Paper Overview

Objective: Enhance multimodal feature preservation and answer reasoning in VQA by introducing an architecture improvement called the Layer-Residual Mechanism (LRM).

Problems Addressed: Transformer-based VQA models lose important feature information across deep layers.

Proposed Solution:

Layer-Residual Mechanism (LRM): A Residual Link between adjacent layers/attention blocks that:

- Preserves multimodal features
- Enhances attention flow
- Stabilizes training and reduces gradient issues

The LRM is applied to 3 Co-Attention architectures: Encoder-Decoder, Pure-Stacking, Co-Stacking

Team Catastrophe

Paper Implementation: LRCN: Layer-residual Co-Attention Networks for VQA

Chaitanya Chakka, Satya Akhil Galla, Gagan Singhal, Astha Rastogi

Data & Preprocessing - VQA v2 Dataset

- VQA v2 has 1.1M image–question–answer pairs over MSCOCO images.
- It balances yes/no, number, and other answers to test real visual reasoning.

Image Feature Processing:

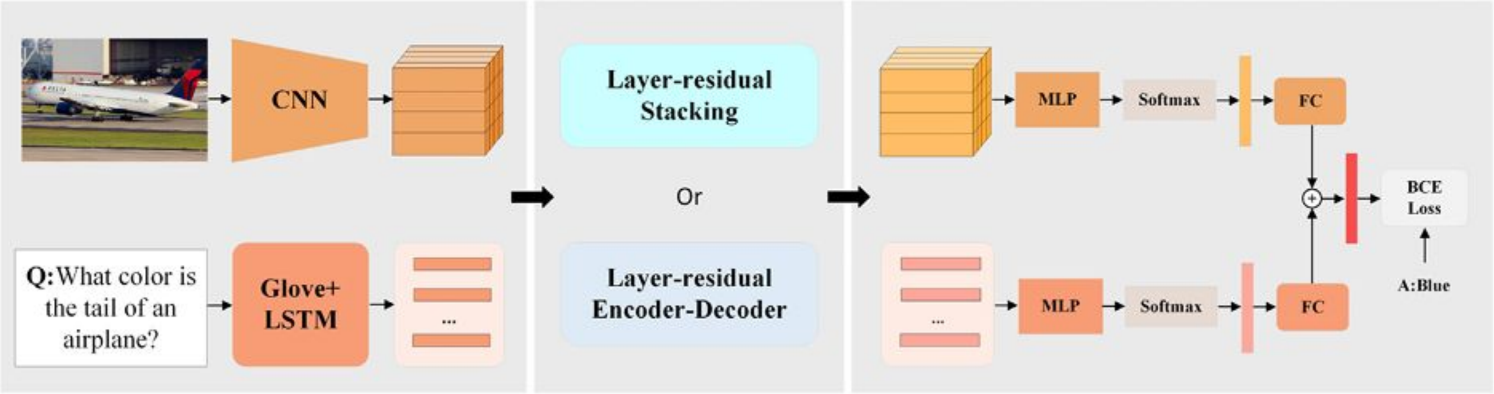
- Resize & center crop to **448×448.**
- Used **ResNet-152** (without classification head) to extract spatial features.
- Output: **14×14** grid of **2048-d** features.
- Padded to **16×16**, then downsampled to **8×8** using stride 2.
- Linearly projected to **512-d** to reduce dimensionality.

Text Feature Processing:

- Questions mapped to **300-d GloVe embeddings.**
- Padded/truncated to **14 tokens.**
- Passed through **1-layer LSTM (512 hidden units).**
- Outputs a fixed **512-d** question representation.

Both image and text features are transformed to 512-d so that they are aligned for LRCN integration

Model Architecture



Question and image feature representation(3.1)

Layer-residual Co-Attention Learning(3.2)

Multimodal Fusion and Answer Prediction (3.3)

Attention Modules

Self Attention

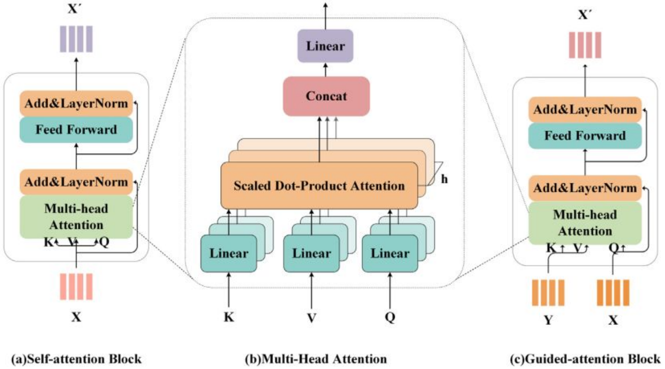
- Captures **intra-modal** relationships (within text or within images).
- Enables each element to attend to all others in its modality, enriching context.

