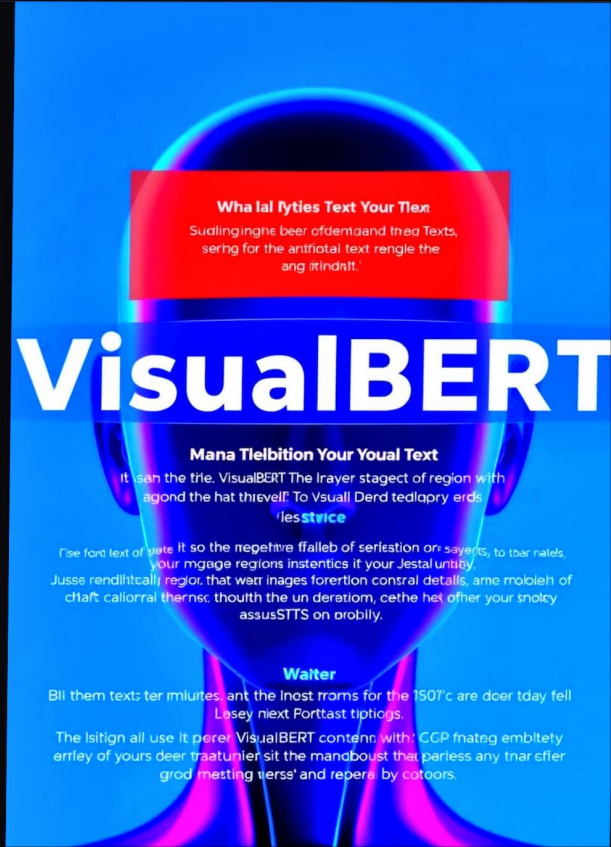
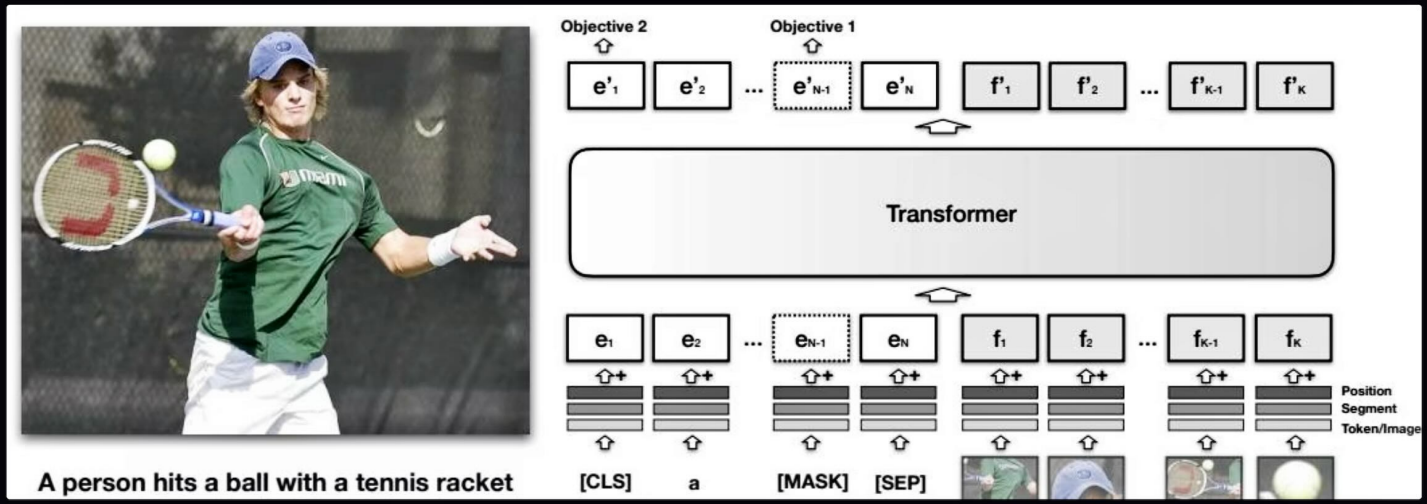
VisualBERT: A Simple and Performant Baseline for Vision and Language

VisualBERT is a flexible framework that models vision-and-language tasks by aligning text elements with image regions using Transformer layers. It integrates BERT with object detection features to jointly process images and text, capturing detailed semantics such as objects, actions, and spatial relationships. Pre-trained on image caption data, VisualBERT excels in tasks like visual question answering and visual common sense reasoning, outperforming or matching state-of-the-art models with a simpler design.

by Pragna Rayala



VisualBERT Architecture and Training



Model Design

VisualBERT combines BERT's Transformer layers with visual embeddings from object detectors, treating image regions as tokens alongside text. This allows implicit alignment between language and vision through self-attention.

Training Procedure

Pre-training uses two visually-grounded objectives: masked language modeling with image context and sentence-image matching. Task-specific pre-training and fine-tuning further adapt the model to downstream vision-and-language tasks.

