Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)

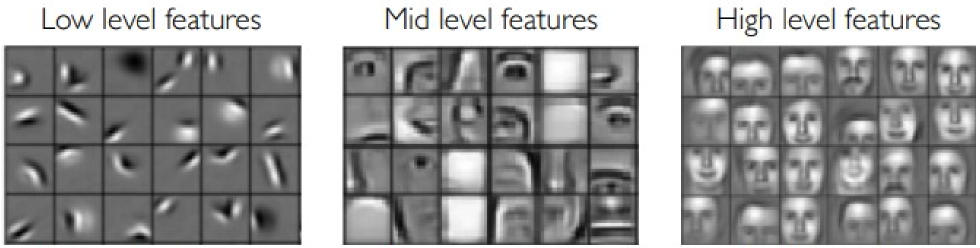
Kim et al. (2018)

Presented by Lucia Gordon, Matthew Nazari, Catherine Yeh

Thoughts on Concept-Based Explanations So Far?

Motivation + Problem Statement

- Interpreting deep learning models is crucial to understanding their behavior, ensuring accurate predictions, and reflecting our values
- But remains a big challenge due to size, complexity, and opacity of ML models
- Additionally, many systems operate on **low-level features** (e.g., pixel values) rather than **high-level concepts** (e.g., face) that are human-interpretable



Motivation + Problem Statement

Image from Been Kim

Problem:

- We can't express these concepts as pixels
- And they weren't our input features

Summary of Contributions

- Introduce **Concept Activation Vectors (CAVs):** way to interpret a neural network's internal state in terms of human-friendly concepts
- Key idea is to use the high-dimensional internal state of a neural net as an aid not an obstacle
- Main contribution is a new linear interpretability method, **Testing with CAV (TCAV)**, that quantifies model sensitivity to a high-level concept learned by a CAV for a particular class

E.g., "striped"

E.g., "zebras"

Goals of TCAV

- **Accessibility:** requires little to no ML expertise
- **Customization:** adaptable to any concept, even outside of training
- **Plug-in readiness:** works without retraining/modifying ML models
- **Global quantification:** can interpret entire classes with a single quantitative measure

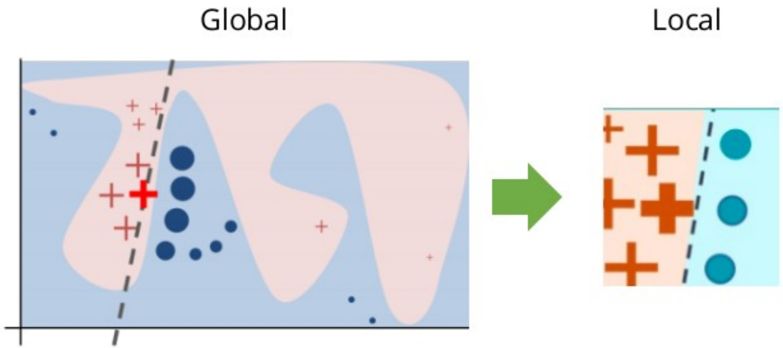
Assessed with experiments + human evaluation

Related Work

Interpretability methods: (Goodman & Flaxman, 2016)

- *Inherently interpretable* models vs. *post-hoc explanations*
  (Kim et al., 2014; Doshi-Velez et al., 2015)

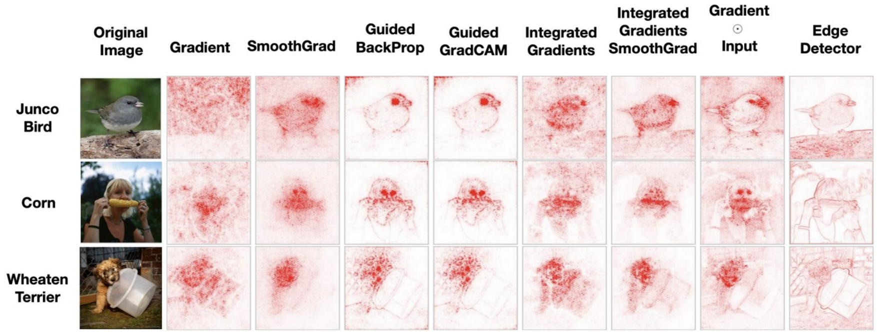
- **Perturbation-based methods:** e.g., LIME/SHAP
  - *local* vs. *global* (Ribeiro et al, 2016; Lundberg & Lee, 2017)



