences like a game. Each "player's" contribution is measured by the confidence of the model with random subsets of the other pixels excluded. Those that were useful in determining the model's class output are colored red, while those that counted against are blue.

LIME

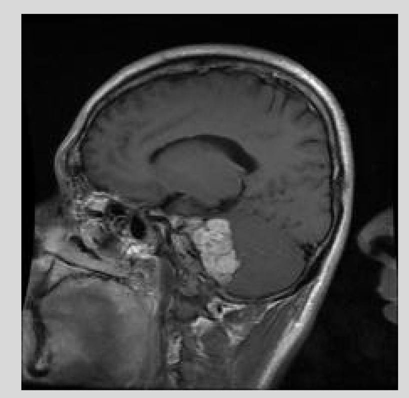
LIME separates the image into a few constituent "superpixel" groups, with similar colors being grouped together. Then, by toggling each one off in turn, it ascertains which ones make the model's confidence change fastest. Sections that the model is confused without are marked in green, while those that the model is surer without (if any) are marked in red.

Anchoring

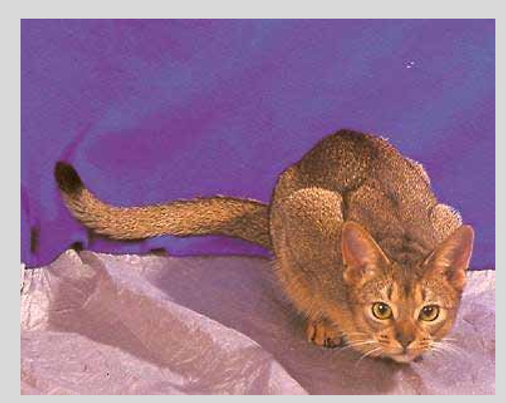
Anchoring, similarly, tries to find which superpixels are the biggest determiners of the model's confidence. Upon finding them, it blacks out the least important superpixels until the model's remaining confidence in the prediction drops to a preset threshold.

Model Confidences
German Shorthaired Pointer: 0.94
Chesapeake Bay Retriever: 0.03
Weimaraner: 0.01
...

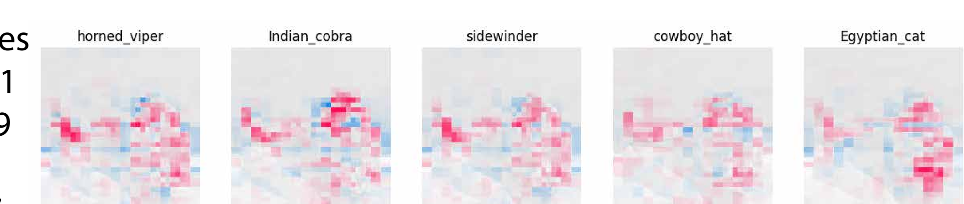
Model is correct and explanations agree why



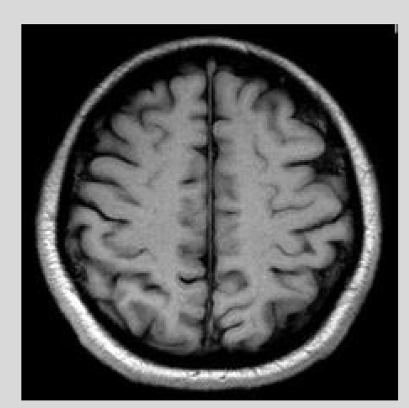
Model is correct, but explanations differ

