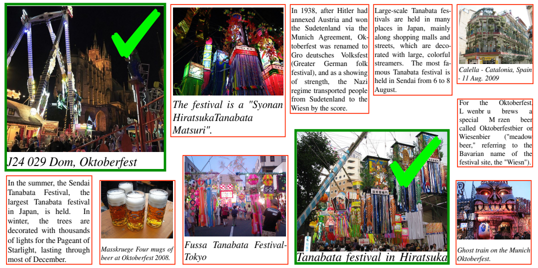
Multimodal Multihop Source Retrieval for WebQA

Nikhil Yadala[1], Paritosh Mittal[2], Saloni Mittal[1], Shubham Gupta[2]
[1]Language Technologies Institute, [2]Robotics Institute

Group 10

Introduction

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



J24 029 Dom, Oktoberfest

The festival is a "Syonan HiratsukaTanabata Matsuri".

In 1938, after Hitler had annexed Austria and won the Sudetenland via the Munich Agreement, Oktoberfest was renamed to Gro deutsches Volksfest (Greater German folk festival), and as a showing of strength, the Nazi regime transported people from Sudetenland to the Wiesn by the score.

Large-scale Tanabata festivals are held in many places in Japan, mainly along shopping malls and streets, which are decorated with large, colorful streamers. The most famous Tanabata festival is held in Sendai from 6 to 8 August.

Calella - Catalonia, Spain - 11 Aug. 2009

For the Oktoberfest, L wenbr u brews a special M rzen beer called Oktoberfestbier or Wiesenbier ("meadow beer," referring to the Bavarian name of the festival site, the "Wiesn").

In the summer, the Sendai Tanabata Festival, the largest Tanabata festival in Japan, is held. In winter, the trees are decorated with thousands of lights for the Pageant of Starlight, lasting through most of December.

Masskruege Four mugs of beer at Oktoberfest 2008.

Fussa Tanabata Festival-Tokyo

Tanabata festival in Hiratsuka

Ghost train on the Munich Oktoberfest.

- 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 and Challenges

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 multihop aspect of selection
  - Resource intensive and difficult to train

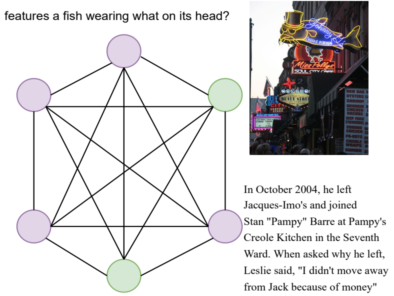
Approach A: Dense Super-Node Graph

Intuition

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

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

Caption: Beale Street, Memphis



In October 2004, he left Jacques-Imo's and joined Stan "Pampy" Barre at Pampy's Creole Kitchen in the Seventh Ward. When asked why he left, Leslie said, "I didn't move away from Jack because of money"

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

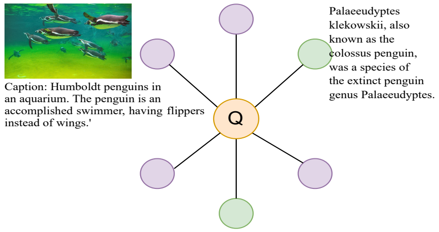
$$x_i' = W_1 x_i + W_2 \cdot mean_{j \in \mathcal{N}(i)} x_j$$

Approach B: Star Graph

Intuition

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

Q = Are colossus penguins and Marple's penguins both extinct?



Caption: Humboldt penguins in an aquarium. The penguin is an accomplished swimmer, having flippers instead of wings.'

Palaeeudyptes klekowskii, also known as the colossus penguin, was a species of the extinct penguin genus Palaeeudyptes.

Marples' penguin (Palaeeudyptes marplesi) was a large species of the extinct penguin genus Palaeeudyptes. It stood between 105 and 145 centimetres (3 ft 5 in and 4 ft 9 in) high in life, larger than the present emperor penguin.

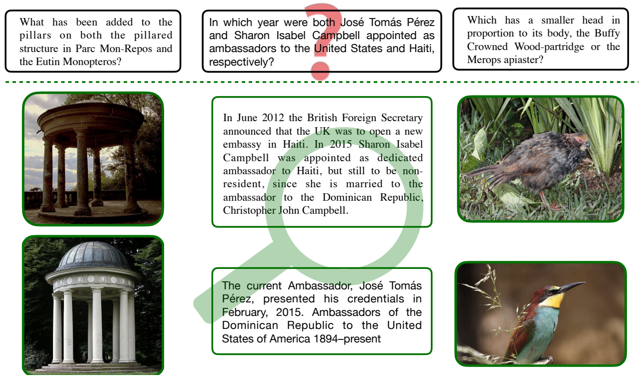
- All sources for a question are connected to a central question node
- We use multiple layers of the GNNs to enable message passing through the question node
- Sparse graph leads to faster training and convergence

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 pillared structure in Parc Mon-Repos and the Eutin Monopteros?

In which year were both José Tomás Pérez and Sharon Isabel Campbell appointed as ambassadors to the United States and Haiti, respectively?

Which has a smaller head in proportion to its body, the Buffy Crowned Wood-partridge or the Merops apiaster?

In June 2012 the British Foreign Secretary announced that the UK was to open a new embassy in Haiti. In 2015 Sharon Isabel Campbell was appointed as dedicated ambassador to Haiti, but still to be non-resident, since she is married to the ambassador to the Dominican Republic, Christopher John Campbell.

The current Ambassador, José Tomás Pérez, presented his credentials in February, 2015. Ambassadors of the Dominican Republic to the United States of America 1894–present

Figure 1:Queries along with Retrieved Sources from Star Graph

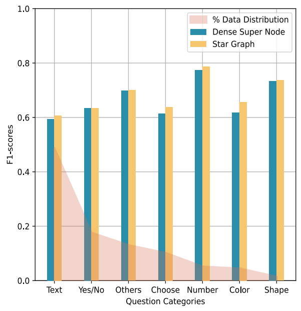
Quantitative Results

Modality	Lexical Overlap	VLP-VinVL	Super Node †	Star Graph †
Image	44.83	68.13	65.59	66.58
Text	33.78	69.48	59.39	60.74

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

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