Multi-modal Information Retrieval via Joint Embedding

KMA Solaiman and Bharat Bhargava

Department of Computer Science, Purdue University, West Lafayette, IN, USA

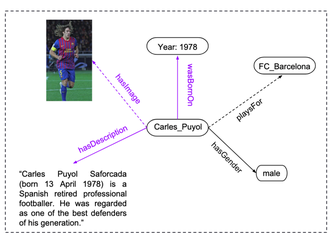
INTRODUCTION

Finding connected information among the vast amount of different data in current world is an interesting and difficult problem. In this work, we propose a solution for the multi-modal information retrieval problem using the concept of Joint Embedding. Video and text are embedded in a low dimensional shared joint embedding space by restraining documents with similar ontology, relations and entities to have similar embeddings. Video frames are extracted as a sequence of actions using Deep Video Captioning framework. For each document, Subject-Relation-Object triplets and topics are extracted using Open IE and LDA. These embeddings would provide a common representation for video and text data, and they can be used to provide users on specific missions with data of their preference without any intervention.

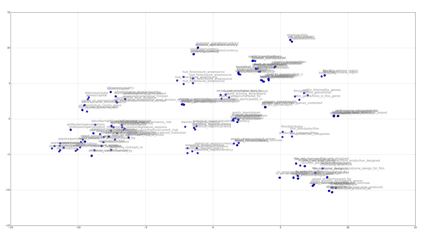
OBJECTIVES

- ❑ Retrieve **knowledge** for multiple users *changing* needs and mission
- ❑ Relate **multi-modal** data and update the existing **knowledge** for users
- ❑ Complete the **unfulfilled data needs** for missions
- ❑ Discover new **knowledge** that can benefit mission

MOTIVATION



Multimodal Knowledge Base

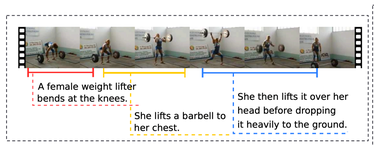


Relation Embeddings (DistMult)

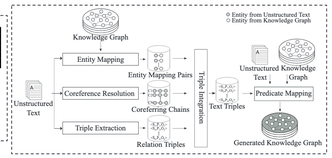
- [1] learned representations for entities and relations for KB represented as a list of relation triplets (subject, relation, object) via a NN.
- Relations are limited and entities come from a fixed, enumerable set of entities that appear in the knowledge base.
- [2] used encoders to learn from text descriptions and image for entities, and then used decoders to generate missing attributes for entities.
- Images, videos and text have more finer details than to just act as an attribute for entities in a KB.

METHODS AND MATERIALS

- Videos and unstructured text (twitter) are used as data sources for demonstrating multi-modality.
- For understanding videos, we use dense video captioning. It localizes distinct events in a long video stream, and generates captions for the localized events.
- [3] uses 3D convolutional network (C3D) to encode all incoming video frames. Convolution and pooling in spatiotemporal space preserves temporal sequence information within the video.
- Using the pooled features from C3D, [3] proposes variable-length temporal events and generates their captions using a two-level hierarchical captioning module that keeps track of context.

