

紹介論文

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

Been Kim and Martin Wattenberg and Justin Gilmer and Carrie Jun Cai and James Wexler and Fernanda Viegas and Rory Abbott Sayres

In *Proceedings of the 35th International Conference on Machine Learning*, pages 2668-2677, 2018
(ICML2018)

1 In a word this paper

ある画像 (動物, 乗り物, ...) に対して, 概念 (縞模様, ...) の重要度の評価手法を提案

2 Motivation

機械学習モデルは, 人間が簡単に理解できるような概念 (Concept) を用いているわけではなく, ピクセル値のような特徴を元にして学習および推論を行う. したがって, モデルの中間表現は人間が理解することのできないものである. ここで, 機械学習モデルの状態を空間 E_m と表す. 一方で, 人間は自身が理解できる概念に対応する空間 E_h とする. すると, モデルの解釈性は関数 $g : E_m \rightarrow E_h$ と見なすことができる. そのような g を構築して, 学習モデルに説明性を与えることを目標とする. ここで, g が線形であるとき, linear interpretability と呼ぶ. 本論文では, そのような liner interpretability な新たな手法を提案する.

3 Goal

- Accessibility : 機械学習の専門知識をほとんど必要としない
- Customization : あらゆる概念 (性別, ...) に適応できる
- Plug-in readiness : 学習モデルの再学習や変更を必要としない
- Global quantification : 個別のデータに対する説明ではなく, データセット全体を单一の指標で評価することができる

4 Concept Activation Vector (CAV)

- E_m と E_h の間を解釈する手法
- 概念のデータセット (concept example) 方向の方向ベクトル
- CAV は concept example と random counterexample との線形分類子の訓練を行い, 決定境界に直交するベクトルをとることで CAV を導く
- このシンプルなアプローチは, local linearity [1, 2, 3] に基づいている

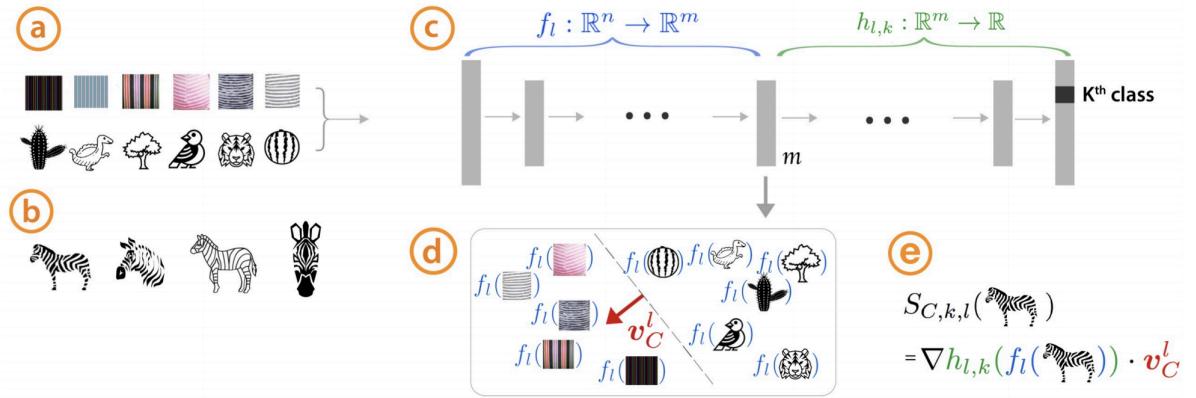


Figure1 Testing with Concept Activation Vectors

5 Testing with CAV (TCAV)

- CAV を用いた liner interpretability な手法 (Figure 1)
- 方向微分を使用して, CAV を用いて, 概念に対する予測の sensitivity を計算する
- 例えば, シマウマを識別するモデルと縞模様の concept example が与えられたとき, TCAV はシマウマの予測に対する縞模様の影響度を单一の値で定量化することができる
- モデルの出力クラスまたは, 状態値を有意かつ相関があると示さない限り, CAV がランダムに再学習または, reject される統計的な検定を行なっている (In addition, we conduct statistical tests where CAVs are randomly re-learned and rejected unless they show a significant and stable correlation with a model output class or state value.)

6 Method

CAV の算出方法および, 概念に対する重要度の値 $TCAV_Q$ を算出する手法を示す.