Related Work

Interpretability methods in neural networks

- Limitations w/ **saliency methods**
  - Local explanation (Erhan et al., 2009; Smilkov et al., 2017)
  - Lack customization
  - Vulnerable to adversarial attacks (Ghorbani et al., 2017)
  - Insensitivity to randomization (Adebayo et al., 2018)



Related Work

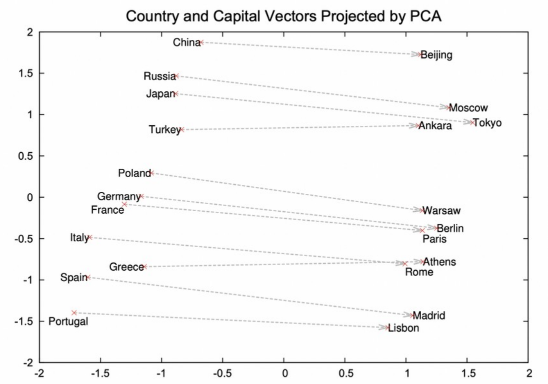
Linearity in neural network + latent dimensions

- Meaningful information can be learned from simple *linear* classifiers

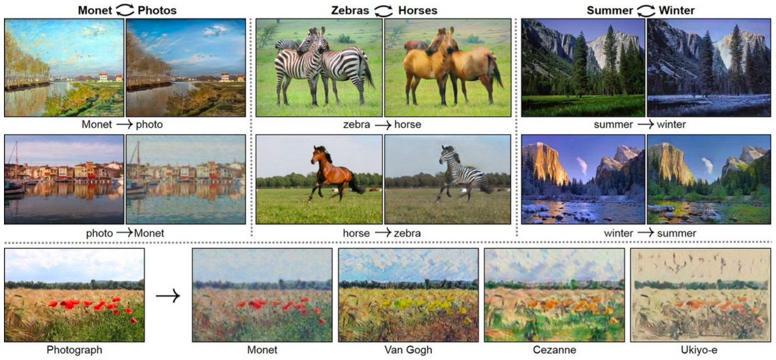
(Bau et al., 2017; Alain & Bengio. 2016)

- Mapping *latent* dimensions to human *concepts*

(Mikolov et al., 2013; Zhu et al., 2017)



word2vec



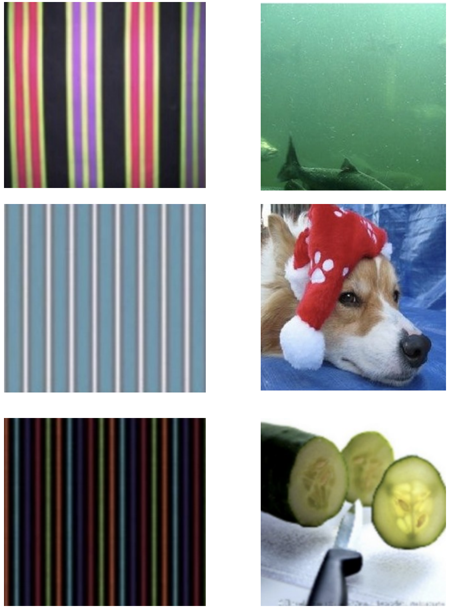
cycleGAN

Approach: Defining the CAV

Consider the fully connected layer $f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and the concept of interest C

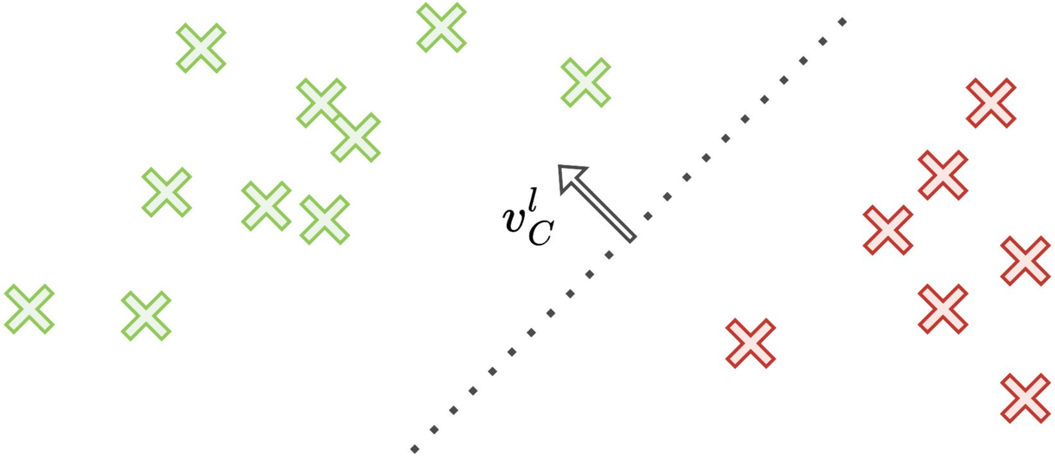
Collect a set of examples P_C of that concept and a negative set of examples N that don't