Guided Attention

- Captures **inter-modal** interactions (e.g., text guiding visual feature focus).
- Allows one modality (usually text) to steer the attention of another (images).

Layer Residual Mechanism

- Adds **residual (skip) connections** across these custom attention layers.
- Stabilizes training, improves gradient flow, and preserves detailed information.



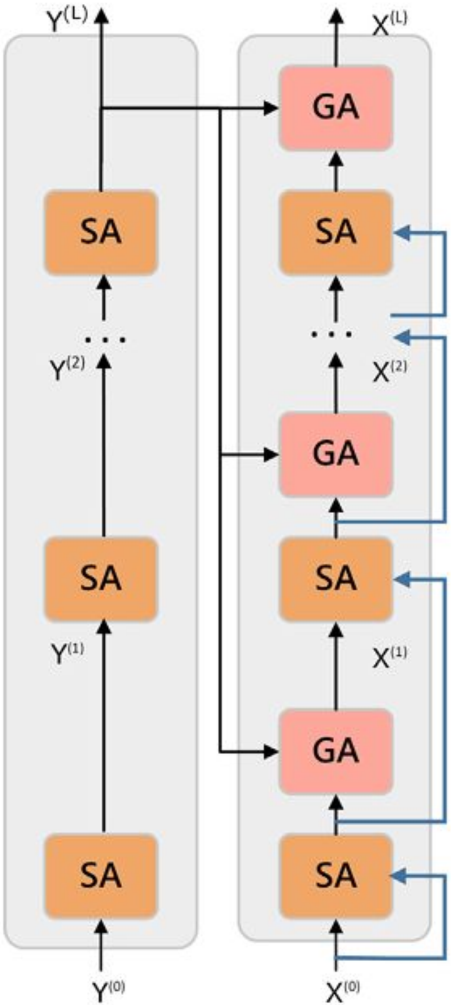
Encoder-Decoder

- **Design:**
  Separates phases: encoder refines text (SA), decoder uses refined text to guide image features (GA).

- **Flow:**
  Fine-grained question output (after L layers of SA) is fed into each image GA layer for guided extraction and multimodal integration.

- **Key Advantage:**
  Clear hierarchical structure improves stability and interpretability.

- **Limitation:**
  Slower interaction; relies on high-quality encoded text before fusion.



(a)Encoder-Decoder with LRM

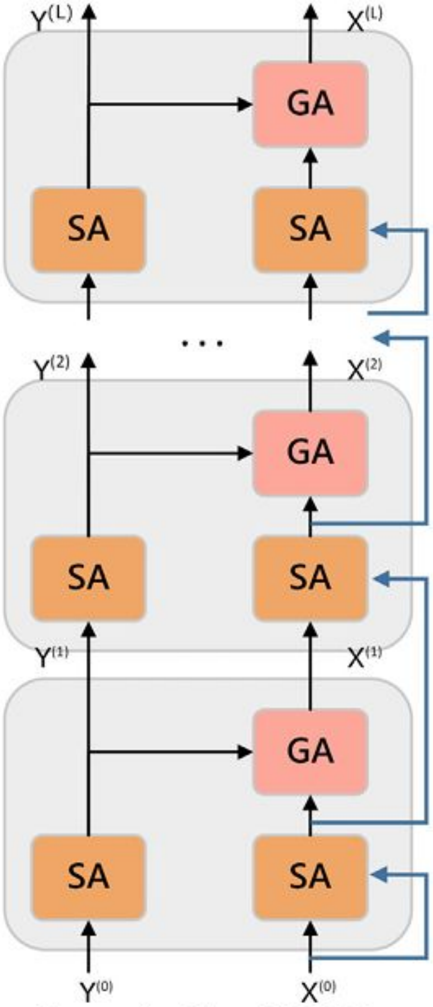
Pure-Stacking

- **Design:**
  Stacks text self-attention (SA); intersperses image SA + guided-attention (GA); late fusion where text guides image blocks.

- **Flow:**
  Output of the question SA layers are progressively fed as guidance to the image GA layers.