6.1 Preparation

- (a) ある概念 C を表現するデータセット (concept example) と概念 C を含まないランダムな画像データセット (random counterexample)
- (b) ラベル付された学習データ
- (c) 訓練済みネットワーク
 - input $\mathbf{x} \in \mathbb{R}^n$
 - output : K class
 - network : $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$ (feedforward layer 1 with m neurons: $f_l : \mathbb{R}^n \rightarrow \mathbb{R}^m$, logit from layer 1 to layer k : $f_{l,k} : \mathbb{R}^m \rightarrow \mathbb{R}$)

6.2 CAV and TCAV

潜在空間 f_l 上で概念を表すベクトルを見つける。concept example と random counterexample を分ける hyperplane の法線として CAV を定義する。

1. concept example を Positive set (P_C), random counterexample を Negative set (N) に分ける
2. ネットワークの 1 層の潜在空間上で 2 つ ($\{f_l(\mathbf{x}) : \mathbf{x} \in P_C\}$ and $\{f_l(\mathbf{x}) : \mathbf{x} \in N\}$) の線形分離を行う
 - $\mathbf{v}_C^l \in \mathbb{R}^m$ は concept C に対する liner CAV である
3. concept C の class k の sensitivity を以下のように定義

$$S_{C,k,l}(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \frac{h_{l,k}(f_l(\mathbf{x}) + \epsilon \mathbf{v}_C^l) - h_{l,k}(f_l(\mathbf{x}))}{\epsilon} = \Delta h_{l,k}(f_l(\mathbf{x})) \cdot \mathbf{v}_C^l \quad (1)$$

4. Testing with CAVs (TCAV) のスコアを以下のように定義

$$\text{TCAV}_{Q_{C,k,l}} = \frac{|\{\mathbf{x} \in X_k : S_{C,k,l}(\mathbf{x}) > 0\}|}{|X_k|} \quad (2)$$

where X_k is all inputs with that given label

$\text{TCAV}_{Q_{C,k,l}}$ は、 $S_{C,k,l}$ の符号のみに依存しているため、全ての入力に対して、概念に対する sensitivity を global に評価することができる。

7 Statistical test

無意味な CAV を学習ことを防ぐために、統計的有意検定を提案する。random examples N の単一なバッチで CAV の一度の学習の代わりに、複数の学習 (typically 500) を実行する。意味のある概念は、学習全体で一貫した TCAV score を示すはずである。

複数サンプルに基づいた TCAV スコアの両側 t-test を行う。TCAV score が 0.5 であるという帰無仮説を棄却できる場合、生じた概念は、有意に予測に関連していると主張することができる。また、仮説に対して Bonferroni correction を行い、誤検出率を制御している (at $p < \alpha/m$ with $m = 2$)。

8 Relative TCAV

実際には意味的に関連する (茶髪 vs 黒髪) は、直交から離れた CAV を生成する。Relative CAVs はこのような比較を行うことを可能にする。ここで、2 つの異なる概念を C, D とする。 $f_l(P_C)$ および $f_l(P_D)$ で訓練すると法線ベクトル $\mathbf{v}_{C,D}^l \in \mathbb{R}^m$ が得られる。 $\mathbf{v}_{C,D}^l$ は、第 l 層の 1-d 次元の subspace を直感的に定義する。この subspace に沿った $f_l(x)$ の射影は、 x が概念 C または D のどちらにより関係しているかを測定する (?).

9 Experiment results

concept C 方向のベクトル \mathbf{v}_C^l と、ある画像データ x_i の特徴量 $f_l(x_i)$ のコサイン類似度を算出して並び替えたのが Figure 2 である。ここで、 x_i は CAV の訓練には使用していないことに注意する。Figure 2 の左図

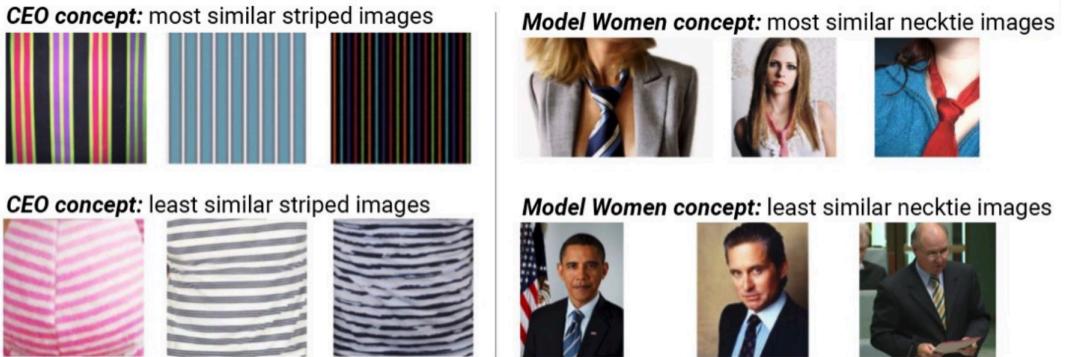


Figure2 The most and least similar pictures of stripes using ‘CEO’ concept (left) and neckties using ‘model women’ concept (right)



