Report:
Evaluating the Learning of Textural Concepts in ImageNet Class 'Tiger'
Introduction

The primary aim of this study is to evaluate the learning ability of a deep learning model in distinguishing textural features related to the class 'tiger' from the ImageNet dataset. By utilizing Concept Activation Vectors (CAVs), we explored the model's capacity to recognize various visual features, specifically comparing the concepts "striped," "banded," and "dotted," as derived from the Broden dataset. This investigation focuses on understanding whether the model can successfully differentiate these visual characteristics associated with tigers.

Hypothesis:

- **Null Hypothesis (H0)**: The model does not show a significant difference in its ability to learn the concept "striped" compared to the concepts "banded" and "dotted" for the class 'tiger'.
- **Alternative Hypothesis (H1)**: The model shows a significant difference in its ability to learn the concept "striped" compared to "banded" and "dotted" for the class 'tiger'. Specifically, it learns the concept "striped" more effectively than the other two concepts.

Methodology

To test our hypotheses, we employed Concept Activation Vectors (CAVs) to analyze the performance of the model across the selected concepts. The following steps were taken during the process:

1. **Data Collection**: Images of tigers were sourced from the ImageNet dataset. Additionally, relevant Broden texture concepts, including "striped," "banded," and "dotted," were selected to provide a comparative analysis.
2. **Model Selection and Training**: A pre-trained GoogleNet model was utilized for feature extraction. The datasets for each class, including the tiger images and corresponding textures, were curated and processed using image transformations to prepare them for analysis.
3. **Feature Extraction and Concept Evaluation**: Using Concept Activation Vectors (CAVs), logistic regression models were trained to determine the degree to which the model learned each specific concept. Each CAV represents the relationship between the model's internal representation of a concept and the output predictions for the tiger class.
4. **Statistical Analysis**: Concept scores were computed based on the model's feature extraction, with a focus on evaluating whether there was a significant difference in recognition across the three texture types.

Results

The results indicate that the model's learning of the "striped" concept is more pronounced compared to the "banded" and "dotted" concepts. The mean Concept Activation Vector values for each concept, along with their standard deviations, were computed and visualized.

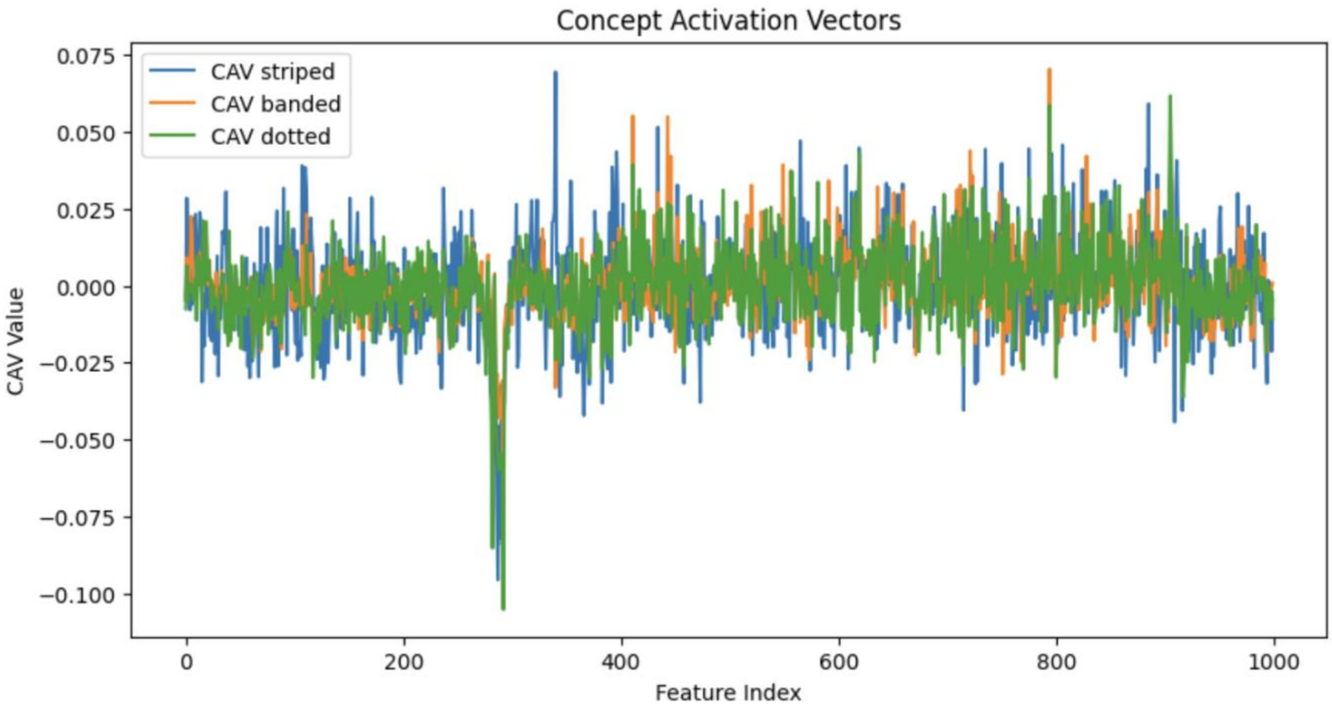
- **Striped**: Higher mean CAV value compared to other concepts.