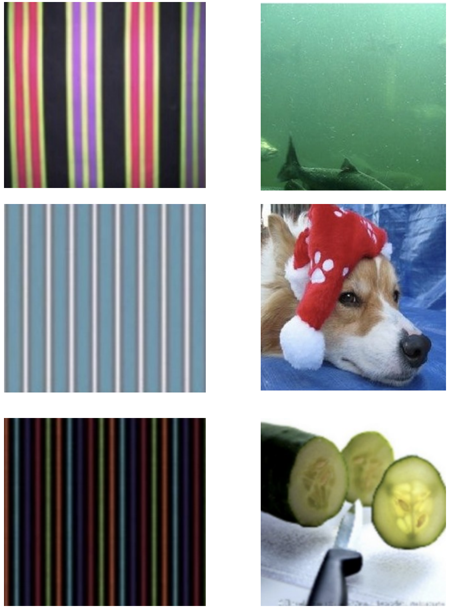
Define the CAV to be a vector orthogonal to a decision boundary between activations $\{f_l(\boldsymbol{x}) : \boldsymbol{x} \in P_C\}$ and $\{f_l(\boldsymbol{x}) : \boldsymbol{x} \in N\}$



Stripes P_C **Not stripes** N

Approach: Visualizing the CAV





Stripes P_C

Not stripes N

$\in \{f_l(x) : x \in P_C\}$

$\in \{f_l(x) : x \in N\}$

Approach: Gauging “concept sensitivity”

Saliency maps gauge sensitivity of prediction $h_k(\boldsymbol{x})$ with respect to per-pixel perturbations

$$S_{C,k,l}(\boldsymbol{x}) \quad = \quad \nabla h_{l,k}(f_l(\boldsymbol{x})) \cdot \boldsymbol{v}_C^l$$

With CAV, we can gauge sensitivity of predictions $h_{l,k}(f_l(\boldsymbol{x}))$ **towards a concept** $\boldsymbol{v}_C^l$ at an entire layer

Approach: Testing with CAV (TCAV)

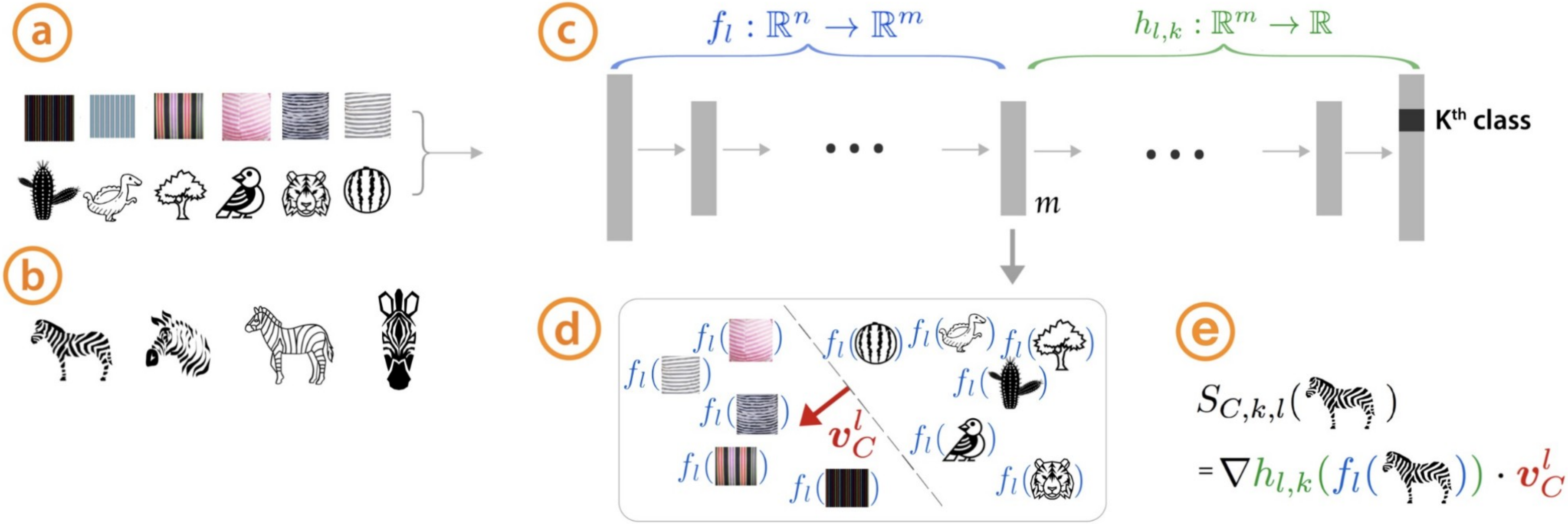
$$\text{TCAVQ}_{C,k,l} = \frac{|\{\boldsymbol{x} \in X_k : S_{C,k,l}(\boldsymbol{x}) > 0\}|}{|X_k|}$$