Figure3 Empirical Deepdream using knitted texture, corgis and Siberian huskey concept vectors (zoomed-in)

から CEO の concept は、横縞模様よりも縦縞模様であることがわかる。また、右図から女性モデルのイメージにあったネクタイを知ることができる。

次に Deep Dream を用いて、CAV を最も activate するパターンを可視化する。Figure 3 は織物、コギー(犬)とシベリアンハスキー(犬)を可視化したものである。

Figure 4 は、GoogleNet および Inception V3 で 3 つの異なる層に Relative TCAV を適用した結果である。赤色に近いほど上層となっているため、予測に直接的に関わる。例えば、消防車は赤の概念が強いことがわかる。また、エプロンは男性に比べて赤ちゃんの概念の方が強いことがわかる。

Figure 5 は、次に概念の種類ごとに線形分類子 (CAV) を用いて分類を行なったときの精度を示したものである。評価は学習データの 1/3 のテストデータセットで行なっている。ここからより抽象的な概念 (objects) の方が上位層で精度が上昇することがわかる。また、単純な概念 (color) は全体的に精度が高い。

Figure 6 のようなキャプションが画像中に含まれており、正解ラベルが既知である場合にも機能するのかを検証する。Figure 6 のようにキャプション付のタクシーときゅうりの画像にノイズ $p \in [0, 1.0]$ を加える。例えば $p = 0.3$ のとき、正しいキャプションからランダムな言葉に 30% 置き換える。

このデータセットを p を変化させて 4 つのネットワークで学習させる。Ground truth として、キャプション無しの画像のみのデータでの精度と比較を行う。学習結果が 7 である。ここから TCAV はキャプションよりも画像を重要視していることがわかる。