- **Key Advantage:**
  Efficient multimodal handling with clear separation.

- **Limitation:**
  Risk of early-stage misguidance if text is under-refined.



(b) pure-Stacking with LRM

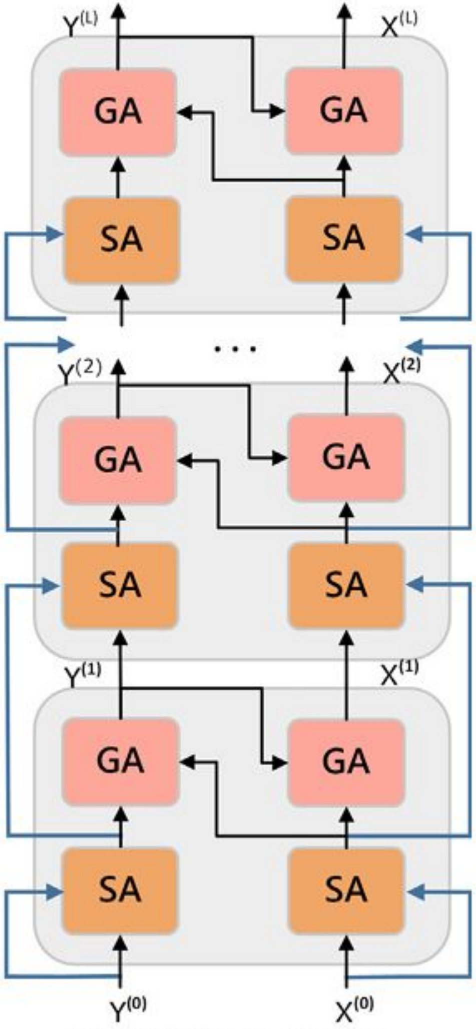
Co-Stacking

- **Design:**
  Applies self-attention (SA) and guided-attention (GA) **simultaneously** across all layers for both text and image.

- **Flow:**
  Text and image features are refined together in parallel; each layer jointly updates both modalities for early and continuous fusion.

- **Key Advantage:**
  Strong multimodal interaction from the start; reduces cumulative error.

- **Limitation:**
  Higher complexity; risk of redundancy without careful inter-modal balancing.



(c)Co-Stacking with LRM

Multimodal Fusion

Align image and question features in a shared space to enable accurate answer prediction.

- **Attention Pooling:**
  The model uses attention mechanisms to focus on the most relevant parts of the image and question. Attention scores are computed using small neural networks (2 FCs with ReLU and Dropout) on both modalities.

- **Feature Aggregation:**
  These attention scores are used to weight and summarize the visual and textual features into compact, meaningful vectors.

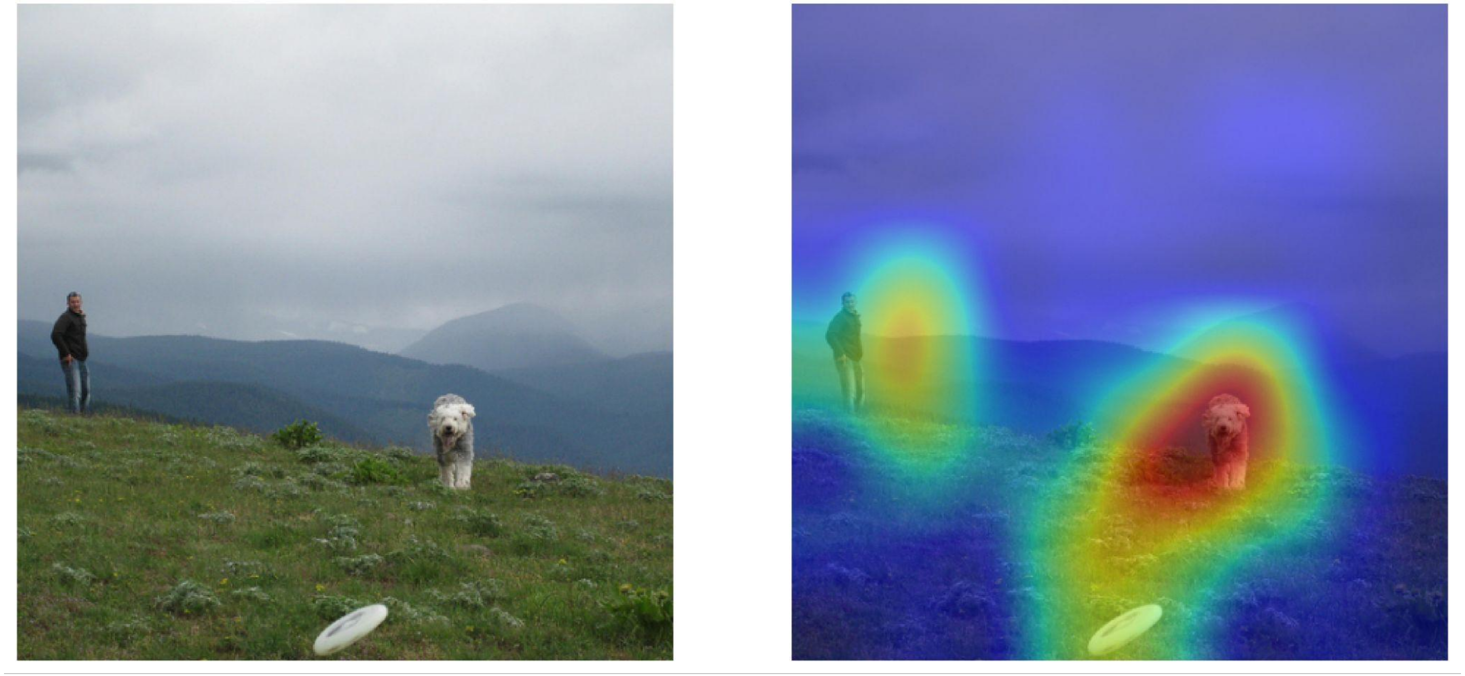
- **Linear Fusion:**
  The summarized features are combined and normalized to produce a single fused representation.

