Err on the Side of Texture: Texture Bias on Real Data

Blaine Hoak, Ryan Sheatsley, Patrick McDaniel

University of Wisconsin-Madison

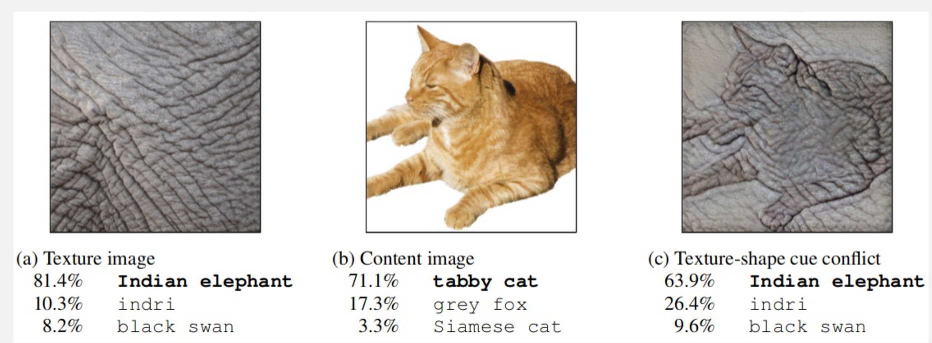
MADS&P

INTRODUCTION

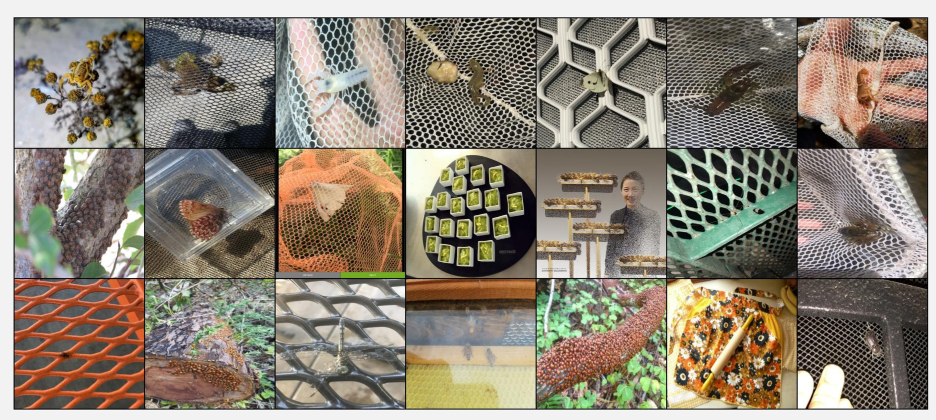
Texture Bias

- Bias serves as a core contributor of poor accuracy and trustworthiness in machine learning models.
- One such bias is *texture bias* – where models strongly rely on texture, rather than shape, when classifying images.
- Existing approaches have not yet been able to capture how naturally occurring texture information influences model classifications.
- We hypothesize that textures serve as a primary and *necessary* signal for driving classification on real data.
- Here, we propose new metrics for quantifying the effect of texture bias on model accuracy, confidence, and robustness.

