1. VisualBERT is a multimodal model that integrates a BERT-style transformer with visual features derived from object detection models like Faster R-CNN. It processes textual and visual data within a unified transformer framework, enabling the model to capture alignments between image regions and language tokens. VisualBERT has been evaluated on a range of benchmarks, including VQA 2.0, VCR, NLVR2, and Flickr30K, showcasing strong adaptability across visual-language tasks. For the VQA task, it leverages pre-training on the COCO image caption dataset to achieve competitive performance.

The training pipeline includes three stages: (1) Task-agnostic pre-training on COCO captions, (2) Task-specific pre-training using data from individual downstream tasks, and (3) Fine-tuning with task-specific objectives and additional layers to optimize performance.

|  |  |  |
| --- | --- | --- |
| **Model** | **Datasets** | **Performance** |
| VisualBERT | VQA v2.0 | 71.00 |
|  | NLVR2 | 67.00 |
|  | VCR | 71.6 |
|  | Flickr30K | 71.33 |

**Table – 1:** Performance table of VisualBERT with different datasets

1. ViLBERT, a model designed for task-agnostic visiolinguistic (vision-and-language) representations. ViLBERT extends BERT to handle both visual and language inputs, enabling it to work across multiple vision-and-language tasks, such as Visual Question Answering (VQA), by pretraining on large paired datasets like Conceptual Captions. Two-Stream Architecture, ViLBERT uses separate "streams" for visual and textual data that interact through co-attentional transformer layers. This setup allows different levels of processing for each modality (visual and language) while enabling cross-modal interactions. The authors trained ViLBERT on two proxy tasks, Masked Multi-modal Learning and Multi-modal Alignment Prediction, this task requires the model to determine if a given text correctly describes the image, enhancing the alignment between vision and language features. ViLBERT was initially pretrained on the Conceptual Captions dataset, which contains about 3.3 million image-caption pairs, and later transferred to various specific tasks, including VQA on the COCO dataset, using fine-tuning. For VQA, ViLBERT was fine-tuned to answer questions about images.

|  |  |  |
| --- | --- | --- |
| ViLBERT | VQA v2.0 | 70.55 |
|  | VCR | 72.42 |

**Table – 2:** Performance table of ViLBERT with different datasets

1. This paper enhances Visual Question Answering (VQA) models for autonomous driving by aligning model attention with human attention in driving scenarios. While humans prioritize critical objects like vehicles and road signs, VQA models often focus on irrelevant details, reducing accuracy in driving-related tasks. To address this misalignment, we introduce a human-guided filter that prioritizes essential driving features (e.g., lanes, traffic lights) before visual data is processed by the vision transformer. This filter is integrated into the LXMERT model’s vision processing pipeline, improving interpretability, and reducing noise from irrelevant elements. Using the NuImages dataset, we evaluate the impact of this approach by comparing the original and filter-enhanced models against human responses. Results show that the filtered model better aligns with human attention patterns, leading to improved accuracy in driving-related VQA tasks. However, while this approach enhances domain-specific performance, it may limit generalization to broader VQA applications.
2. Pixel-BERT Pixel-BERT aligns image pixels with text in an end-to-end framework, using a CNN to process pixel-level data instead of relying on region-based features from object detectors like Faster R-CNN. This preserves fine-grained visual details such as spatial relationships, object shapes, and background context. Pixel and text embeddings are combined in a multi-modal Transformer to enable rich cross-modal attention. The model is pre-trained using Masked Language Modeling (MLM) and Image-Text Matching (ITM), helping it learn contextual and relational understanding across modalities. By operating directly at the pixel level, Pixel-BERT overcomes the limitations of bounding box-based approaches.

|  |  |  |
| --- | --- | --- |
| **Models** | **Datasets** | **Performance** |
| ViLBERT | VQA v2.0 | 70.55 |
| VisualBERT | VQA 2.0, NLVR2 | 70.80 (VQA) / 67.4 (NLVR2) |
| VL-BERT | VQA 2.0 | 72.22 |
| LXMERT | VQA 2.0 | 72.54 |
| UNITER | VQA 2.0, Flickr30K, COCO | VQA 72.27 / 63.3 (COCO) |
| Pixel-BERT (x152) | VQA 2.0, Flickr30K, COCO | VQA: 74.45 |

**Table – 3:** Performance table of different models with their datasets

1. The Med-VQA paper presents the ITLTA (Image to Label to Answer) framework to tackle challenges in medical VQA, such as limited data and interpretability. Instead of using an end-to-end model, ITLTA splits the task into multi-label image classification and label-based question answering, reducing complexity and improving adaptability in data-scarce settings. A CNN-based model (DenseNet) is first pretrained on external medical datasets to classify image attributes like modality, organ, and abnormality. For answering, ITLTA uses prompting with large language models (e.g., GLM-6B, Baichuan-13B) in a zero-shot setup, avoiding end-to-end training. This design reduces reliance on annotated Med-VQA datasets while lowering computational cost.
2. The Hi-VT5 model, introduced in the DocVQA paper, addresses Document Visual Question Answering (DocVQA) on multi-page documents using a hierarchical Transformer architecture capable of processing up to 20,480 tokens. Each page is encoded independently, producing [PAGE] tokens that summarize key information based on the question. Built on T5, Hi-VT5 concatenates these page-level embeddings and feeds them into a decoder to generate answers, effectively managing document length and structure. It is pretrained on a layout-aware de-noising task to align textual layout with semantics, enhancing the contextual value of [PAGE] tokens. A secondary module predicts the page containing the answer, improving explainability. This design reduces memory usage by summarizing pages rather than processing the full sequence at once.
3. MMFT-BERT enhances video-based Visual Question Answering (VQA) using separate BERT-based encoders for each modality: Q-BERT for text, V-BERT for visual features, and S-BERT for subtitles. Each modality is processed independently, capturing modality-specific information before fusion. The MMFT module aggregates these outputs using a trainable [FUSE] vector that attends to relevant modalities based on the question. Multi-head attention is applied across the concatenated outputs, enabling adaptive focus on informative features. Unlike models with fixed feature extractors, MMFT-BERT is trained end-to-end, allowing dynamic learning of multimodal representations. While this improves performance in complex scenarios, the multi-stream architecture increases computational demands, posing challenges for real-time use.