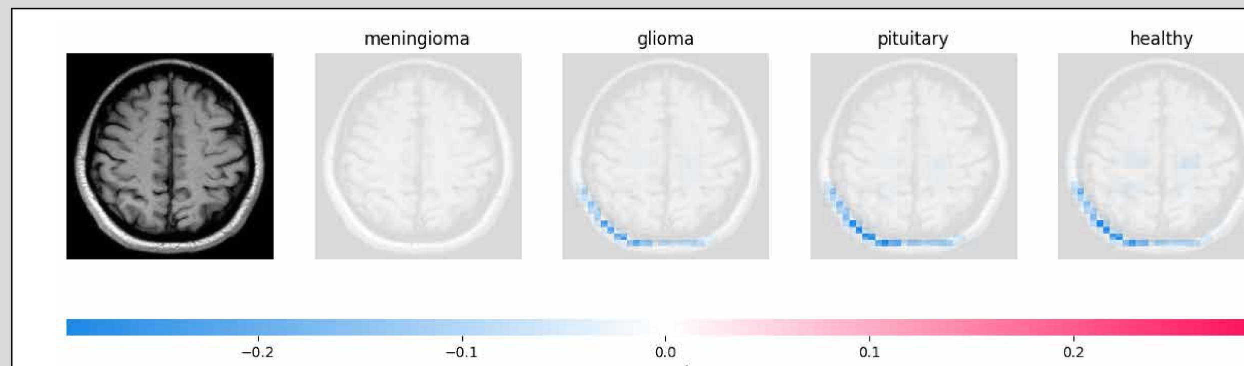


Model Confidences
Horned Viper: 0.21
Indian Cobra: 0.09
Sidewinder: 0.08
Cowboy Hat: 0.07
Egyptian cat: 0.06
...



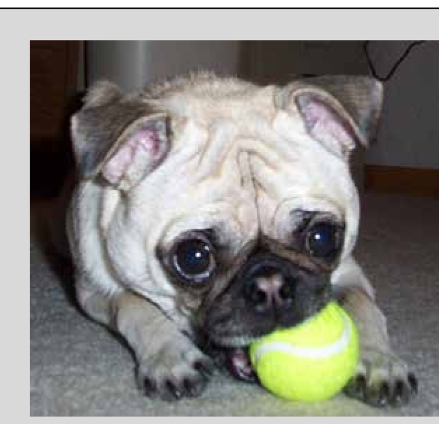
Cannot explain confidences





Model Confidences
Polar Bear: 0.56
Golden Retriever: 0.17
...

Sometimes the explanations of the same prediction disagree greatly



Model Confidences
Tennis Ball: 0.59
Pug: 0.37
...

Different explanations may be due to human-set values, like confidence thresholds

References & Acknowledgements

Our group was led and mentored by Anna Rafferty. In addition, we drew heavy inspiration from Ribearo et al. for the design of our user study.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778

Besse, Philippe & Castets-Renard, Céline & Garivier, Aurélien & Loubes, Jean-Michel. (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version). 10.13140/RG.2.2.22973.31207.

Further Citations

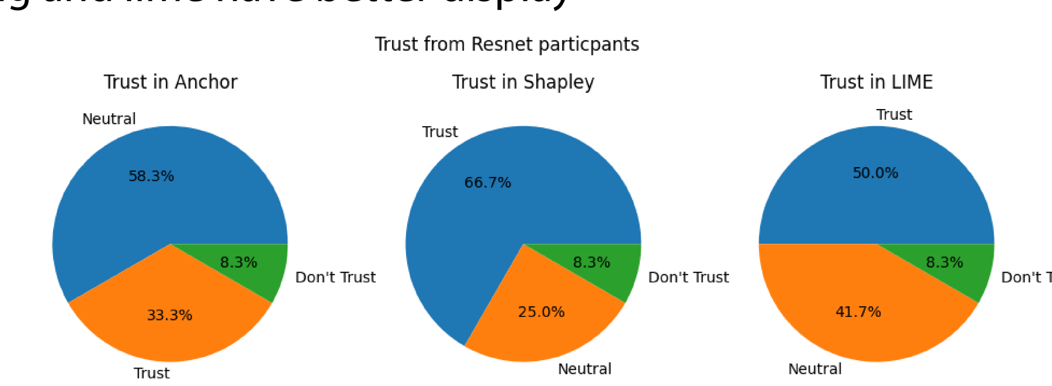
One classic example of the power of explainable AI comes from a model that was trained to differentiate between wolves and huskies. The model had high accuracy, until it was tested in the field. As it turned out, all the images of wolves in the training data had snow in them as well, so the model had only learned to recognize snow. (Besse et al.)

User Study

We asked 12 Carleton students to review curated pictures of cats and dogs, along with Resnet's prediction about the subject of the image and our three explanations of the predictions. For each explanation, we asked how it changed their understanding of the model's thought process, and which visualizations they preferred. Overall, we aimed to determine which of these techniques – applied "out of the box" to image data – were most useful at building understanding of the model.