- **Answer Prediction:**
  The fused vector goes through ReLU and sigmoid activations to generate answer scores as probabilities across the vocab.

Training

- Training Specifications
  - Epochs: 13
  - Optimizer: Adam
  - Learning Rate Scheduler: Warm-up for 3 epochs and linear increase with milestone decay
  - Criterion: Binary Cross entropy
  - Train Data size: 443757 questions
  - Validation Data size: 214354 questions
- Hardware specifications
  - GPU: Nvidia L40 48 GB
  - Training time ~ 9 minutes/epoch
  - CPU count: 16

Qualitative Observations

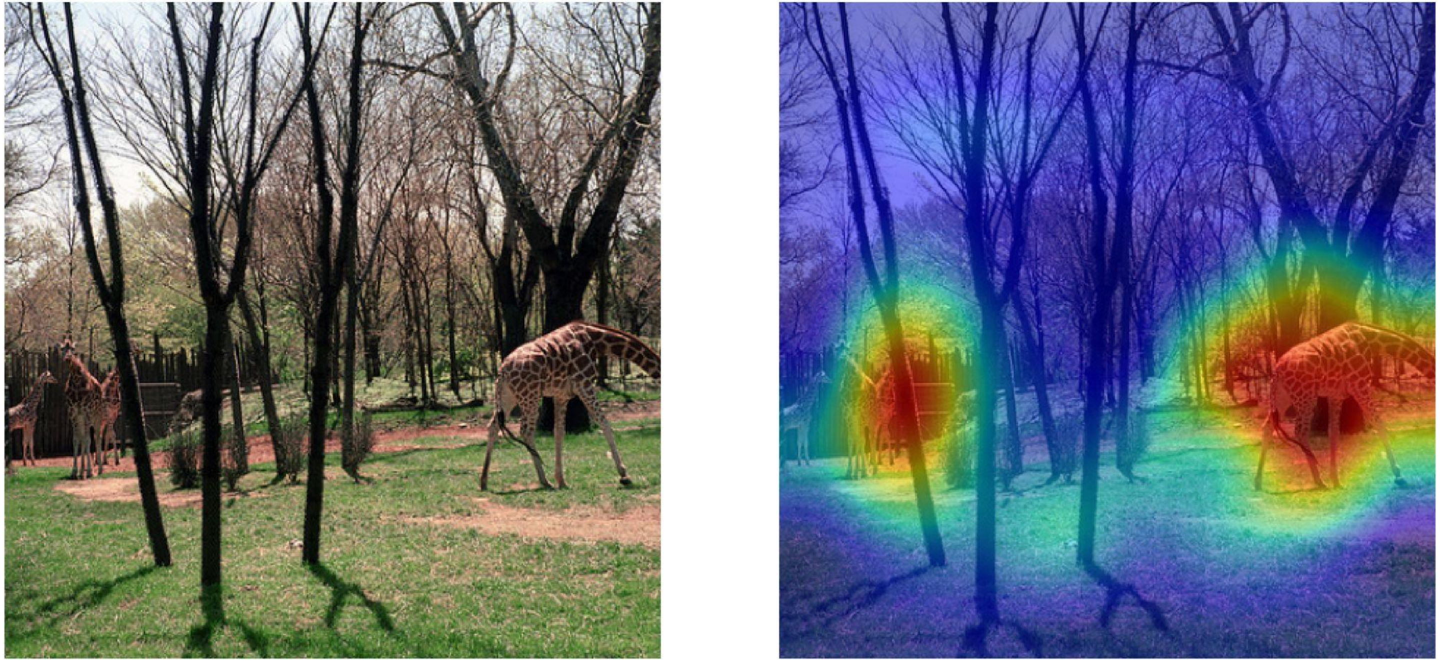


q_id: 119120010, word: animals, ans: dog

What animal is in this picture? - Single instance object

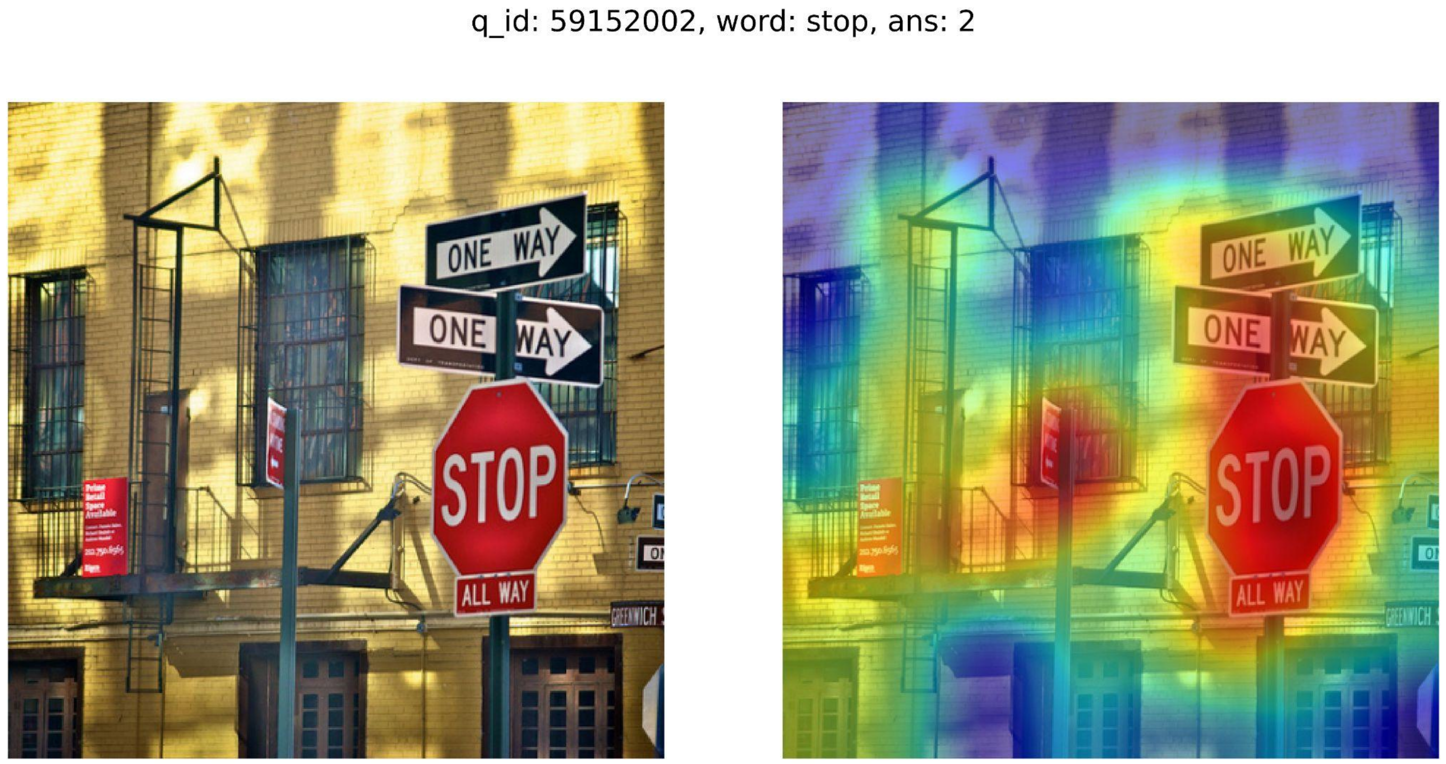
Qualitative Observations

q_id: 53016000, word: <unk>, ans: yes



Is this a giraffe? - Multi instance object

