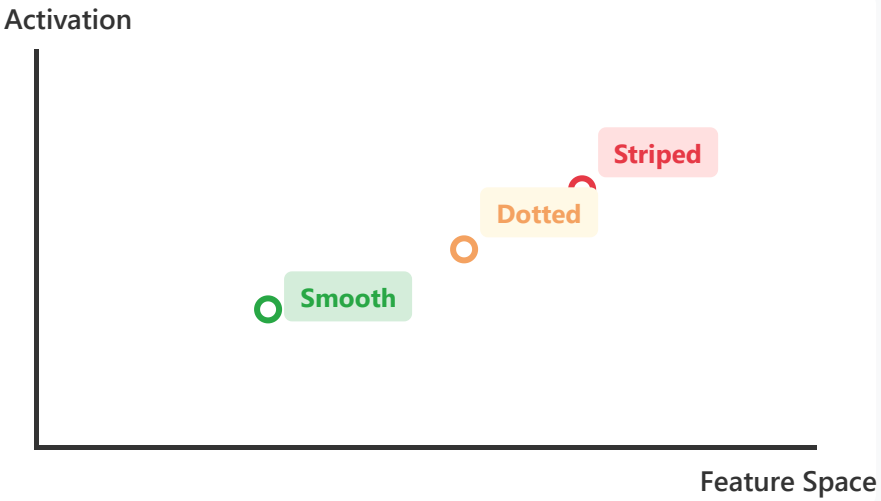
Concept-based Explanations

Human-understandable concepts vs individual features

Concept Space Visualization



Example: Texture Classification

Concept: "Striped pattern"
Model Sensitivity: How much does prediction change when moving in the direction of the striped concept?

$$S_C = \nabla_C f(x) \cdot v_C$$

Directional derivative along concept vector

Theoretical Foundation

TCAV Score:

$TCAV_{C,k,l}(x) = 1$ if $S_C^{k,l}(x) > 0$, else 0

TCAV

Testing with Concept Activation Vectors

Measures model sensitivity to user-defined concepts

- Define concept examples
- Train linear classifier
- Compute directional derivatives

ACE

Automated Concept-based Explanation

Automatically discovers important concepts without human input

- Segment activation space
- Cluster similar patterns
- Identify meaningful concepts

Concept Bottleneck