Findings

Shapley was the favorite technique, but not by a consistent margin.
"If Shapley always had the edge in accuracy I would choose that, but anchoring and lime have better display"



There were a few classes of problems that the techniques couldn't explain. (Regarding the third image from the to) "I can see why the algorithm would think its a snake, but can't see why it doesn't think its a cat"

Participants' CS experience levels did not strongly correlate with preferred techniques, but those with more experience were much less accepting of the discrepancies between different techniques.
"If I was given any single one, I'd say they help, but they confuse in total"

Conclusions

- These techniques can be applied, with little customization, to a range of AI models and on both numerical/image data.
- However, without a preexisting understanding of the data used by the model, these techniques may only confirm biases or confuse.
- Confirmation bias lessens the ability of these techniques to report model failures.
- The explanations are unable to compare confidences between different classes, explore global data, or find where the machine learned its judgements.
- Despite the limitations, they do provide more insight than the black box.

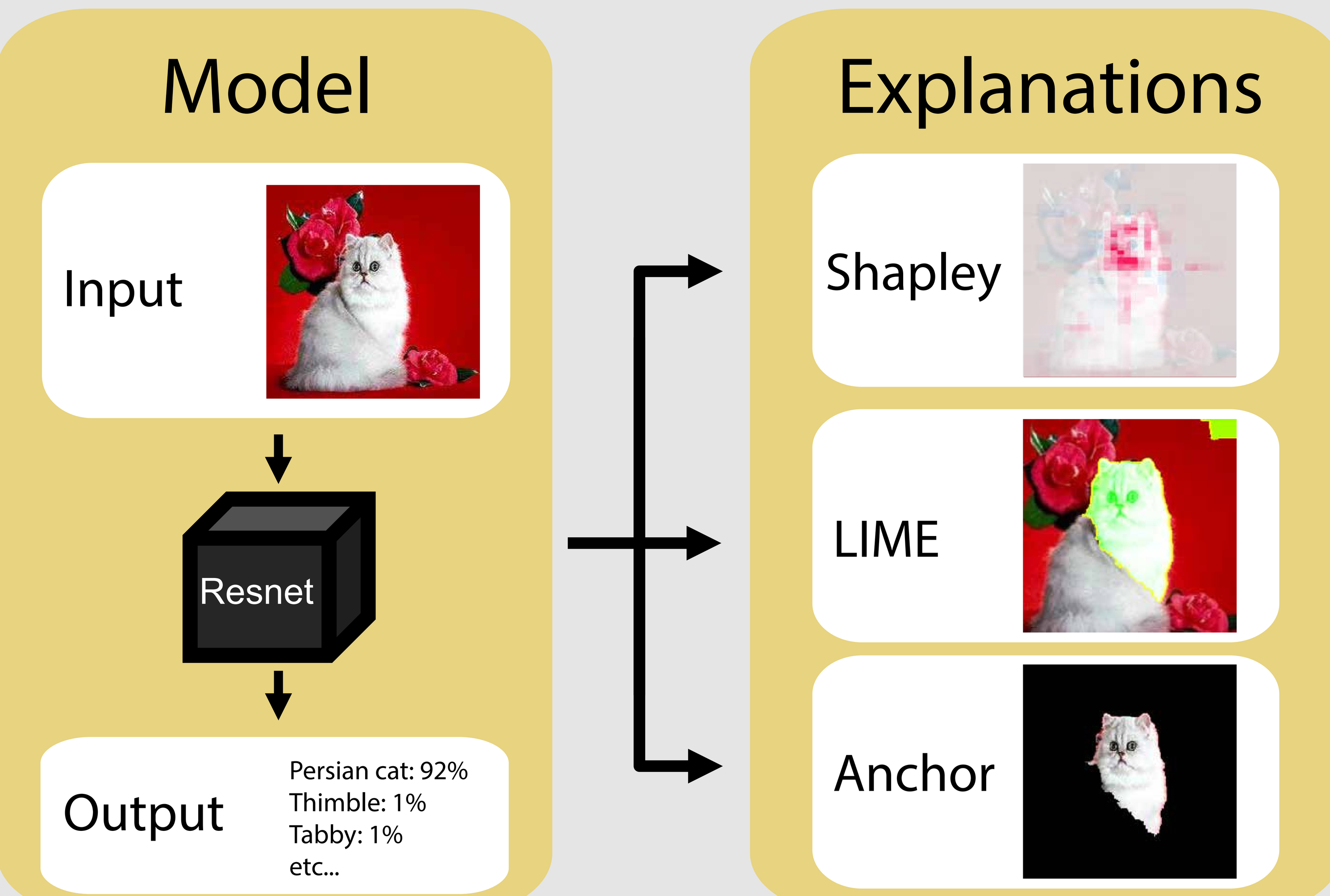
Explainable AI: Breaking Down the Black Box

Josh Moore, Adrian Boskovic, Chris Melville,
Lev Shuster, Sam Johnson-Lacoss, Thomas Pree

Introduction

Machine learning models are more capable than ever before of performing complicated tasks like pattern recognition. But with their growth has come an increased need for transparency.