$\text{TCAVQ}_{C,k,l}$ measures the fraction of inputs whose activations were influenced by a concept

This provides interpretation global to a particular class

A *t*-test can safeguard against meaningless CAVs

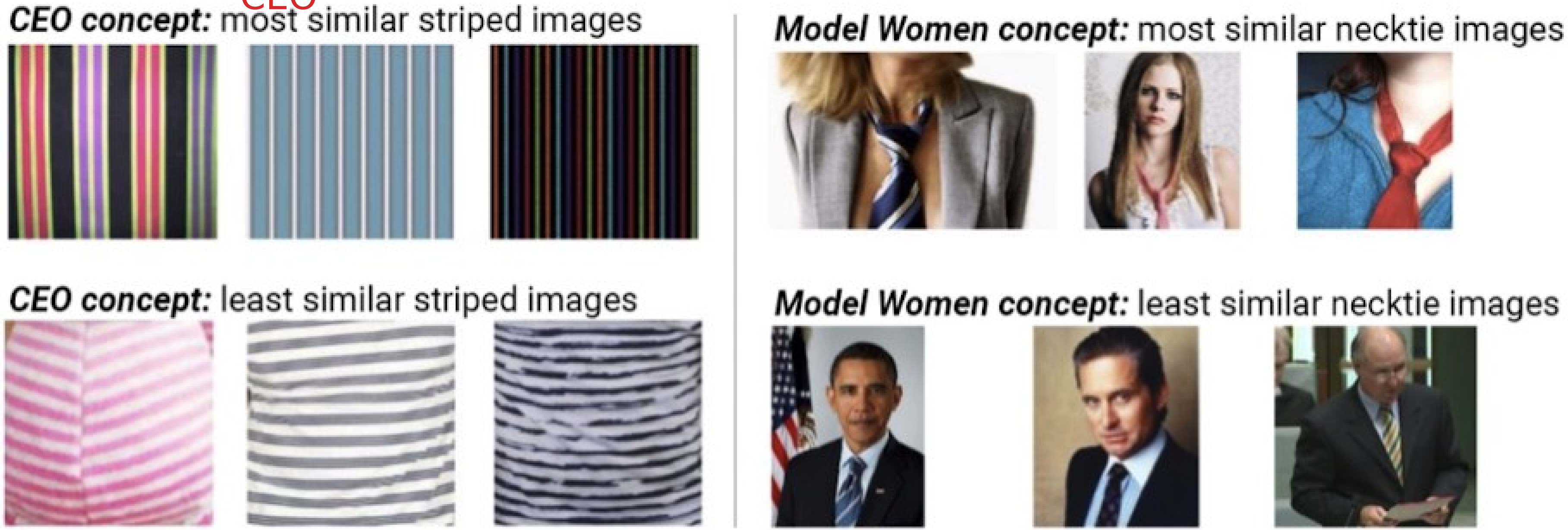
Approach Summary



Results: Sorting Images with CAVs

Class: Stripes
Concept: CEO

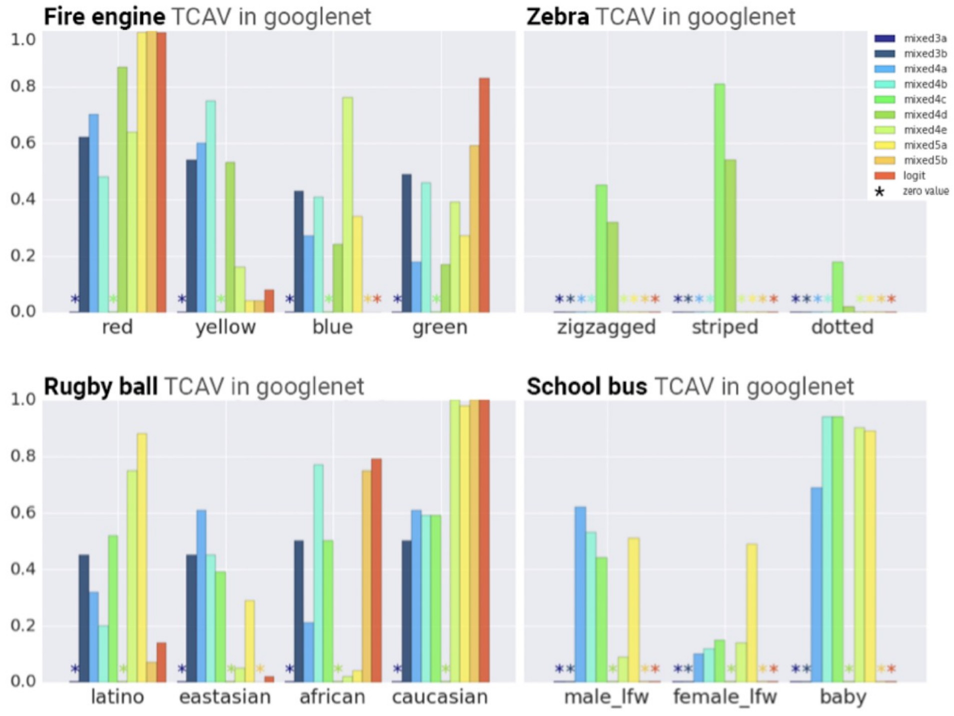
Class: necktie
Concept: model woman



- Confirmation that the CAVs correctly reflect the concept of interest
- Sorting procedure can reveal biases used to learn the CAV

Results: Gaining Insights with TCAV

- x-axis: CAV
- y-axis: TCAV score = conceptual sensitivity of classification to concept
- Each bar is a different layer of the NN with a statistically significant CAV
- Layers closer to logit layer have a greater influence on the prediction