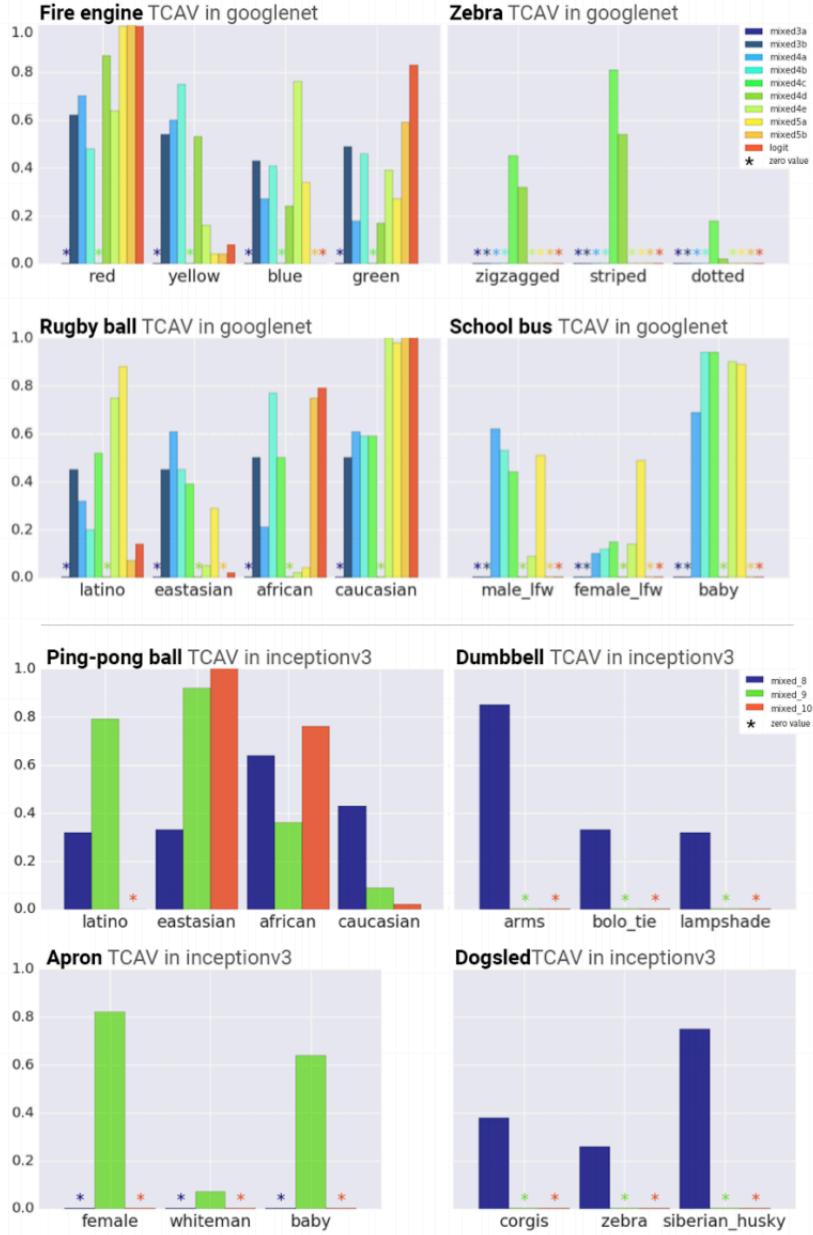


Figure4 Relative TCAV for all layers in GoogleNet and last three layers in Inception V3. TCAVQs in layers close to the logit layer (red) represent more direct influence on the prediction than lower layers. “**” is mark CAVs omitted after statistical testing.

次に医療分野での応用結果を示す。糖尿病網膜症 (diabetic retinopathy, DR) は 4 段階 (Level 0 ~ 4) に分かれている。医師はこの 4 つの段階を microaneurysms (MA), pan-retinal laser scars (PRP) などの概念を元に判断して分類する。ここで、TCAV を用いて各段階の概念の重要度を評価する (Figure 8)。

DR level 1において、TCAV は医師の経験則とは異なる場合がある (Figure 8 bottom)。HMA は比較的に高い DR level の診断結果のはずであるが、DR level 1において HMA は相対的に TCAV が高い。この結果と

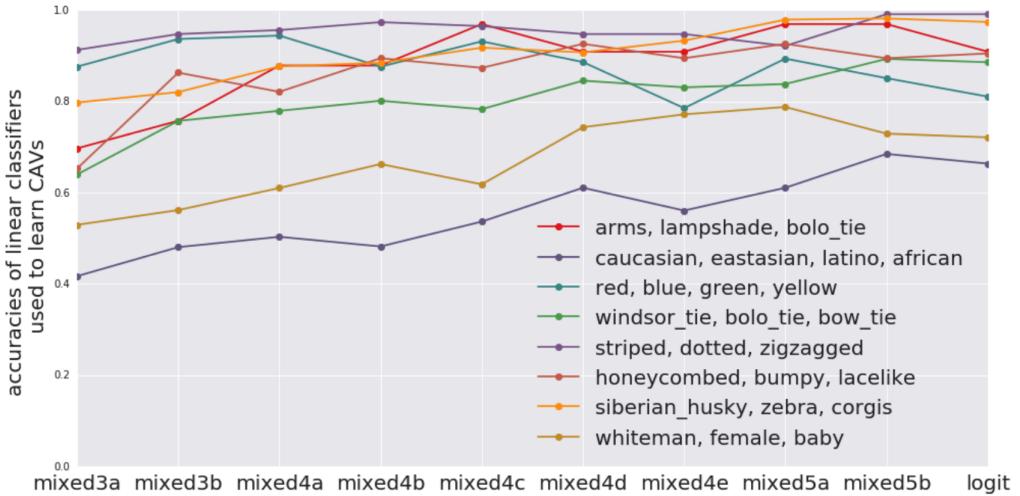


Figure5 The accuracies of CAVs at each layer. Simple concepts (e.g., colors) achieve higher performance in lower-layers than more abstract or complex concepts (e.g. people, objects)



Figure6 A controlled training set: Regular images and images with captions for the cab and cucumber class.

