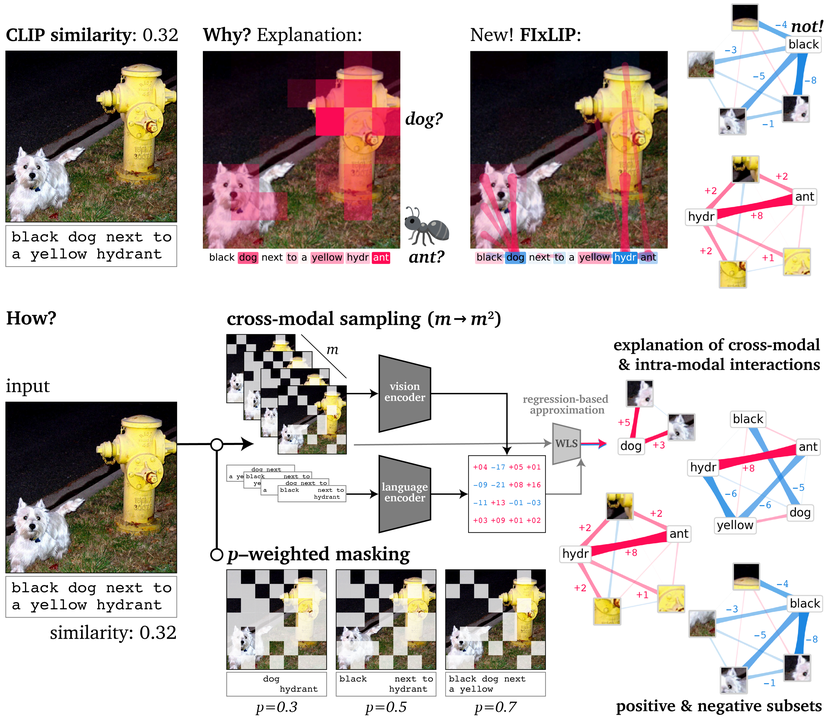
NEURAL INFORMATION PROCESSING SYSTEMS
UNIVERSITY OF WARSAW
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
mcml Munich Center for Machine Learning
UNIVERSITÄT BIELEFELD
Warsaw University of Technology
CCAI Centre for Credible Artificial Intelligence

Explaining similarity in vision-language encoders with weighted Banzhaf interactions

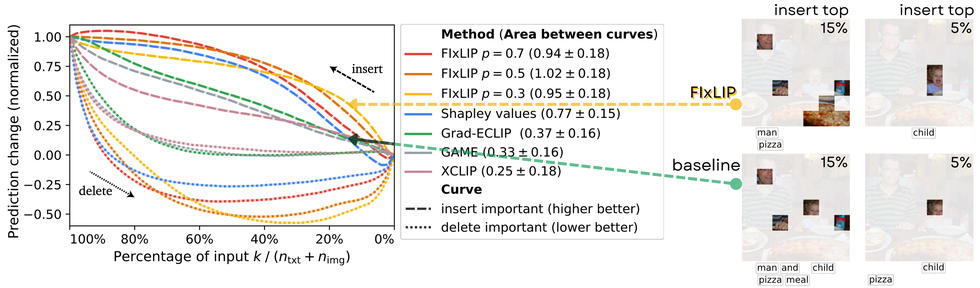
Hubert Baniecki, Maximilian Muschalik, Fabian Fumagalli, Barbara Hammer, Eyke Hüllermeier, Przemyslaw Biecek

TL;DR: A cool application of game-theoretic interactions in explaining a popular representation learning model architecture – **CLIP**.

