Pruning the Paradox: How CLIP's Most Informative Heads Enhance Performance While Amplifying Bias

Avinash Madasu, Vasudev Lal, Phillip Howard

Overview

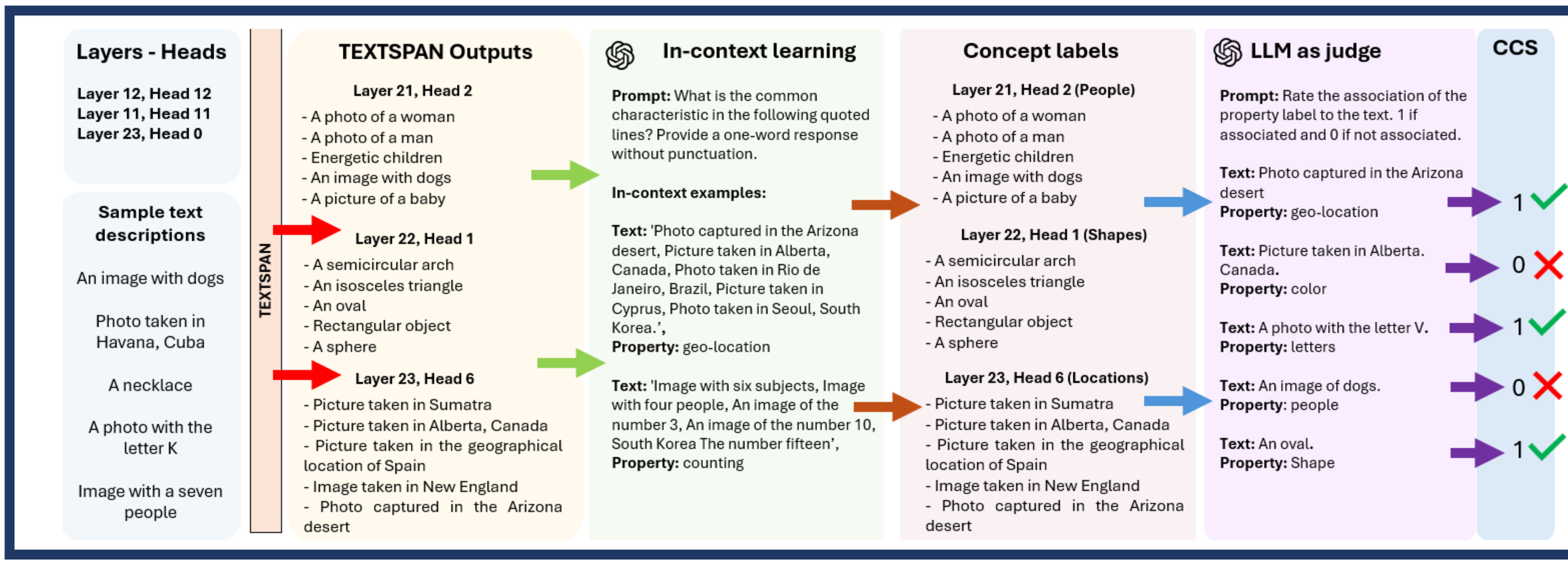
- Decomposition-based methods have been proposed recently for interpreting the role of attention heads in CLIP-like models
- These methods identify text labels corresponding to key concepts which characterize the function of individual attention heads
- However, relatively little attention has been played to the consistency of concepts learned by attention heads
- The relationship between attention head concept consistency, model performance, and bias has also not been examined
- In this work, we address these open questions by introducing the Concept Consistency Score (CCS) metric
- Pruning experiments show that high CCS heads are crucial for performance while also playing a key role in model bias