一致して予測モデルは Level 1 を Level 2 と予測してしまうことが多い。

ここから、医師は Level 1 の HMA の重要度を強調したくないと主張するように、TCAV は専門家によってモデル予測を解釈することができて、モデルに同意しない場合に、モデルを修正するのに役立つ可能性がある。

10 Future work

- TCAV の画像データ以外 (audio, video, sequences, ...) での適用
- TCAV の解釈性以外の活用 (AE の検知) (Appendix A)
- 概念の自動抽出

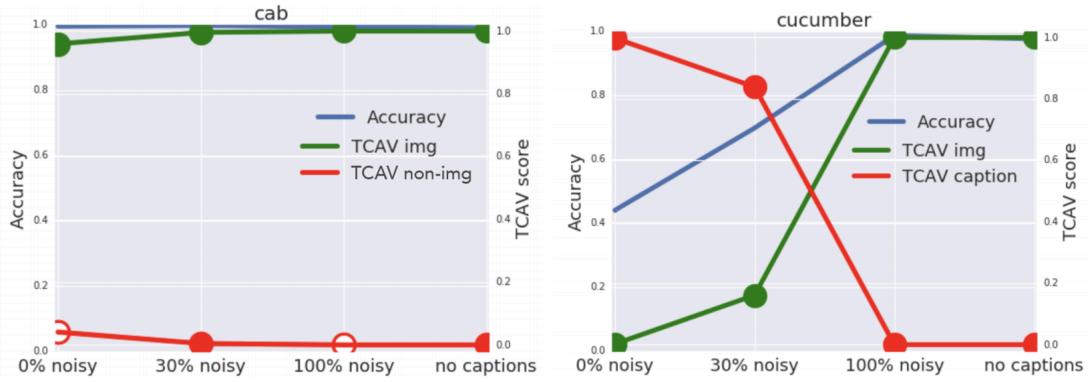


Figure7 TCAV results with approximated ground truth: Both cab and cucumber classes, TCAV closely matches the ground truth. For the cab class, the network used image concept more than the caption concept regardless of the models

11 TCAV on adversarial examples (Appendix A)

AE によってシマウマと分類される 2 つデータの TCAV スコアと通常のシマウマの TCAV スコアを比較する (Figure 9). AE によって生成されたスコアの分布と元の分布が異なることがわかる。例えば、通常の画像に対する TCAV スコアを予め知っていれば、AE による攻撃を検知することができる。

References

- [1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.
- [2] Gilmer Justin Yosinski Jason Raghu, Maithra and Jascha SohlDickstein. Svcca singular vector canonical correlation analysis for deep understanding and improvement. *arXiv preprint arXiv:1706.05806*, 2017.
- [3] Zaremba Wojciech Sutskever Ilya Bruna Joan Erhan Dumitru Goodfellow Ian Szegedy, Christian and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

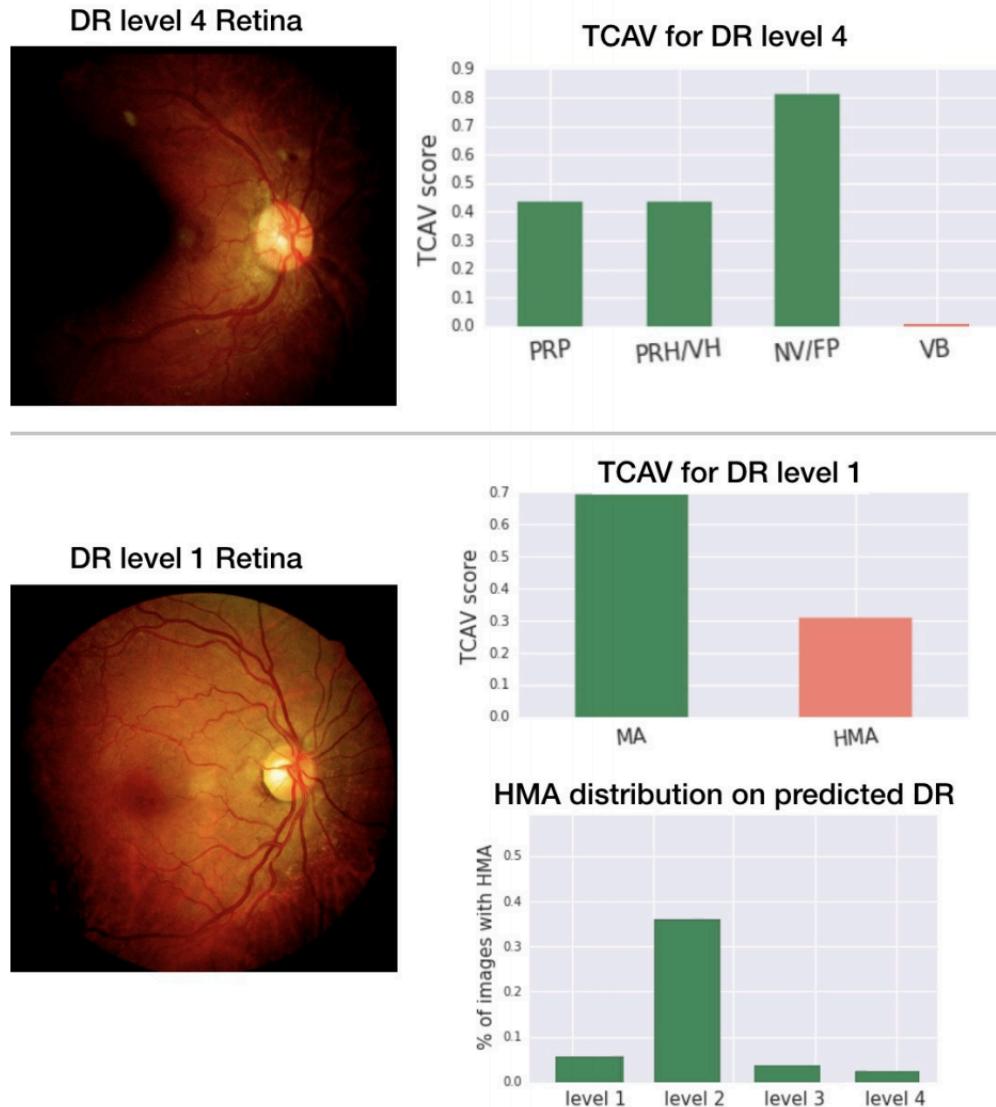


Figure8 Top: A DR level 4 image and TCAV results. TCAVQ is high for features relevant for this level (green), and low for an irrelevant concept (red). Middle: DR level 1 (mild) TCAV results. The model often incorrectly predicts level 1 as level 2, a model error that could be made more interpretable using TCAV: TCAVQs on concepts typically related to level 1 (green, MA) are high in addition to level 2-related concepts (red, HMA). Bottom: the HMA feature appears more frequently in DR level 2 than DR level 1.

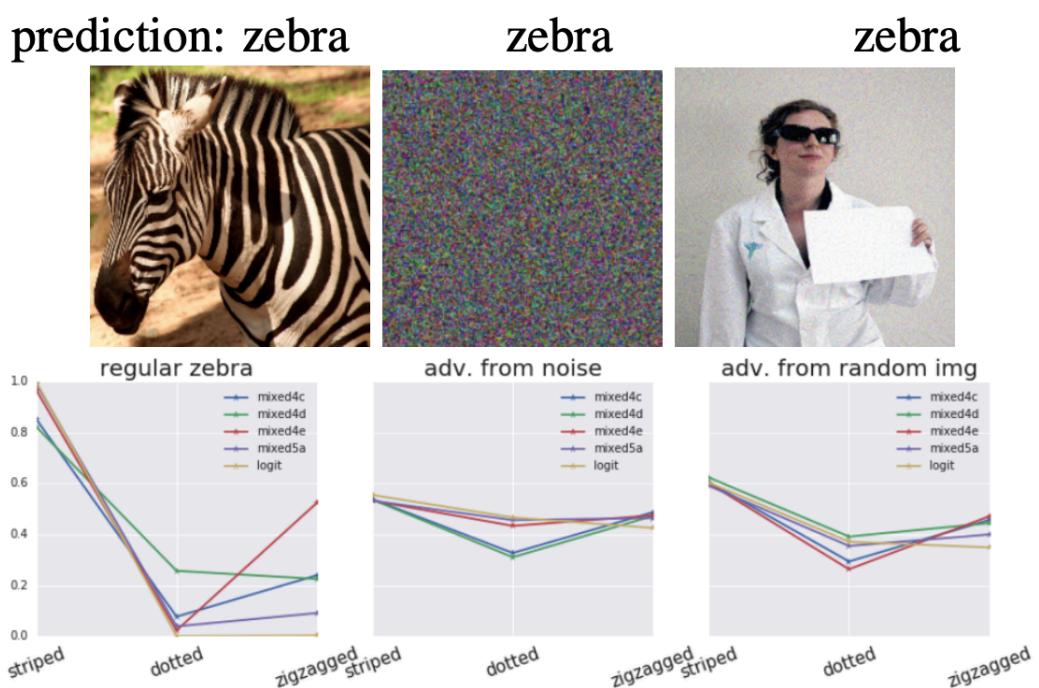


Figure9 Two types of adversarial images that are classified as zebra. In both cases, the distribution of TCAVQ are different from that of a normal zebra.