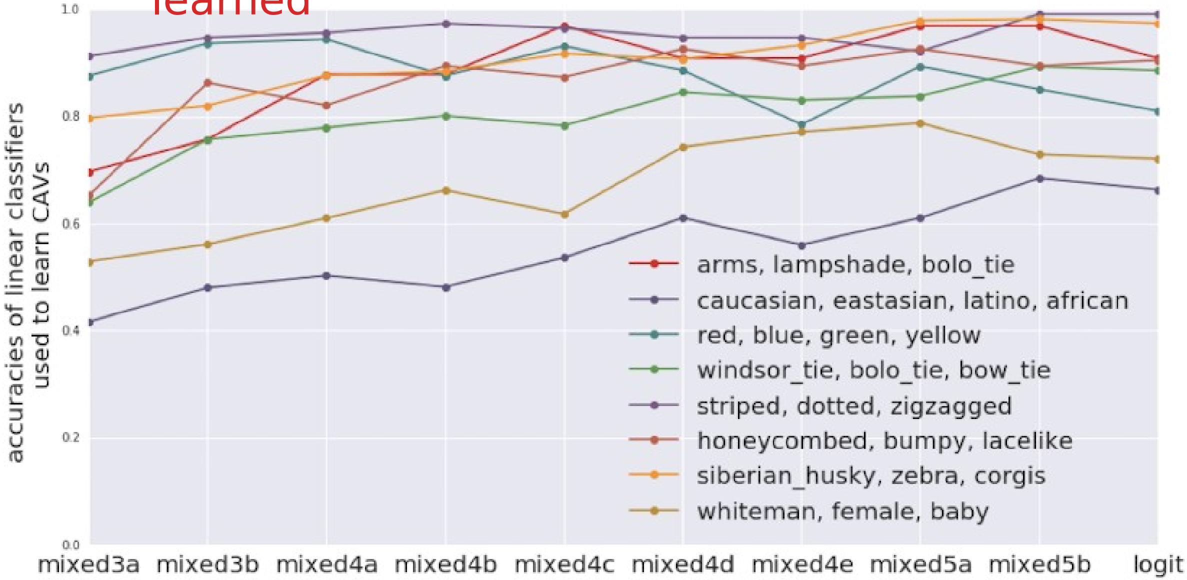


- Matches intuition
  - “Red” is important for “fire engine”
  - “Striped” is important for “zebra”
- Biases
  - “Caucasian” is important for “rugby ball”
- Statistical significance test successfully removed spurious CAVS
  - “Dotted” is not important for “zebra”
- Quantitatively confirm qualitative

Results: TCAV for Where Concepts are Learned

- Simple concepts (ex. colors, patterns) reach a high accuracy at low layers
- Complex concepts (ex. age, sex, objects) don’t reach high accuracy until higher layers
- Confirming past findings that lower layers = feature detectors and higher layers = classifiers

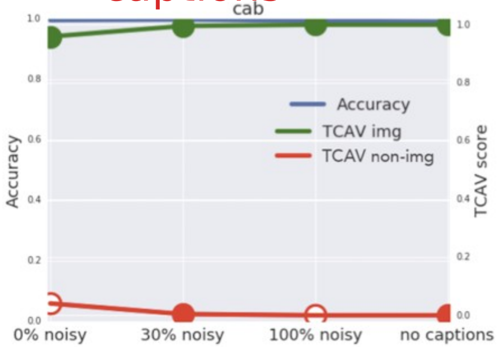
Accuracies of different CAVs at different layers, showing where each concept is learned

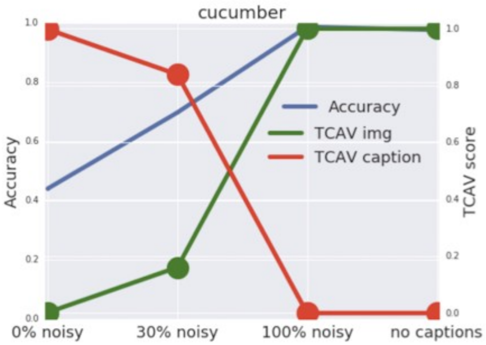


Results: Controlled Experiment with Ground Truth

cab image | cab image with caption | cucumber image | cucumber with caption

- TCAV img → image concept CAV
- TCAV non-img → caption concept CAV
- Accuracy → tested on images without captions

