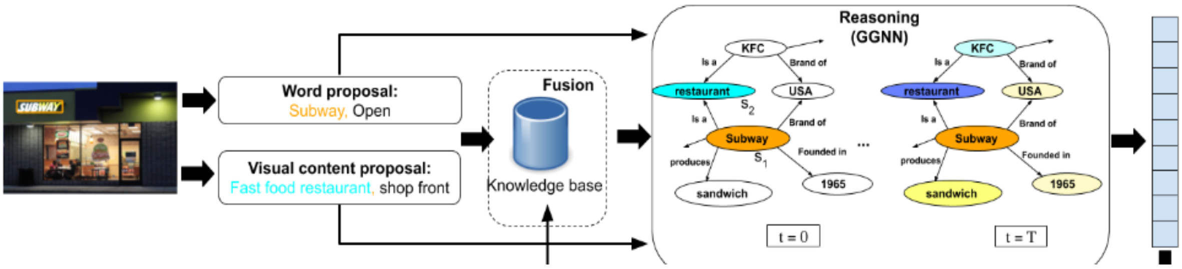
Knowledge-base for TextCaps and CTC

Scene-text contains explicit semantics that can be used as an additional modality to further improve image understanding. Recent works (TextCaps and Coco-Text Captioned) have been focusing on capturing the interplay between these three modalities, images, language and scene-text.

In this project, we will explore Cross-modal Retrieval, where the task is to provide a matching caption to a given image or vice-versa. The main idea is to create an external knowledge base or graph that captures relations between objects, scenes and scene-text. Later, the learned representation can be employed to assess which modalities can be employed to yield an improved retrieval performance.



The project will start with state-of-the-art models and define an integration strategy to query or exploit the information acquired in the knowledge base.
We will define different training strategies and compare the results. In this project, you will acquire knowledge of different text embeddings employed for understanding tasks, a combination of vision and NLP as well as a practical insight of how cross-modal retrieval systems work.