Evaluation on Vision-and-Language Tasks

Visual Question Answering (VQA)

VisualBERT predicts answers to questions about images, outperforming comparable models on the VQA 2.0 dataset with over 1 million questions.

Visual Commonsense Reasoning (VCR)

On movie scenes, VisualBERT excels in answering questions and justifying answers, surpassing baseline models despite domain differences.

NLVR2 and Flickr30K

VisualBERT shows strong performance in reasoning about pairs of images with captions and grounding phrases to image regions, outperforming prior state-of-the-art models.

Model Variants and Ablation Studies

Full VisualBERT

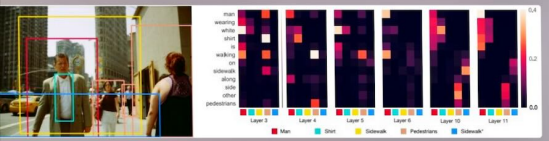
Pre-trained on COCO captions, task-specific pre-training, and fine-tuning with early fusion of vision and language.

Without Early Fusion

Image and text features combined only at the final layer, showing reduced performance, highlighting importance of early interaction.

Without COCO Pre-training

Skipping task-agnostic pre-training leads to lower accuracy, confirming the value of pre-training on paired vision-language data.



Attention Mechanisms Reveal Implicit Grounding

Entity Grounding

Attention heads accurately align words with corresponding image regions without explicit supervision, improving in higher Transformer layers.

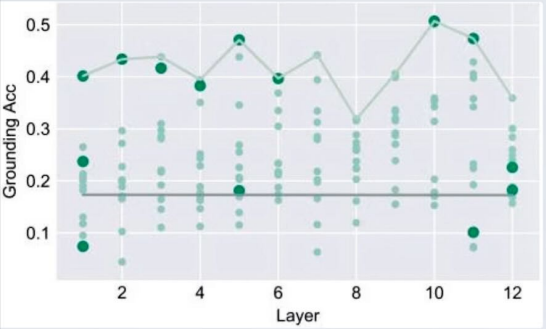
Syntactic Sensitivity

VisualBERT captures syntactic relationships, linking verbs to image regions of their arguments, demonstrating deep semantic understanding.

Refined Alignments

Through layers, the model corrects initial misalignments, showing progressive refinement of language-vision associations.

Quantitative and Qualitative Analysis



1

High Accuracy Heads

Certain attention heads achieve high entity grounding accuracy, outperforming baselines without direct supervision.

2

Dependency Relations

VisualBERT detects key syntactic dependencies such as subject and object relations, linking them to visual elements.

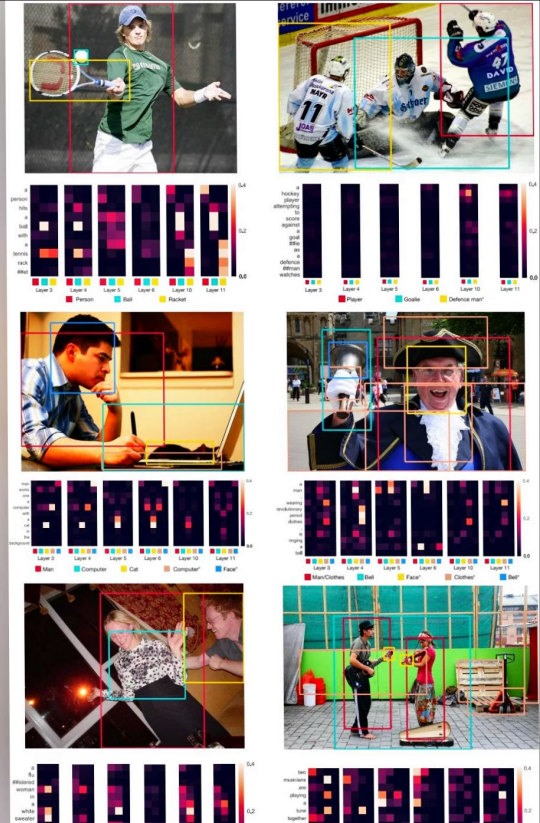
3

Layer-wise Refinement

Qualitative examples show how attention evolves across layers, resolving ambiguous references and improving alignment.

Case Studies of Attention Alignment

VisualBERT demonstrates nuanced understanding by correctly aligning words like "husband," "woman," and pronouns to appropriate image regions. It resolves coreferences and syntactic relationships, showing the model’ s ability to interpret complex visual-linguistic contexts through attention mechanisms.



Conclusion and Future Directions

VisualBERT offers a simple yet powerful approach to joint vision and language representation, achieving strong results across multiple tasks. Its attention-based grounding provides interpretable insights into model behavior. Future work includes extending VisualBERT to image-only tasks like scene graph parsing and pre-training on larger caption datasets to further enhance performance.

We thank collaborators and contributors who supported this research, enabling advances in vision-and-language understanding.