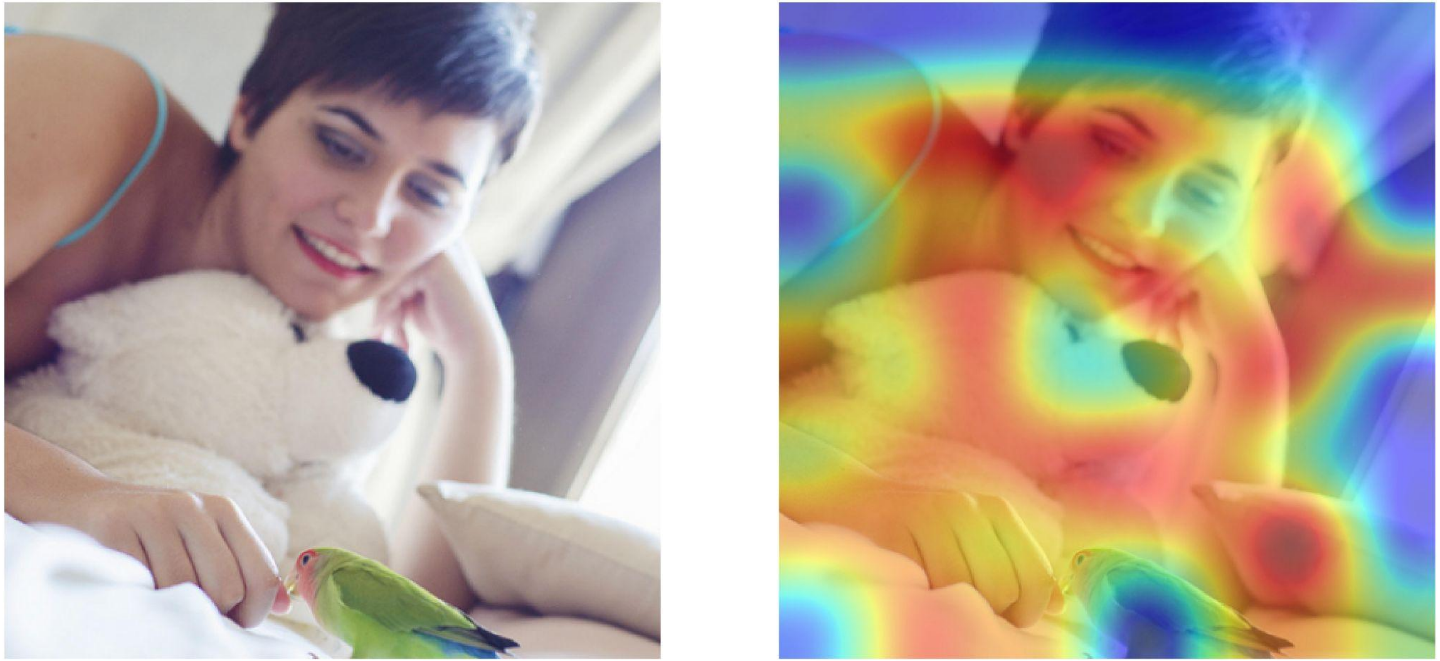
Qualitative Observations



q_id: 59152002, word: stop, ans: 2

How many stop signs ? - Multi instance object

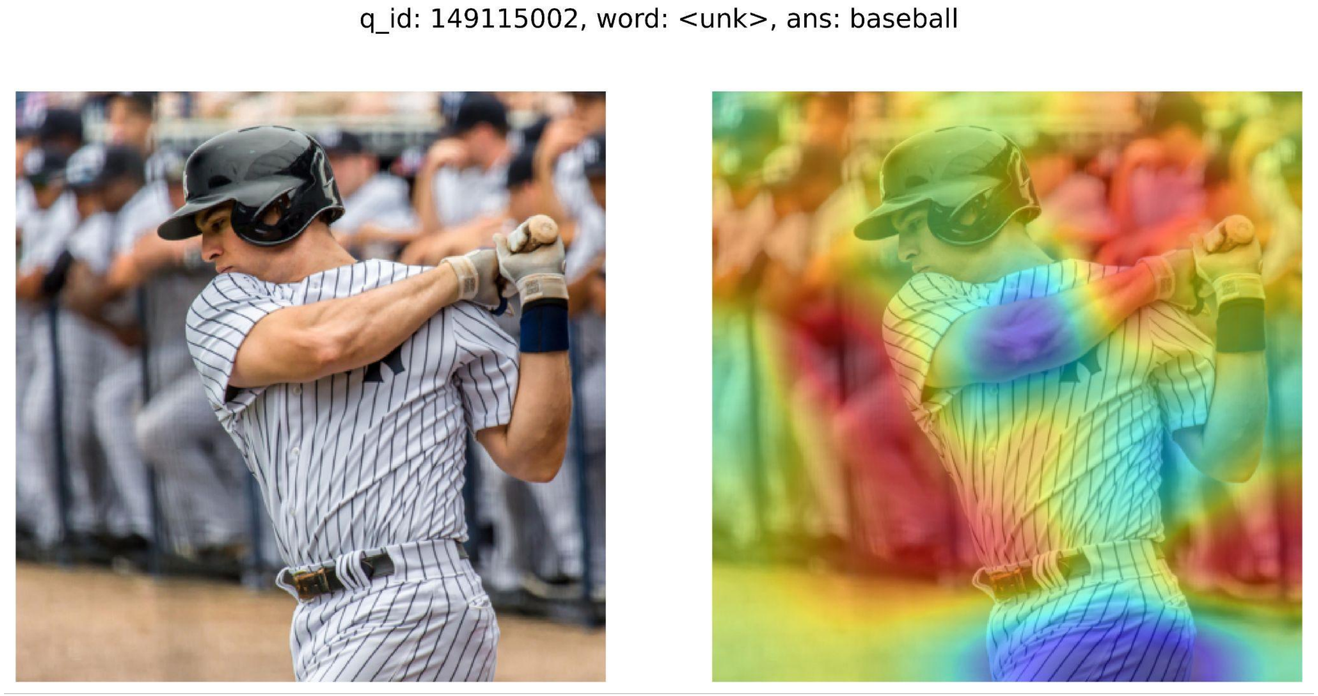
Qualitative Observations



q_id: 97800001, word: color, ans: pink

What color is the bird's face? - Abstract concepts

Qualitative Observations



q_id: 149115002, word: <unk>, ans: baseball

What sport is being played? - Indirect Reasoning