Transparency is important because these models are being used in sensitive areas, and we want to know if our models are biased or are learning the wrong things. In the EU, users are even guaranteed an explanation of any algorithm that significantly affects them. But for image models, what does an "explanation" even mean?



Shapley

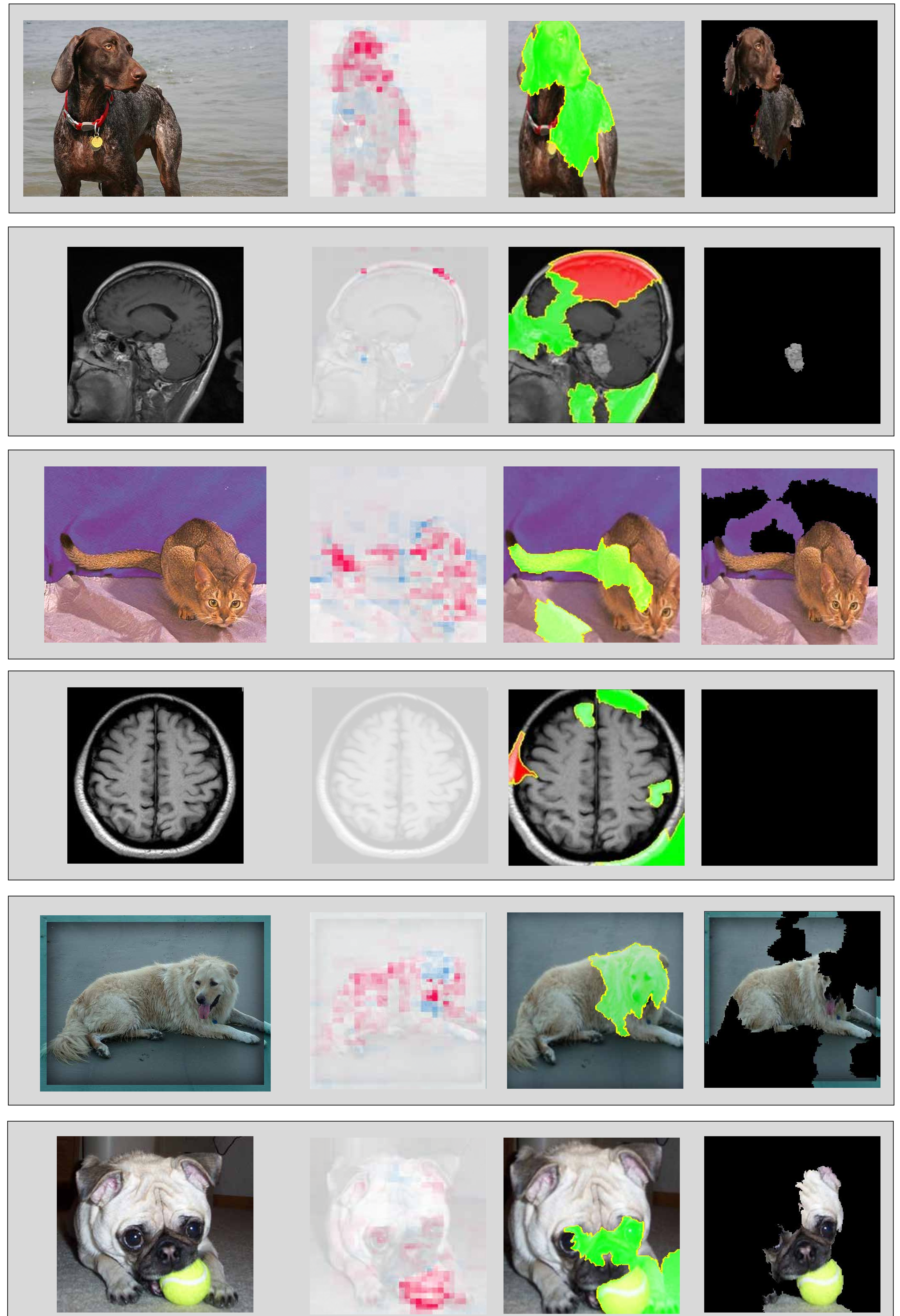
Shapley values, originating from cooperative game theory, treat each group of pixels like a player and class confidences like a game. Each "player's" contribution is measured by the confidence of the model with random subsets of the other pixels excluded. Those that were useful in determining the model's class output are colored red, while those that counted against are blue.