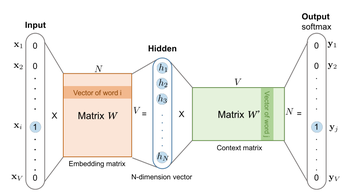


Joint Event Detection and Description in Continuous Video Streams

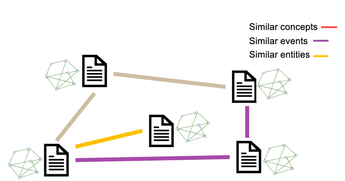


Architecture of the T2KG System

- Proposed method follows similar methodology as first half of the T2KG architecture [7] to generate the formulated text triplets.
- **Extracting triplets from text:** Open IE 5.0 [4] system is used to extract triplets consisting of [Subject-relation-Object] structure from both the unstructured text and the event descriptions from video.
- **Entity and Relation Mapping:** If an extracted entity or relation can be mapped to an identical entity in any KG, the URI of such an entity is used as a representative. Otherwise, a new URI is given to the entity or relation.
- **Concept Extraction:** For each of the text description, we generated topics using Latent Dirichlet Allocation (LDA) [6].
- **Joint Embeddings:** Following the intuition of Word2Vec [7], we train a neural network to restrict the low dimensional embeddings of documents and videos by following these objective functions –
  - ❑ Docs and videos with similar entities should have similar embeddings
  - ❑ Docs and videos with similar concepts should have similar embeddings
  - ❑ Docs and videos with similar relations should have similar embeddings

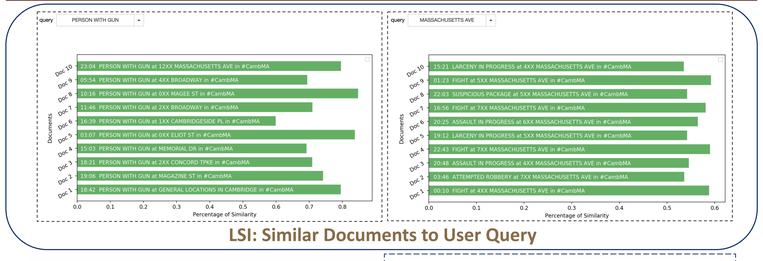


Model architecture of Word2Vec

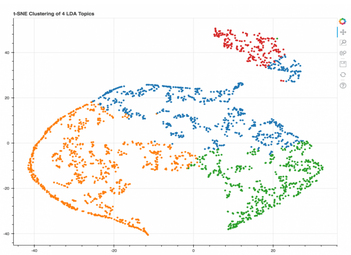


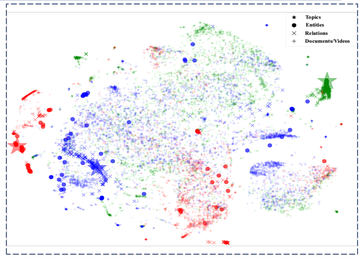
Information network for Joint Embedding

RESULTS



LSI: Similar Documents to User Query





Expected t-SNE Embeddings

CONCLUSIONS

- ❑ Proposed an end-to-end DNN architecture to integrate fine grained information from multiple data sources into a Knowledge Base
- ❑ Derived embeddings can be used as indices for finding events or information of interest
- ❑ Embeddings in same vector space depicts to similar interest to the mission and the user

FUTURE DIRECTIONS

- ❑ Complete and experiment the proposed architecture in Semantic Text Understanding and Information Similarity task
- ❑ Improvement of the Open IE task for triplet detection
- ❑ Experiment with other datasets in REALM

REFERENCES

1. Yang, Bishan & Yih, Wen-tau & He, Xiaodong & Gao, Jianfeng & Deng, li. Embedding Entities and Relations for Learning and Inference in Knowledge Bases, in ICLR 2015.
2. Pezeshkpour, Pouya & Chen, Liyan & Singh, Sameer. (2018). Embedding Multimodal Relational Data for Knowledge Base Completion.
3. H. Xu, B. Li, V. Ramanishka, L. Sigal and K. Saenko, "Joint Event Detection and Description in Continuous Video Streams," 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 2019, pp. 25-26.
4. Mausam. "Open Information Extraction Systems and Downstream Applications". Invited Paper for Early Career Spotlight Track. International Joint Conference on Artificial Intelligence (IJCAI). New York, NY. July 2016.
5. Kertkeidkachorn, N., & Ichise, R. (2017). T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. AAAI Workshops.
6. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3 (March 2003).
7. T Mikolov, K Chen, G Corrado, J Dean. Efficient estimation of word representations in vector space.

Acknowledgement: This research is supported by the Northrop Grumman Mission Systems University Research Program.