Geirhos et al. ICLR 2019

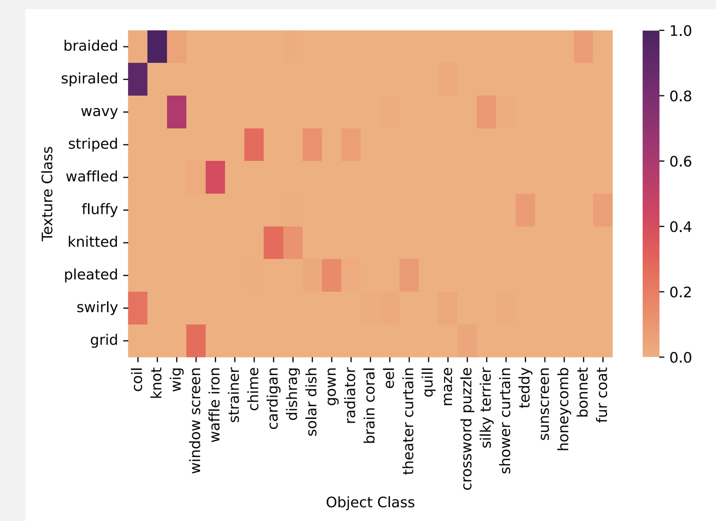


Natural adversarial examples classified as honeycombs



METHODS

Texture Association and Identification

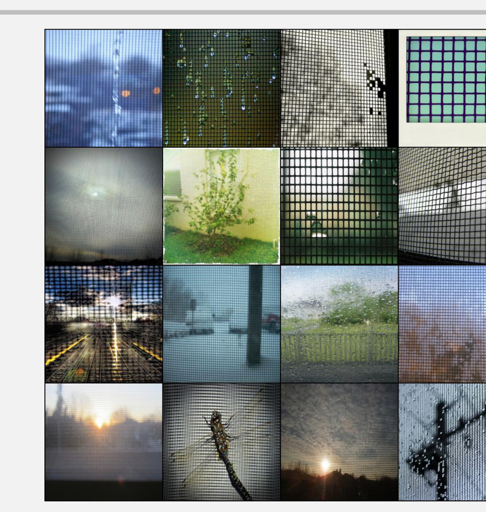


Texture Object Association Value (TAV) quantifies the relationship between textures and the object classes a model predicts, which is captured by analyzing model predictions on texture data.

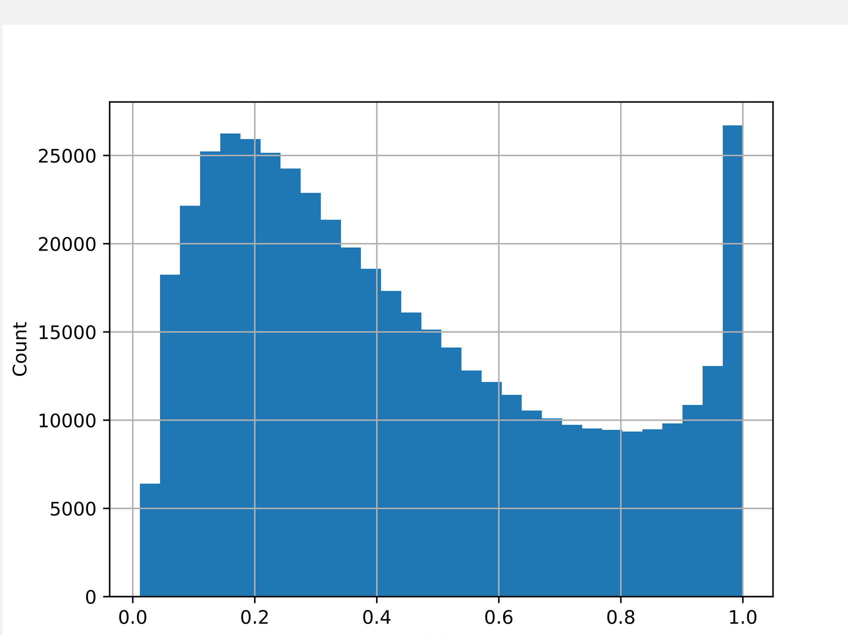
$$\mathrm{TAV}_{ij} = PT_{ij} \cdot (1 - TH_i) \cdot PO_{ij} \cdot (1 - OH_j)$$

Using *TAV*, we identify textures present in real images by comparing similarity between response to textures and validation data.

$$\mathrm{TID}(x) = \arg\max_i \frac{f_\theta(x) \cdot \mathrm{TAV}_i}{\|f_\theta(x)\| \cdot \|\mathrm{TAV}_i\|}$$