LIME

LIME separates the image into a few constituent "superpixel" groups, with similar colors being grouped together. Then, by toggling each one off in turn, it ascertains which ones make the model's confidence change fastest. Sections that the model is confused without are marked in green, while those that the model is surer without (if any) are marked in red.

Anchoring

Anchoring, similarly, tries to find which superpixels are the biggest determiners of the model's confidence. Upon finding them, it blacks out the least important superpixels until the model's remaining confidence in the prediction drops to a preset threshold.



References & Acknowledgements

Our group was led and mentored by Anna Rafferty. In addition, we drew heavy inspiration from Ribearo et al. for the design of our user study.
Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/2939672.2939778
Besse, Philippe & Castets-Renard, Céline & Garivier, Aurélien & Loubes, Jean-Michel. (2018). Can Everyday AI be Ethical? Machine Learning Algorithm Fairness (english version). 10.13140/RG.2.2.22973.31207.

Further Citations

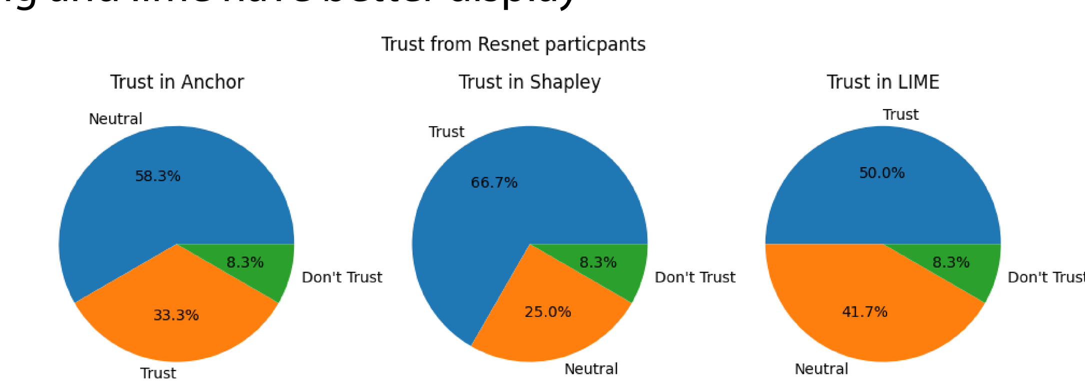
One classic example of the power of explainable AI comes from a model that was trained to differentiate between wolves and huskies. The model had high accuracy, until it was tested in the field. As it turned out, all the images of wolves in the training data had snow in them as well, so the model had only learned to recognize snow. (Besse et al.)

User Study

We asked 12 Carleton students to review curated pictures of cats and dogs, along with Resnet's prediction about the subject of the image and our three explanations of the predictions. For each explanation, we asked how it changed their understanding of the model's thought process, and which visualizations they preferred. Overall, we aimed to determine which of these techniques – applied "out of the box" to image data – were most useful at building understanding of the model.

Findings

Shapley was the favorite technique, but not by a consistent margin.
"If Shapley always had the edge in accuracy I would choose that, but anchoring and lime have better display"



There were a few classes of problems that the techniques couldn't explain. (Regarding the third image from the to) "I can see why the algorithm would think its a snake, but can't see why it doesn't think its a cat"

Participants' CS experience levels did not strongly correlate with preferred techniques, but those with more experience were much less accepting of the discrepancies between different techniques.
"If I was given any single one, I'd say they help, but they confuse in total"

Conclusions

- These techniques can be applied, with little customization, to a range of AI models and on both numerical/image data.
- However, without a preexisting understanding of the data used by the model, these techniques may only confirm biases or confuse.
- Confirmation bias lessens the ability of these techniques to report model failures.
- The explanations are unable to compare confidences between different classes, explore global data, or find where the machine learned its judgements.
- Despite the limitations, they do provide more insight than the black box.