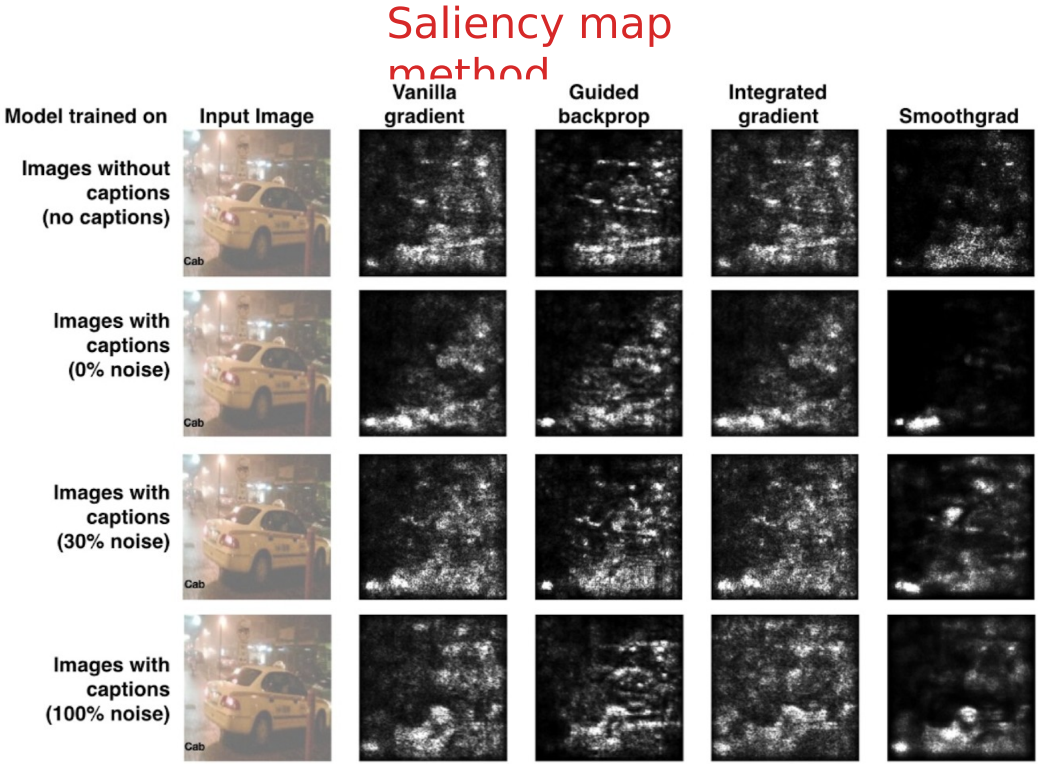




- Cab: image concept more important than caption concept regardless of noise
- Cucumber: caption concept more important when caption is likely to appear
- TCAV reflects ground-truth
  - Only image is important → high accuracy
  - Only caption is important → low accuracy

Results: Evaluation of Saliency Maps

We know that for “cab” the “image” concept is most important, but this is not reflected in saliency maps → superiority of TCAV interpretability method