RESULTS

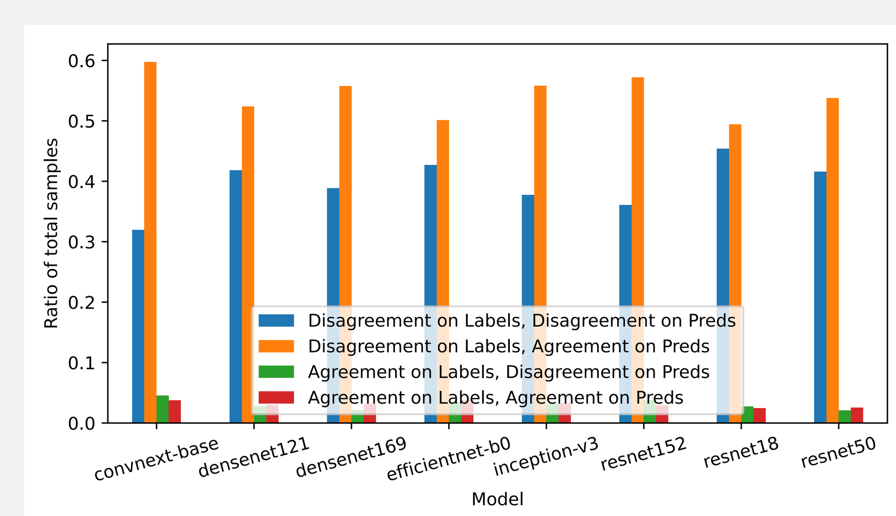


Textures are Predictive Features

Analyzing how models respond to textures alone, we find that texture images are classified as objects at rates *far* above random guessing (0.001).

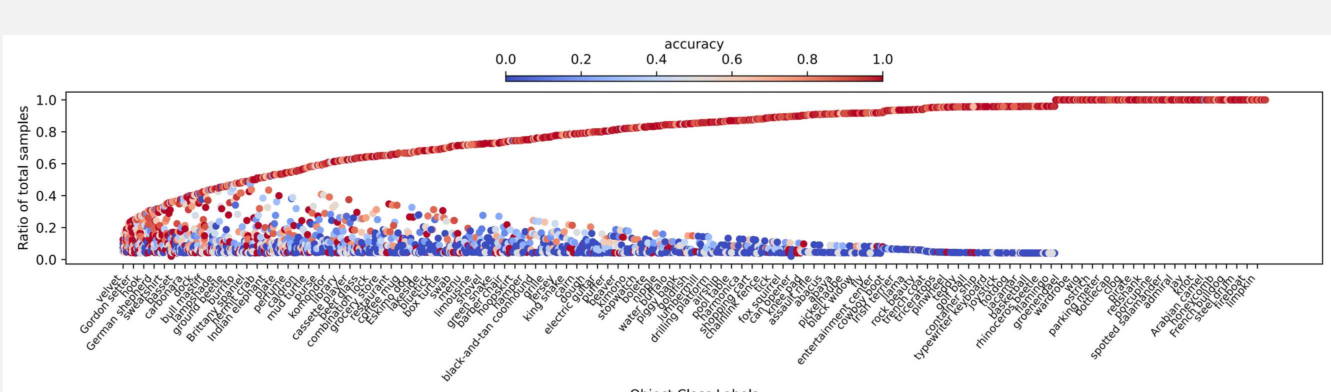
We also find that many images are even classified at or close to 100% confidence, even though these samples are OOD and missing all object information.

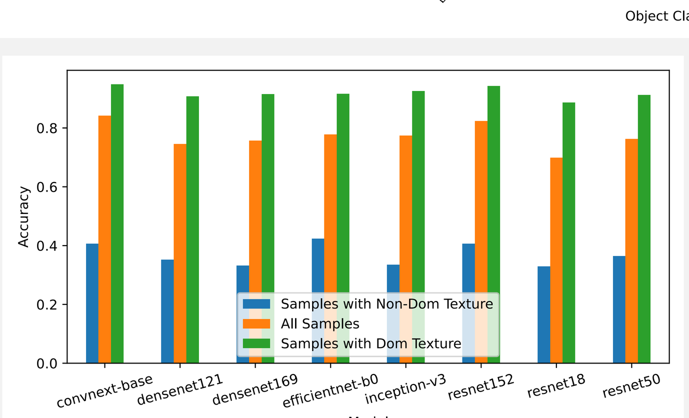
Natural Adversarial Examples



Looking at natural adversarial examples – samples that models are confidently incorrect on – we find that in up to 90% of these images contain textures that disagree with the dominant texture of their true label. This suggests that texture misalignment can explain confident mispredictions.