Well you have been deceived...

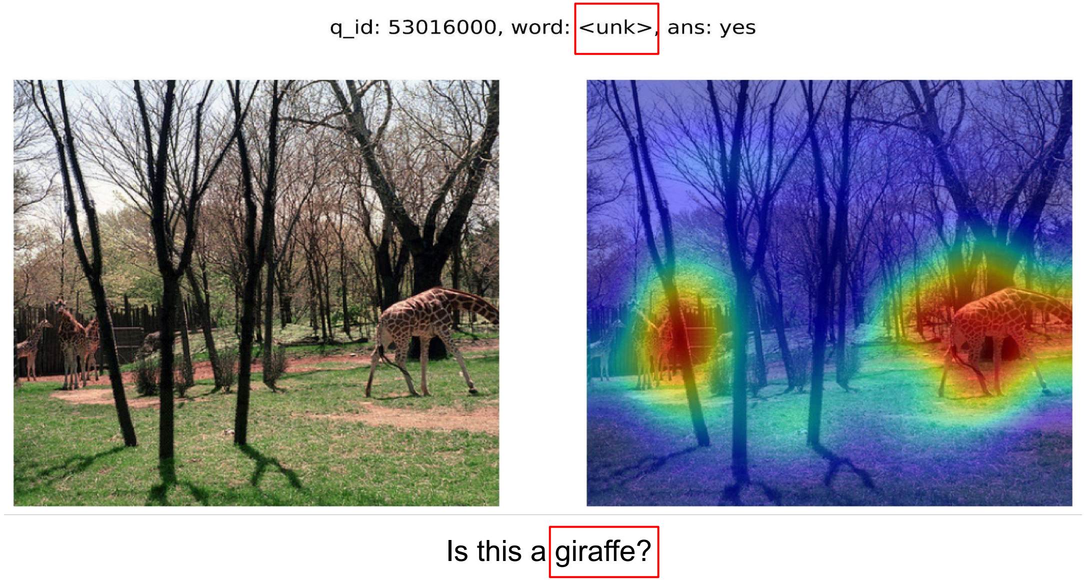
Results

	Encoder-Decoder		**Co-Stacking**		**Pure-Stacking**	
	Original	**Ours**	Original	**Ours**	Original	**Ours**
yes/no	87.74	69.09	87.31	74.75	87.57	69.47
number	53.99	36.07	54.43	40.85	53.61	35.88
other	62.29	25.95	62.33	45.74	61.99	25.89
overall	**71.83**	**44.87**	**71.72**	**57.61**	**71.58**	**44.98**

So what went wrong...

