|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ref | Method | Dataset | Task | Model | Category | Approach |
|  |  |  |  |  |  |  |
| 1. | CLIP | MS-COCO, Visual Genome, YFCC100M, WIT (self-made) |  | Resnet or ViT, CBOW or text encoder |  | Encode text and encode image. Minimise cosine similarity bw pairs and maximise bw the rest |
| 2 | SHAP-CAT |  |  |  |  | Take different vector representation from discriminative task. Apply shap and combine only imp features to get SHAP CAT representation |
| 3 | EXplainable AI (XAI) approach to image captioning | MSCOCO, Flickr30K | Image captioning,  Object Detection | VG-16,  LSTM |  | Generation (cnn + rnn) and explainable part. EX part use object detection output and caption. Attn is applied to give details and further this attn in passed in FFN for loss. |
| 4 | Grounding DINO | COCO, LVIS, ODinW | Object Detection | Swin T, Swin L |  | Feature enhance with multiple cross attn bw image and text embedding  Decoder layer with self and cross attn.  Contrastive loss.  Outputs bounding box and noun phrase |
| 5 | Segment Anything |  |  |  |  | Image embeddings and prompt embeddings; combine them.  Mask decoder. Loss with valid masks. |
| 6 | ICEv2 | ImageNet, ILSVRC, CUB-200-2011, ECSSD, DUTS, DUT-OMRON, Pascal VOC 07/12 | Segmentation |  |  | Predicts each path embedding among C+1 classes.  Adversarial network for class and background separation.  Tldr; give pixels not in background |
| 7 | 3VL |  |  |  |  | Break down captions in parts. Make negative samples. Make tree of negative and positive samples. Calculate contrastive loss for each layer of tree.  Patches of image matched with tokens. Token removal |
| 8 |  |  |  |  |  |  |

[Learning Transferable Visual Models From Natural Language Supervision](https://arxiv.org/pdf/2103.00020) [1]

[SHAP-CAT: A INTERPRETABLE MULTI-MODAL FRAMEWORK ENHANCING WSI CLASSIFICATION VIA VIRTUAL STAINING AND SHAPLEY-VALUE-BASED MULTIMODAL FUSION](https://arxiv.org/pdf/2410.01408) [2]

[EXplainable AI (XAI) approach to image captioning](https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/joe.2019.1217) [3]

[Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection](https://arxiv.org/pdf/2303.05499) [4]

[Segment Anything](https://arxiv.org/pdf/2304.02643) [5]

[ICEv2: Interpretability, Comprehensiveness, and Explainability in Vision Transformer](https://doi.org/10.1007/s11263-024-02290-6) [6]

[3VL: Using Trees to Improve Vision-Language Models’ Interpretability](https://arxiv.org/pdf/2312.17345) [7]

Surveys I’m still reading:

[EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI): FROM INHERENT EXPLAINABILITY TO LARGE LANGUAGE MODELS](https://arxiv.org/pdf/2501.09967)

[A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future](https://arxiv.org/pdf/2412.14056)