|  |  |
| --- | --- |
| **Models** | **Performance (Q+V+S)** |
| Two-Stream | 67.70 |
| Single BERT | 72.20 |
| WACV20 | 72.45 |
| MMFT | 74.97 |

**Table – 4:** Performance comparison of experiments on TVQA dataset

1. ViLT (Vision-and-Language Transformer) removes convolutional layers and object detectors, using a patch projection technique that divides images into fixed-size patches (e.g., 32×32), linearly embeds them, and feeds them directly into a Transformer alongside text tokens. This unified architecture reduces complexity and boosts speed. ViLT is pretrained with Masked Language Modeling (MLM) and Image-Text Matching (ITM) to learn token prediction and cross-modal alignment. Its simplified design allows inference speeds up to 60× faster than region-based models, while still achieving competitive results on standard VQA and retrieval tasks.

|  |  |  |
| --- | --- | --- |
| **Models** | **Datasets** | **Performance** |
| VilBERT | MSCOCO, Visual Genome | 70.55 |
| VisualBERT | MSCOCO, Conceptual Captions | 70.80 |
| LXMERT | MSCOCO, Visual Genome, GQA | 72.42 |
| UNITER | MSCOCO, Visual Genome, Conceptual Captions, SBU | 72.70 |
| PIXEL-BERT | MSCOCO | 74.45 |
| ViLT | MSCOCO, Visual Genome, Conceptual Captions, SBU | 70.33 |

**Table – 5:** Performance comparison of different models with their datasets

1. VBFusion introduces a multi-modal transformer-based model for Visual Question Answering (VQA) in remote sensing (RS) images, enabling natural language queries over satellite data. It uses separate encoders: a BoxExtractor for images, which generates random bounding boxes without predefined object labels, and a BERT-based tokenizer for text. Fusion occurs via a VisualBERT-based module with self-attention across image and text features, followed by an MLP for answer prediction. The model was evaluated on RSVQA-LR and RSVQAxBEN, with enhanced accuracy on the latter after including additional spectral bands, which improved performance on complex questions by enriching spatial and spectral representation.
2. DQMA (Question-Driven Multiple Attention) enhances VQA performance by focusing on question-relevant image regions. Visual features are extracted using Faster R-CNN, while questions are embedded via GloVe and processed by an LSTM. A question-driven attention mechanism computes relevance-based attention scores to highlight essential visual areas. A co-attentive network follows, with Self-Attention (SA) capturing dependencies within the question and Guided Attention (GA) aligning text and image features. The final attended features are passed through an MLP with softmax for answer prediction. Evaluated on VQA v2.0, DQMA outperforms other models in handling complex, crowded visual scenes.
3. VLC-BERT enhances Visual Question Answering (VQA) by integrating commonsense knowledge for tasks requiring contextual reasoning. Using Faster R-CNN for image region encoding and COMET for generating commonsense inferences, VLC-BERT dynamically incorporates contextual knowledge from ConceptNet and ATOMIC. Inferences are filtered using SBERT for semantic ranking, with the top selections fused with visual and textual features via a multi-head attention mechanism. The final fused representation is processed by a single-stream transformer (VL-BERT). Evaluated on the OK-VQA and A-OKVQA datasets, VLC-BERT outperforms models relying on static knowledge bases, demonstrating superior accuracy in tasks requiring complex reasoning.

|  |  |  |
| --- | --- | --- |
| **Model** | **Dataset** | **Performance** |
| ViLBERT | A-OKVQA | 25.85 |
| LXMERT | A-OKVQA | 25.89 |
| VLC-BERT | A-OKVQA | 38.05 |

**Table – 6:** Performance table of different models with their datasets

1. UNITER is a universal image-text representation model for various vision-and-language tasks, including VQA, image-text retrieval, and visual entailment. It uses Faster R-CNN for image region extraction and a BERT-based model for text embedding, creating joint representations through a Transformer architecture. UNITER is pretrained using Masked Language Modeling (MLM), Masked Region Modeling (MRM), and Image-Text Matching (ITM) for alignment, along with Word-Region Alignment (WRA) using Optimal Transport (OT) for fine-grained word-region mapping. Pretraining is conducted across multiple datasets (COCO, Visual Genome, Conceptual Captions, and SBU Captions), ensuring broad generalizability across V+L tasks.

|  |  |
| --- | --- |
| **Models** | **Performance** |
| SOTA | 70.90 |
| VilBERT | 70.92 |
| VLBERT | 72.22 |
| VisualBERT | 71.00 |
| LXMERT | 72.54 |
| UNITER | 74.02 |

**Table – 7:** Performance table of models with their datasets