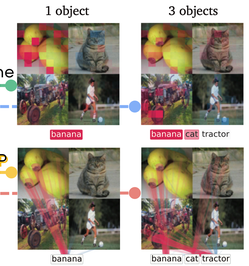
1. Faithful interaction explanations of CLIP-like models (FIxLIP)



2. We extend explanation evaluation metrics to interaction-based methods: area between the insertion/deletion curves and pointing game recognition

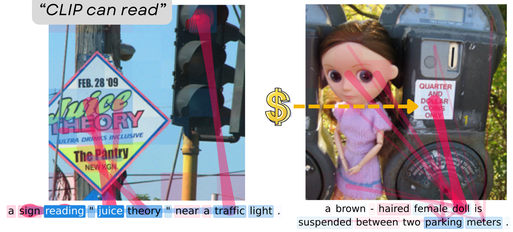


Explanation Method	Pointing Game Recognition (↑) 1 object	2 objects	3 objects	4 objects
GAME [12]	.61±.12	.43±.03	.33±.02	.28±.01
Grad-ECLIP [77]	.68±.15	.45±.04	.33±.02	.28±.01
Shapley values	.70±.11	.56±.06	.46±.05	.37±.04
Banzhaf values	.64±.12	.52±.06	.43±.05	.35±.04
exCLIP [48]	.73±.20	.88±.08	.89±.06	.92±.05
FIxLIP (Shapley interactions)	.83±.10	.82±.08	.84±.06	.86±.06
FIxLIP (w. Banzhaf interactions $p = 0.3$)	.78±.13	.78±.11	.80±.08	.81±.07
FIxLIP (w. Banzhaf interactions $p = 0.5$)	.81±.12	.80±.09	.81±.07	.83±.06
FIxLIP (w. Banzhaf interactions $p = 0.7$)	.83±.12	.81±.08	.83±.07	.85±.06

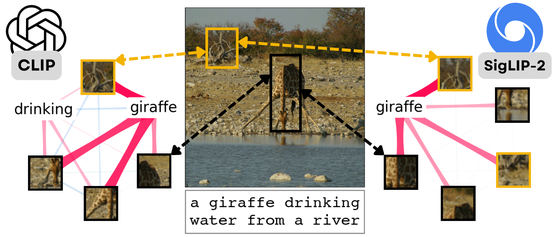


3. Intriguing examples

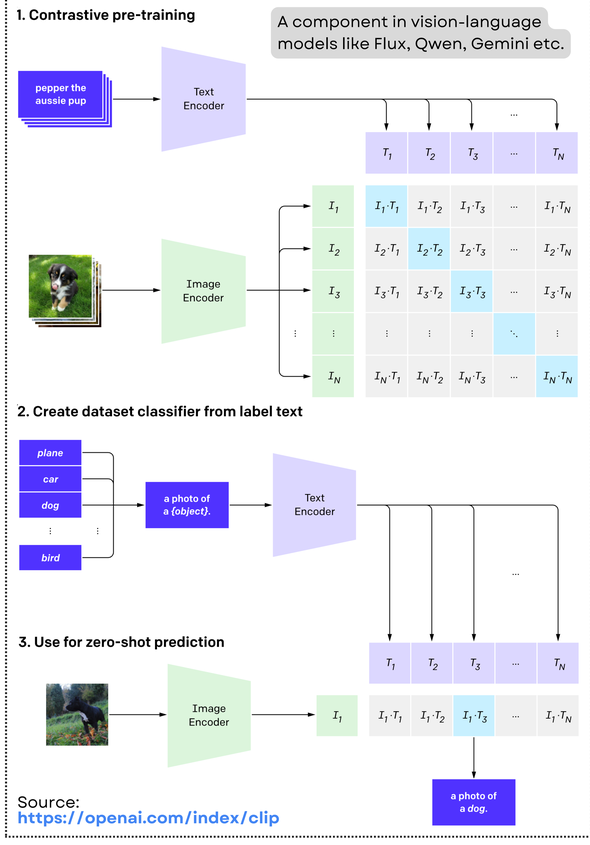
"The model is right for the wrong reasons"



4. Model comparison



0. A quick introduction to contrastive language–image pre-training (CLIP)



Source: https://openai.com/index/clip