**Introduction**

Q: At which festival can you see a castle in the background: Oktoberfest in Domplatz, Austria or Tanabata festival in Hiratsuka, Japan?

- Information can be scattered across multiple sources; the proposed system should be capable of identifying and collating information critical to answering a question.
- We aim to develop a system capable of selecting ‘relevant’ multimodal sources that can be combined to generate natural language answers to questions.

**Motivation**

- Information is rarely localized within individual sources.
- Information can come from any combination of modalities.
- Modality agnostic approach to generalize and scale with web data.

**Challenges**

- Significant data imbalance between positive and negative sources.
- Need for collective reasoning and ‘smart’ information aggregation.

**Baselines**

- **Lexical Overlap**: A trivial baseline that outputs the top 2 sources with the highest lexical overlap between question and caption.
- **VLP**: A transformer-based model trained on MLM and VQA is used for source retrieval.
  - Processes each source independently and hence poor in capturing multimodal aspect of selection.
  - Resource intensive and difficult to train.

**Approach A: Dense Super-Node Graph**

Q = The sign for Johnny’s Oyster Kitchen features a fish wearing what on its head?

### Intuition

- Unlike VLP, graphs can perform multimodal reasoning on multiple sources.
- It can learn meaningful connections between sources.

### Message Passing

- Super node contains all information about source and question.
- All nodes pertaining to a question are connected together (dense).
- Source selection is reduced to node classification (+/-).
- Message passing mathematical formulation:

\[ x'_i = W_1 x_i + W_2 \cdot \text{mean}_j N(x_j) \]

**Approach B: Star Graph**

Q = Are colubrid penguins and Maples’ penguin both extinct?

### Intuition

- Dense graph has a large number of uninformative connections (90% negative sources).
- Dropping irrelevant connections can improve learning.

- All sources for a question are connected to a central question node.
- We use multiple layers of the GNNs to enable message passing through the question node.
- Sparse graph leads to faster training and convergence.

**Primitive Representations**

Sentence embeddings from BERT to represent textual modality and ResNet-152 features to represent image modality while SOTA uses VinVL, X101fpn and VLP based feature representations.

**Results**

*Qualitative Results*

- What has been added to the pillars on both the pillars structure in Dom, Munich, and Austria?
- In which year were the first Tanabata Festival and Skyscrapers held as the capital city of Japan?
- Which has a smaller heap on the heap lifts in the red, the real or the yellow wood package on the message?

*Quantitative Results*

<table>
<thead>
<tr>
<th>Modality</th>
<th>Lexical Overlap</th>
<th>VLP-VinVL</th>
<th>Super Node</th>
<th>Star Graph</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>44.83</td>
<td>68.13</td>
<td>65.59</td>
<td>66.58</td>
</tr>
<tr>
<td>Text</td>
<td>33.78</td>
<td>69.48</td>
<td>59.39</td>
<td>60.74</td>
</tr>
</tbody>
</table>

Table 1: F1-score comparison of baselines with our methods.

**Insights**

- Even with ‘primitive’ representations, graph based approach has comparable performance to SOTA due to inherent ‘multihop’ reasoning ability.
- Intuition-based sparse connections are faster and improve the performance.

**Ongoing Work**

- Edge classification using graph attention networks.
- Experiment with gated graphs for better information flow.
- Using richer VinVL/CLIP features as node inputs.