We messed up preprocessing

- Forgot to remove question mark - so ‘giraffe’ got stored as ‘giraffe?’
- Led to important words mapped to <unk>.
- Which, in turn, led to wrong embeddings.

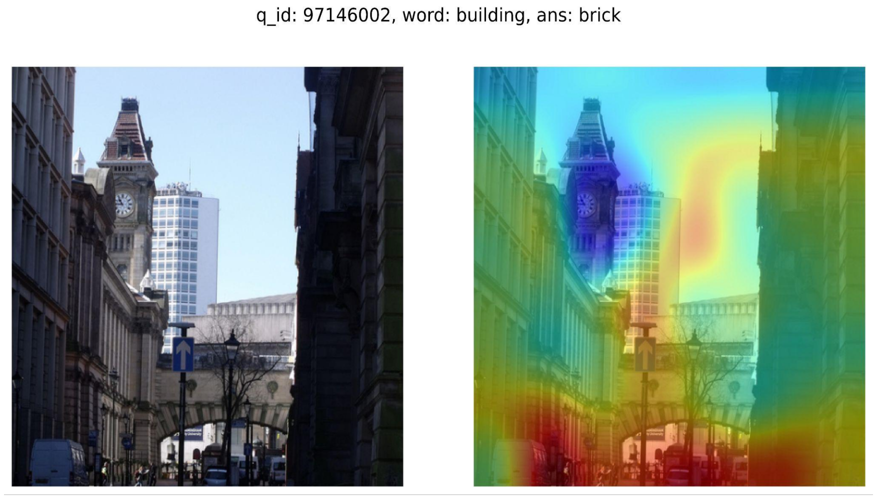


q_id: 53016000, word: <unk>, ans: yes

Is this a giraffe?

Why do the heat maps look so convincing?

- Accidently discovered the strength of these networks
  - Important words always mapped as <unk> - acted like masking
  - Model learned to focus on matrix object when it encounters <unk>.
- Ablation study
  - This multimodal architecture learns efficient modality alignment
- Due to this, the outputs looks exactly what we want
  - Looks like focusing on giraffe.
  - In reality, focuses on matrix object - generally coincides with the question subject.
- We see that when the no. of instances of the matrix objects became high in the image, the heatmaps were not that convincing.
  - We see in the image with many buildings, the model left out the clock building and even the non prominent buildings are not as strongly focused.



*Matrix object - The Important objects in the image

Observations

- **Counting weakness**: accuracy drops sharply on number-of-instances questions; fine-tuning on counting-focused VQA splits is recommended.
- **Variant ranking**: co-stacking still outperforms pure-stacking and encoder-decoder, matching the original paper.
- **Type accuracy**: best on yes/no, decent on “other,” poorest on numeric queries.
- **Training curve**: loss plateaus after the first few epochs, showing rapid early convergence.

Challenges

- **CLEVR Preprocessing Gaps:** Missing official steps required reverse-engineering data handling and normalization.
- **Dimensional Mismatches**: Ambiguous architecture specs caused frequent tensor shape errors and alignment fixes.
- **Reproduction Gap:** Matching reported benchmarks was difficult despite architectural details, hinting at hidden tuning.
- **Token Dictionary Scope**: Building a unified dataset-wide token ID dictionary would improve generalization over split-wise vocabularies.

Future Work

- Implement the model with **CLEVR dataset**.
- Provide **official code** to the paper and open-source for developments
- **Blind-navigation VQA**: apply the model to guide visually-impaired users through spoken scene questions and route suggestions.
- **Pre-training boost**: pre-train on large image-caption pairs to raise overall accuracy.
- **ViT backbone**: swap the CNN feature extractor for a Vision Transformer to capture richer image cues.
- **Video VQA**: extend the approach to understand and answer questions about video clips.

Questions?