Visual Answering		
paper 1	ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks https://arxiv.org/abs/1908.02265	ViLBERT (short for Vision-and-Language BERT) is a model proposed for learning task-agnostic joint representations of image content and natural language. The model is designed to perform visual grounding from paired visiolinguistic data and uses a BERT-style pretraining architecture to extract features in visual and linguistic streams and attend to each other. ViLBERT extends the popular BERT architecture to a multi-modal setting and has been shown to outperform existing models on several vision and language benchmarks, including VQA, VCR, and NLVR2. The model is trained on image-text pairs, where the text is encoded with the standard transformer process using tokenization and positional embeddings, and images are decomposed into non-overlapping patches projected in a
paper 2	An effective spatial relational reasoning networks for visual question answering https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9704574/	The paper "An Effective Spatial Relational Reasoning Networks for Visual Question Answering" proposes a novel spatial relationship reasoning network (SRRN) that effectively models visual objects' spatial position relationship and object attribute relationship. The SRRN model is designed based on a deep co-attention mechanism, which combines visual object semantic reasoning and spatial relationship reasoning to achieve fine-grained multi-modal reasoning and understanding. The experimental results on the VQA 2.0 and GQA datasets demonstrate that the SRRN model outperforms existing state-of-the-art approaches, achieving an overall accuracy of 71.18% on the VQA 2.0 dataset and 57.59% on the GQA dataset. The SRRN model's success can be attributed to its ability to
paper 3	Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks https://paperswithcode.com/paper/image-as-a-foreign-language-beit-pretraining	The paper "Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks" proposes a unified pre-training approach called BEiT (Bridging Encoder and Transformer) for all vision and vision-language tasks. The paper argues that large foundational vision-language models follow two main training strategies, typically exemplified by disjoint architectures. Some vision-language pre-training approaches apply a contrastive loss, in a dual-encoder style architecture, while others use a single transformer-based encoder. The BEiT model unifies these two approaches by using a single transformer-based visual encoder for both image and text inputs. The paper shows that BEiT outperforms existing models on several vision and vision-language benchmarks, including VQA, GQA, and
Object Detection		
paper 1	You Only Look Once (YOLOv3): Object Detection and Recognition for Indoor Environment http://ijdri.com/me/wp-content/uploads/2021/06/17.pdf	This paper introduces a system for object detection and recognition in indoor environments, leveraging the YOLOv3 algorithm. The authors employ Computer Vision (CV) principles, emphasizing the significance of object detection in applications such as aiding the visually impaired and autonomous vehicles. They select YOLOv3 for its efficiency in processing entire images quickly. OpenCV is used for video frame capture and processing. A custom dataset, featuring six object categories, Table, Person, TV, Bottle, Chair, and Laptop, is created and used to train the model. The system generates Arabic voice messages to inform users about detected objects' presence and positions. It achieves high accuracy, with a mean Average Precision (mAP) of 100%, and the paper discusses potential future
paper 2	A Convolutional Neural Network based Live Object Recognition System as Blind Aid https://www.semanticscholar.org/reader/3453cf912f0b0b148125c595811b0d9a588018b2	This paper introduces a Convolutional Neural Network (CNN)-based live object recognition system designed to assist visually impaired individuals. The system captures real-time digital images using a camera, processes them through a CNN model implemented in Python and TensorFlow, and presents object information in audio or Braille text. Through image preprocessing and classification techniques, the model achieves a mean Average Precision (mAP) of 50% and a top-1 accuracy of 70.6% on a 200-object dataset. The system aims to enhance environmental awareness for the blind and identifies potential areas for future improvement, such as increased accuracy and real-time depth recognition. Overall, this research presents a promising approach to assistive technology for the visually
paper 3	Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection https://arxiv.org/pdf/2303.05499.pdf	In this research paper, the authors introduce Grounding DINO, a model designed for open-set object detection and referring object detection tasks. They conduct extensive experiments using two model variants, Grounding-DINO-T and Grounding-DINO-L, which combine image and text understanding. The results demonstrate the model's excellence in various scenarios: it outperforms other models in zero-shot transfer settings on COCO, LVIS, and ODinW benchmarks, sets records in COCO object detection without prior exposure to COCO data, and excels in referring expression comprehension (REC) tasks. Ablation studies emphasize the importance of specific model components, and the research explores the efficient transfer of pre-trained weights from DINO to Grounding DINO
voice activation		
paper 1	TOWARDS DATA-EFFICIENT MODELING FOR WAKE WORD SPOTTING https://arxiv.org/pdf/2010.06659.pdf	Focus on improving models wake word detection. Uses a small dataset of high quality data and added noise to simulate noisy environments. Also use self supervised learning to improve the performance
paper 2	Low-resource Low-footprint Wake-word Detection using Knowledge Distillation https://arxiv.org/pdf/2207.03331.pdf	Paper uses knowledge distillation to improve wake word detection. Allows to reduce the energy consumption of models Test on a variety of datasets and find that KD reduces error rate
paper 3	https://arxiv.org/pdf/2102.04488.pdf WAKE WORD DETECTION WITH STREAMING TRANSFORMERS	Authors evaluate a streaming transformers model for wake word detection. THeir mdoel performs better than most the benchmark which was CNNs Authors explore various techniques of providing input to the model such as looking ahead to the next chunk of data, stopping the gradient during back-propagation, using different positional embedding methods. Applying a variety of these techniques to improve results