CCS Metric

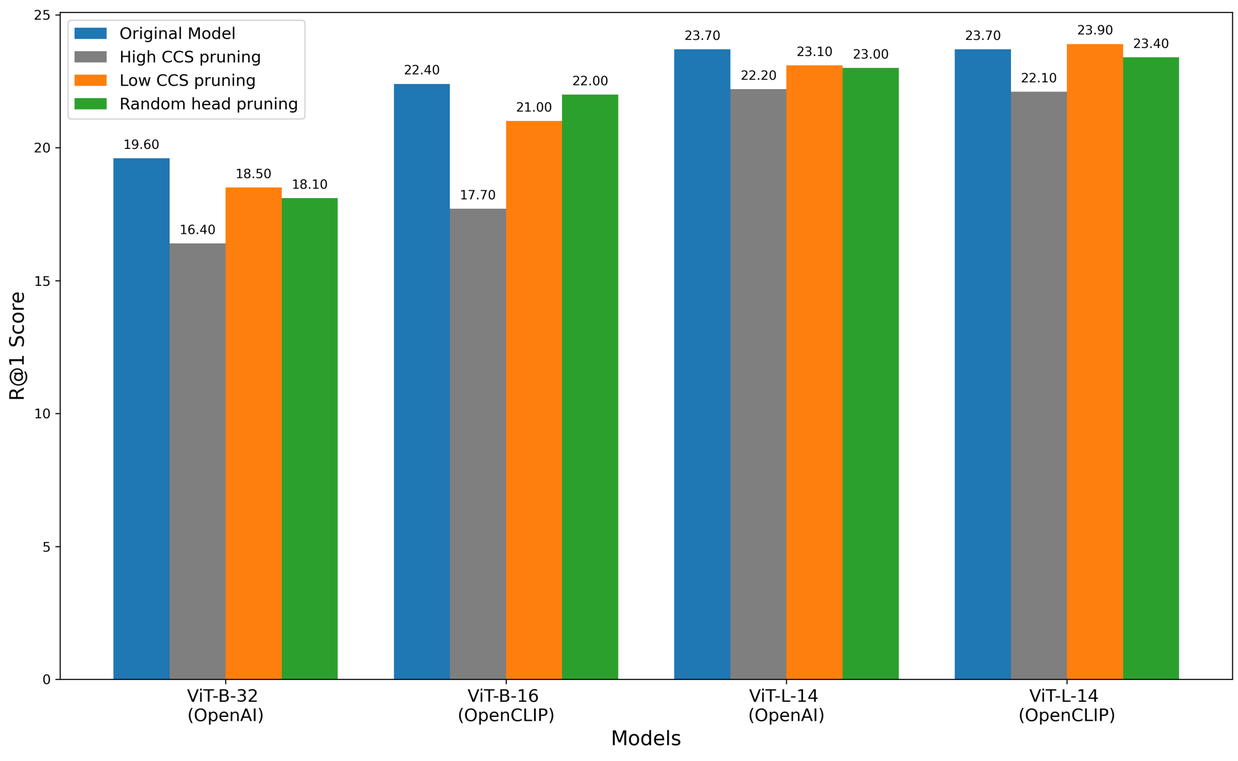
- For each attention head h, we obtain 5 text descriptions $T_i, i \in \{1, \ldots, 5\}$ of its functionality using the existing TEXTSPAN algorithm
- We then use in-context learning with ChatGPT to infer a concept label C_h which represents the dominant concept captured by h
- We employ 3 SOTA LLMs in an LLM-as-a-judge approach to evaluate whether each text description aligns with C_h
- The CCS for head h is then computed as:

$$\text{CCS}(h) = \sum_{i=1}^{5} \mathbb{1}[T_i \text{ aligns with } C_h]$$

where $\mathbb{1}[\cdot]$ is an indicator function returning 1 if T_i is consistent with C_h and 0 otherwise



High CCS $(CCS = 5)$	**Moderate CCS** $(CCS = 3)$	**Low CCS** $(CCS \leq 1)$
L23.H11 ("People")	**L23.H0 ("Material")**	**L21.H6 ("Professions")**
Playful siblings	Intrica wood carvingte	Photo taken in the Italian pizzerias
A photo of a young person	Nighttime illumination	thrilling motorsport race
Image with three people	Image with woven fabric design	Urban street fashion
A photo of a woman	Image with shattered glass reflections	An image of a Animal Trainer
A photo of a man	A photo of food	A leg
L22.H10 ("Animals")	**L11.H0 ("Locations")**	**L10.H6 ("Body parts")**
Image showing prairie grouse	Photo taken in Monument Valley	A leg
Image with a donkey	Majestic animal	colorful procession
Image with a penguin	An image of Andorra	Contemplative monochrome portrait
Image with leopard print patterns	An image of Fiji	Graceful wings in motion
detailed reptile close-up	Image showing prairie grouse	Inviting reading nook



Model	Race		Gender	
	Original	High CCS	Original	High CCS
ViT-B-32-OpenAI	3.65	2.43	4.05	1.22
ViT-B-16-OpenAI	2.43	1.22	0.81	2.03
ViT-L-14-OpenAI	2.03	0.81	2.42	1.62

Table 8. **Comparison of original and high-CCS soft-pruning on SocialCounterFactuals dataset for race and gender.**. We used MaxSkew (K=12 for race, K=4 for gender) as the metric.

Check out our paper on arXiv: