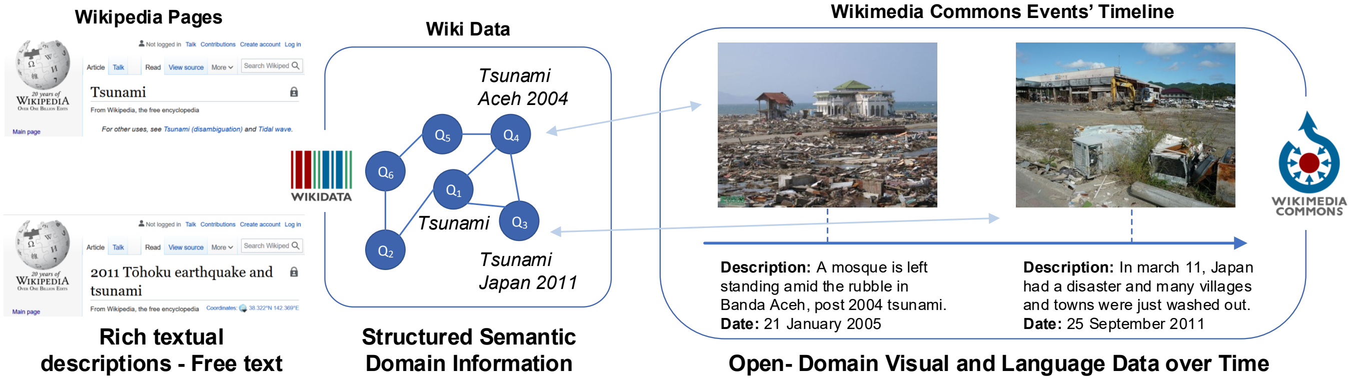
Towards Open-domain Vision and Language Understanding with Wikimedia

David Semedo
NOVA LINCS, Universidade NOVA de Lisboa, Caparica, Portugal
df.semedo@fct.unl.pt


Wikipedia Pages
Tsunami
2011 Tōhoku earthquake and tsunami
WIKIDATA
Wiki Data
Tsunami Aceh 2004
Q5
Q4
Q6
Q1
Q3
Q2
Tsunami
Tsunami Japan 2011
Wikimedia Commons Events' Timeline
WIKIMEDIA COMMONS
Rich textual descriptions - Free text
Structured Semantic Domain Information
Open- Domain Visual and Language Data over Time

Description: A mosque is left standing amid the rubble in Banda Aceh, post 2004 tsunami. **Date:** 21 January 2005

Description: In march 11, Japan had a disaster and many villages and towns were just washed out. **Date:** 25 September 2011

Figure 1: Wikimedia Open-domain Media Data Framework for Media Understanding overview.

Abstract

Current state-of-the-art task-agnostic visio-linguistic approaches, such as ViLBERT [2], are limited to domains in which texts have a visual materialization (e.g. a person running). This work describes a project towards achieving the next generation of models, that can deal with open-domain media, and learn visio-linguistic representations that reflect data's context, by jointly reasoning over media, a domain knowledge-graph and temporal context. This ambition will be leveraged by a Wikimedia data framework, comprised by comprehensive and high-quality data, covering a wide range of social, cultural, political and other type of events. Towards this goal, we propose a research setup comprised by an open-domain data framework and a set of novel independent research tasks.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → **Multimedia and multimodal retrieval**.

Keywords

Media Understanding, Vision and Language, Artificial Intelligence

1 Introduction

The Wikimedia[1] library can be seen as a digital mirror of humanity's footprint, covering not only the most impactful real-world events, that changed economies and societies, to general knowledge about a large variety of topics. Automatically understand media (images and texts) from such events, requires framing them in time, and learning how they relate with other media. Thus, multiple information types, ranging from free-text descriptions, images and semantic knowledge-graphs, have to be jointly considered. While Wikipedia pages describe a specific topic/concept in a semi-structured free-text form, Wikidata provides a knowledge-graph that enables semantic reasoning, and Wikimedia Commons provides free high-quality media. In this work, we posit that this framework has the potential to unlock a new line of research towards truly task-agnostic and open-domain visio-linguistic representation learning systems.

Achieving such fully context and knowledge-aware models is key to provide users with tools that facilitate the access to years of information, and enable the study of the evolution of certain topics and concepts of interest. For example, it will help answering questions such as - "How was this visual concept described textually 10 months ago, and how did it change?" - or - "How any two photos, from distinct yet of the same nature events, and taken years apart, are related?". Answering such questions requires explicitly accounting for the topic semantics, framing it in space and time, and unveil visual-textual relations for each particular context.

Current state-of-the-art task-agnostic vision and language representation learning systems [2], trained on paired data (image + text), perform quite well at bridging vision and language in closed domains. However, they operate in an highly restrictive setting, in which the concepts involved must be directly materialized in an image (e.g. a bus next to a building), and no open-domain domain-knowledge is explicitly considered. While some initial attempts to incorporate domain-knowledge in such systems have been made [1, 3], the scale and diversity of both the tasks and datasets considered are still quite limiting.

[1] https://www.wikimedia.org/

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452346

Towards evolving these existing models to be fully context and knowledge-aware, while coping with open-domain data, we decompose the general problem and propose a set of novel independent tasks to be tackled by researchers. Furthermore, we describe a Wikimedia-based open-domain media data framework, illustrated in Fig. 1, that will enable tackling each of the proposed tasks.

2 Wikimedia as a Large-scale Open-domain Media Repository

Fig. 1 describes a full framework where rich media event data, with images, their textual descriptions, and timestamps, can be mapped to the domain knowledge-graph from Wikidata. More comprehensive textual descriptions can then be obtained from event Wikipedia pages. This poses a rich and connected environment to pursuit open-domain and knowledge-grounded representation learning systems. The scale of information available enables both deep learning, symbolic and other types of approaches.

3 New Generation of Visio-linguistic Models

Current task-agnostic self-supervised deep learning models, based on the transformer architecture, have established the state-of-the-art in multiple visio-linguistic tasks [2], such as retrieval, image captioning, visual question answering, and others. These tasks not only were approached by sticking to the visual domain and the concepts that can be materialized in it (e.g. a tree, person sitting in a chair), but also assume that images and texts are *context-less*.

Leveraging on the full open-domain data framework presented in Fig. 1 will allow taking the next step towards generalizing these models knowledge to an open-domain setting, which includes event domain-specific concepts. In such scenario, the same concept (visual or not) may be described differently, depending on its context (semantic and temporal). Therefore, this establishes a new line of research that will seek to evolve existing approaches, mostly based on deep learning, to such open-domain scenario, that must be fully context-aware. Having not only paired data (i.e. image + description), but also the domain semantic knowledge will be key to achieve such goal. Namely, with Wikidata, a new generation of media understanding models can be pursuit, **that can jointly reason over media, the knowledge graph and the temporal context**, to semantically categorize each media asset.

4 Emerging Tasks and Use-cases

Now we introduce the novel research tasks, resulting from the decomposed problem.

4.1 Open-Domain Media Captioning

This task aims at designing models that can produce a textual description/paragraph for an image. Unlike the standard formulation of this task, the goal is to be able to describe an image using not only what is *visible*, but also its high-level context. This results in a context-dependent image captioning task, where depending on the image domain, the event and timestamp, the same visual concepts may be described differently (see Fig. 1 for an example).

4.2 Open-domain Media Conversational Agents

Usually a multimodal conversational agent attends to user information needs through a conversation, where in each turn, the user intent is expressed through textual and visual inputs. Relevant initiatives are the TREC-CAsT[2] track, which benchmarks open-domain but text-only conversational agents, and multimodal open-domain visual dialog [1]. Our proposed task, combines these initiatives by involving both open-domain text and images during a conversation, and aiming to ground answers in the knowledge-graph, similar to the approach of [3].

4.3 Event Understanding and Social Media

This tasks builds upon previous efforts in real-world event categorization and summarization, by seeking models that through the exploitation of image-text relations and knowledge-graph information, can structure event media and automatically create visually illustrated event storyline digests. These storylines are expected to convey the timeline of how events unfolded. Media linking between Wikimedia and social media could a) be used to further study how an event impacted society, by analyzing users reactions, and b) as a complimentary (and more immediate) information source, that in turn can enrich both the generated storylines and Wikimedia.

4.4 Cross Space and Time Media Visualization

Given a model that can capture media context, that understands how it relates to other topics or events, and that is able to ground that context in the knowledge-graph, how can we empower the user, by making all this information available through comprehensive yet compelling visualizations? The purpose of this task is to research interactive visualization systems that will enable users to navigate over years of media documents, either by interactively exploiting visual and knowledge-graph relations (space), or data timelines (time) such as the one in Fig. 1.

5 From Preliminary to Future Work

Preliminary experiments using Flickr[3] CC multimodal event media[4], have shown the importance of accounting for temporal context when bridging vision and language in an open-domain setting. We recently proposed a diachronic cross-modal embedding model [4], that explicitly frames media in time. However, the proposed approach is limited by two main aspects: a) the quality of data, due to its social media nature, and b) the lack of a semantic knowledge-graph where concepts and topics relations can be exploited.

Towards addressing these limitations, the first step will be to collect a representative snapshot of Wikimedia, comprising an initial set of real-word events, where images, their corresponding descriptions, and event knowledge-graphs will be collected. Then each of the proposed tasks will be formally defined. The collection will then be shared with the community to foster progress in each of the previously defined tasks. We will start by researching how to adapt models such as ViLBERT to the open-domain setting, and evaluate it on the defined tasks.

6 Acknowledgments

This work has been partially funded by the project PT2020 SmartyFlow, Ref. LISBOA-01-0247-FEDER-017283, the iFetch project, Ref. 45920 co-financed by ERDF, COMPETE 2020, NORTE 2020, and by the project NOVA LINCS Ref. UID/CEC/04516/2013.

[2] https://www.treccast.ai/
[3] https://www.flickr.com/
[4] https://novasearch.org/multimodal-diachronic-models/

References

[1] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

[2] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-Task Vision and Language Representation Learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] David Semedo and Joao Magalhaes. 2019. Diachronic Cross-Modal Embeddings. In *Proceedings of the 27th ACM International Conference on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 2061–2069.