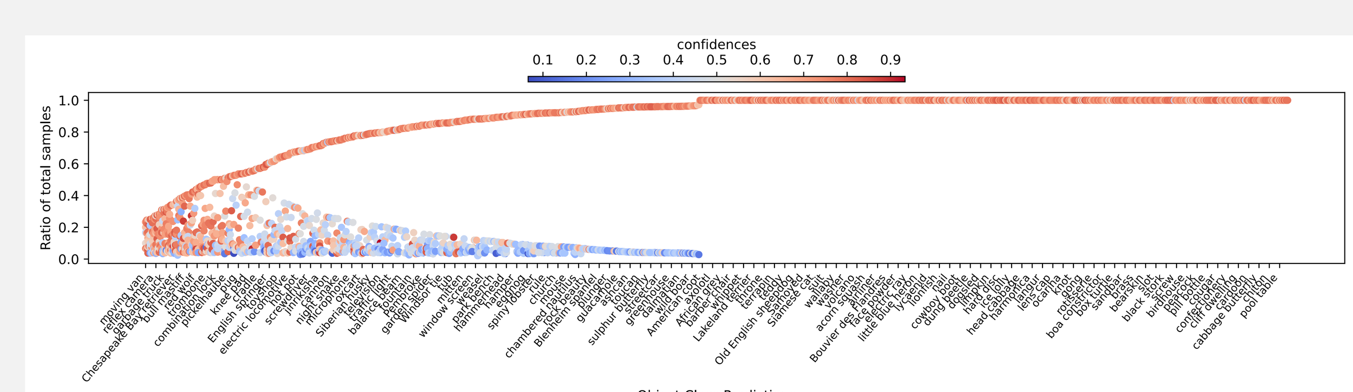
Accurate, Confident Classifications Require Texture

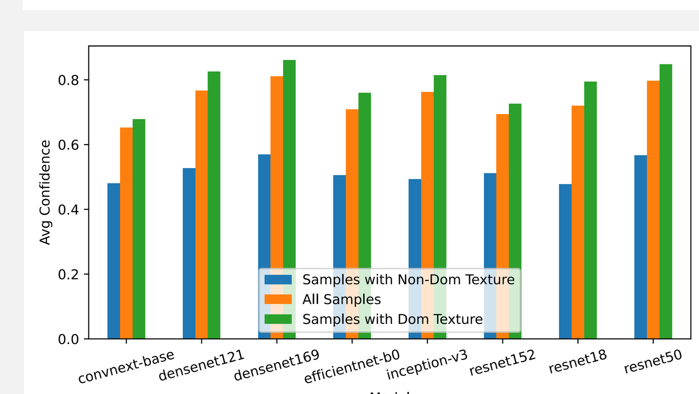




Models are up to 66% more accurate on samples that contain the dominant texture of the object class than those that contain a different texture.

Takeaway: The separation in both model accuracy and confidence between samples containing different textures highlights that models *learn* and *rely on* the presence of specific textures.





Models are up to 40% more confident on samples that contain the dominant texture of the object class than those that contain a different texture.

CONCLUSION

https://hoak.me
@blaine_hoak
blainehoak
bhoak@cs.wisc.edu

Summary

- We find that textures are highly predictive features that models learn and rely on when classifying objects. This bias towards texture departs from human visual processing and undermines model trustworthiness.
- The presence of specific textures in an image can determine how accurate and confident a model is, showing that texture bias plays a key role on real data classifications.
- Model robustness is influenced by textures. Confident mispredictions can be explained by the fact that these images contain textures not associated with their label.

Future Work

- Further investigation into the interplay between security and texture bias is needed – specifically with how robust models may differ in their reliance on textures and how other security phenomena may be explained by texture bias.
- In some cases, texture may be a truly necessary feature to learn, future work should aim to uncover when texture bias may be desirable or disastrous.