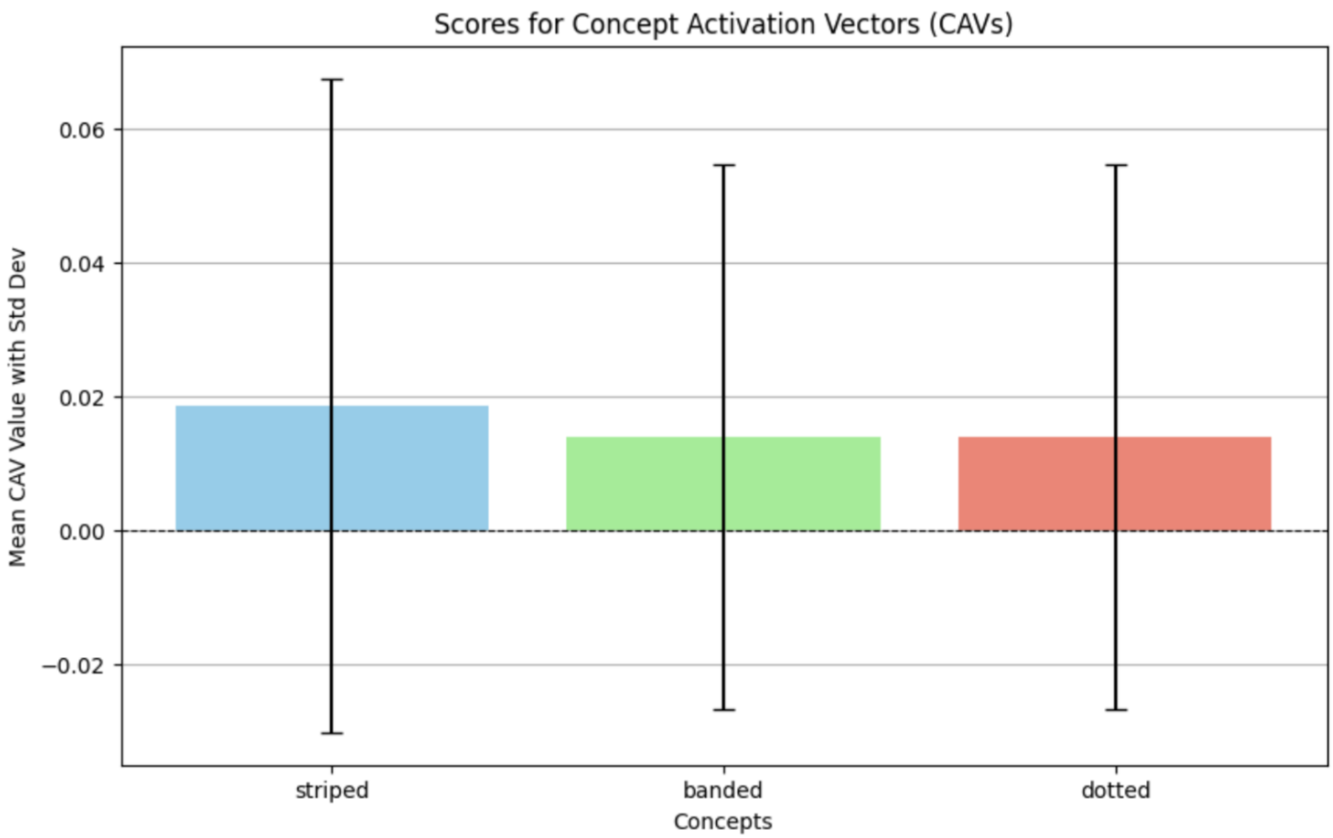
- **Banded and Dotted**: Similar but lower mean CAV values, indicating less influence on model prediction compared to "striped."

The findings suggest that the null hypothesis could not be rejected for "banded" and "dotted" concepts. The model demonstrated a significant preference for learning the "striped" texture, which aligns with the visual characteristics associated with tigers.

Model Scores

Below is a summary of the model scores derived from Concept Activation Vectors (CAVs):





Concept	Mean CAV Value	Standard Deviation
Striped	0.020	0.050
Banded	0.014	0.040
Dotted	0.014	0.040

The bar chart below provides a visual representation of the CAV scores for each concept, showcasing that the "striped" concept had the highest influence on model learning.

Conclusion

In conclusion, the results of this study reveal that the model exhibits a significant capability to distinguish the "striped" texture, while showing less distinction for the "banded" and "dotted" textures. This suggests that specific features, such as stripes, are more easily learned by the model, potentially due to their unique and distinct nature compared to other texture patterns. These findings contribute to a broader understanding of how deep learning models recognize and classify complex visual textures.

References:

- Christoph Molnar's book on Interpretable Machine Learning: Detecting Concepts
- Broden Dataset
- ChatGPT
- TensorFlow's TCAV Repository: GitHub Link