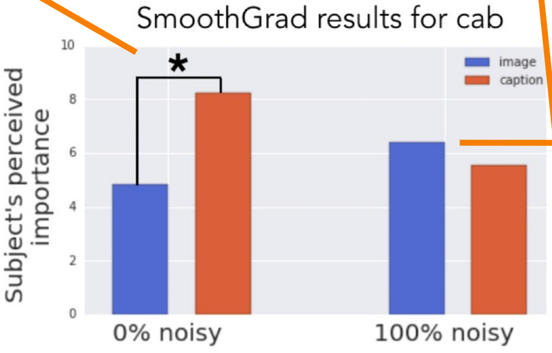


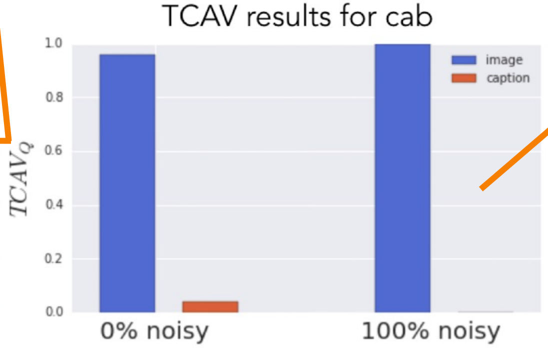
Results: Evaluation of Saliency Maps with Human Subjects

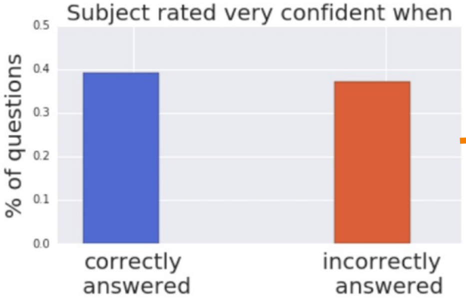
Subjects could not discern a difference in importance between the image and caption

Subjects incorrectly thought the caption was more important than the image

TCAV score correctly reflects which concept is most important





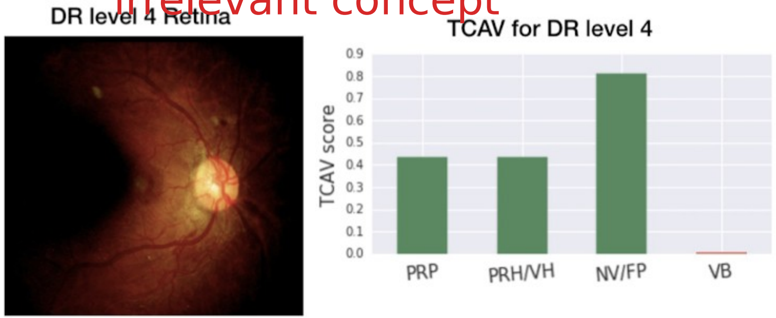


Saliency maps are misleading!

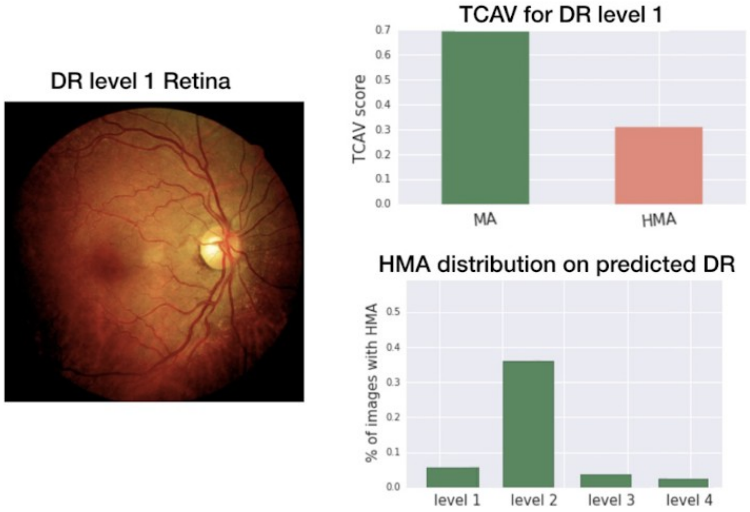
Results: TCAV for a Medical Application

Green = relevant concept

Red = irrelevant concept



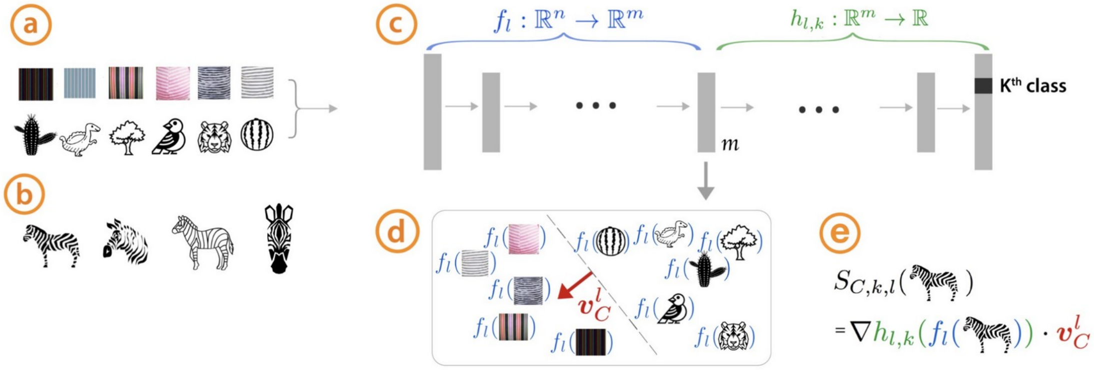
TCAV score shows model successfully distinguishes relevant and irrelevant concepts for level 4 diagnosis



TCAV score shows model is giving too much importance to HMA concept for level 1 diagnosis; HMA is relevant to level 2 not level 1 → use this to debug model

Conclusions

- **Roadmap:** Gradient → Attention → Concept based approaches (TCAV!)
- **Limitations**
  - Evaluated only on computer vision tasks
  - Statistical significance testing — is this rigorous?



Discussion Questions

- How does this method compare to e.g., saliency maps?
- How would you adapt TCAV for other types of data (e.g., audio/video) and what would that look like?
- Thoughts on TCAV for adversarial example identification (Appendix A)?
- Do you think there could be adversarial images that allow meaningless CAVs to pass the statistical significance test?

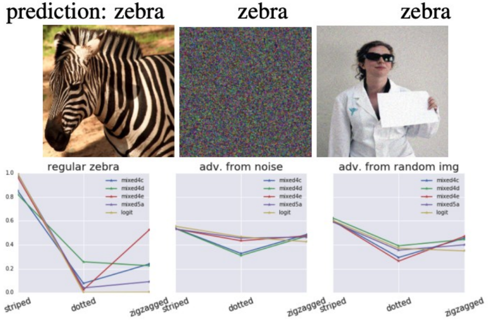


Figure 11. Two types of adversarial images that are classified as zebra. In both cases, the distribution of TCAVQare different from that of a normal zebra.