Web Image Context Extraction with Graph Neural Networks and Sentence Embeddings on the DOM tree

Chen Dang, Hicham Randrianarivo, Raphaël Fournier-S'niehotta, Nicolas Audebert

ECML PKDD 2021
VIRTUAL
13-17 September

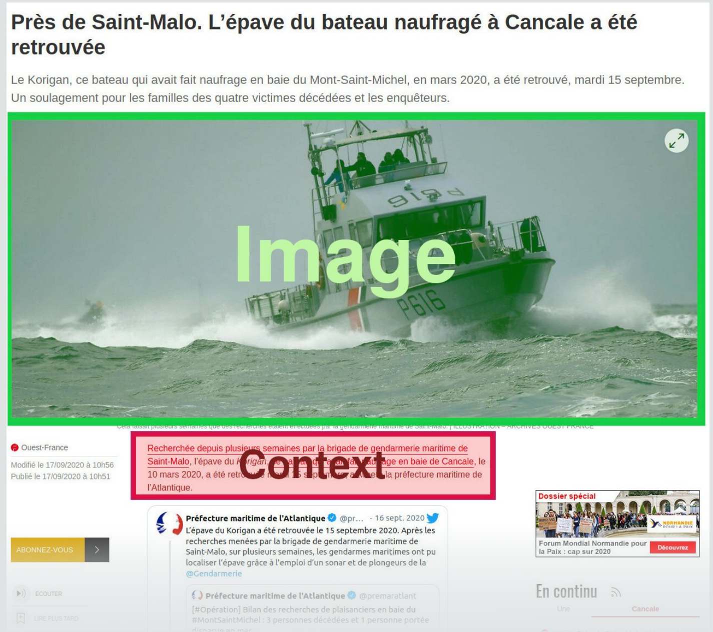
This paper introduces a novel approach for Web Image Context Extraction (WICE) that combines Graph Neural Networks (GNNs) and Natural Language Processing models.

Introduction

- ❖ Identifying the text in a webpage that best describes an **image** is a key step for efficiently **indexing images** in a **search engine**
- ❖ Visually rendering the webpage facilitates the extraction of an image's context, but isn't tractable on a large scale

Our Contribution

- ❖ Use state-of-the-art language models to generate sentence embeddings for each text node in the DOM tree



Près de Saint-Malo. L'épave du bateau naufragé à Cancale a été retrouvée

Le Korigan, ce bateau qui avait fait naufrage en baie du Mont-Saint-Michel, en mars 2020, a été retrouvé, mardi 15 septembre. Un soulagement pour les familles des quatre victimes décédées et les enquêteurs.

Cela faisait plusieurs semaines que des recherches étaient effectuées par la gendarmerie maritime de Saint-Malo. | ILLUSTRATION – ARCHIVES OUEST FRANCE

Ouest-France
Modifié le 17/09/2020 à 10h56
Publié le 17/09/2020 à 10h51

Recherchée depuis plusieurs semaines par la brigade de gendarmerie maritime de Saint-Malo, l'épave du Korigan, [illegible] en baie de Cancale, le 10 mars 2020, a été retrouvée [illegible] la préfecture maritime de l'Atlantique.

Préfecture maritime de l'Atlantique @pr... · 16 sept. 2020
L'épave du Korigan a été retrouvée le 15 septembre 2020. Après les recherches menées par la brigade de gendarmerie maritime de Saint-Malo, sur plusieurs semaines, les gendarmes maritimes ont pu localiser l'épave grâce à l'emploi d'un sonar et de plongeurs de la @Gendarmerie

Préfecture maritime de l'Atlantique @premaratlant
[#Opération] Bilan des recherches de plaisanciers en baie du #MontSaintMichel : 3 personnes décédées et 1 personne portée disparue en mer

Fig 1: An example of the WICE setting

- ❖ Use sentence embeddings as node features to train a GNN, which can combine both structural and semantic information
- ❖ Use graphe models for large-scale processing of highly diverse news websites

Architecture

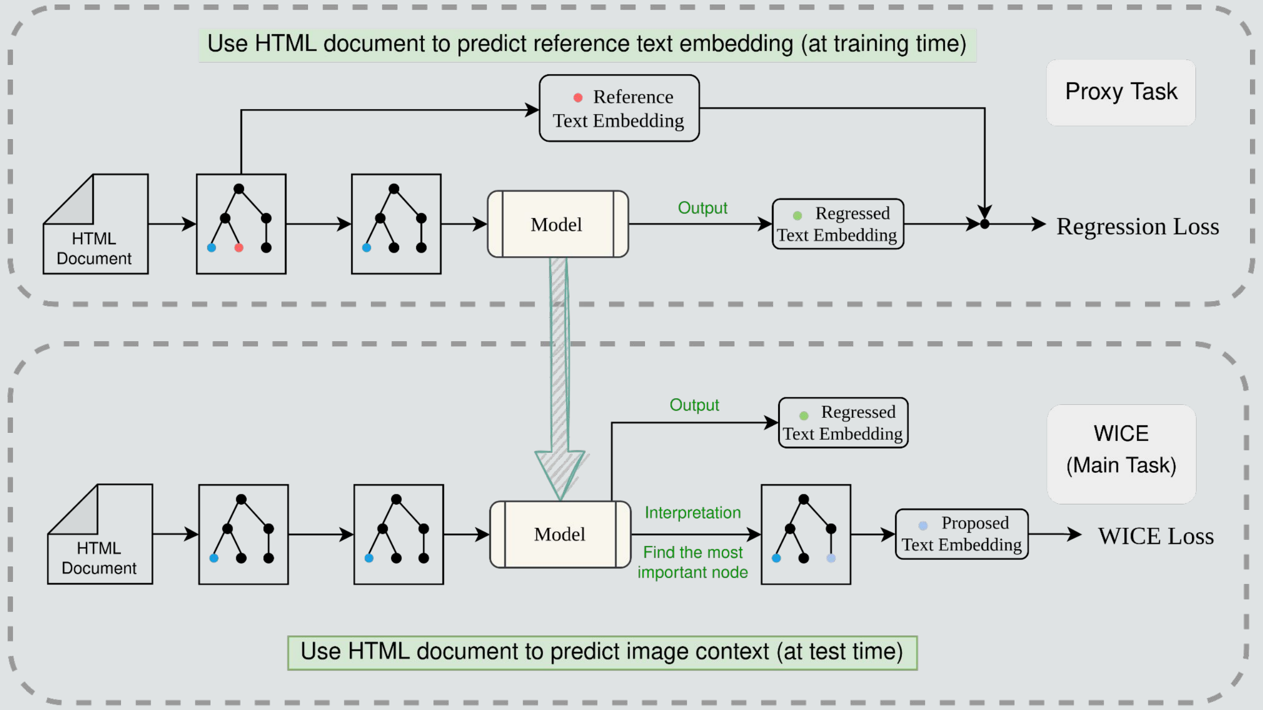


Fig 2: Our pipeline

- ❖ **Proxy task:** train a GNN model to predict the input document's reference text (red dot)
- ❖ **Main task:** interpret the trained model to choose the most predominant textual node (green dot) is then used as the context of the image

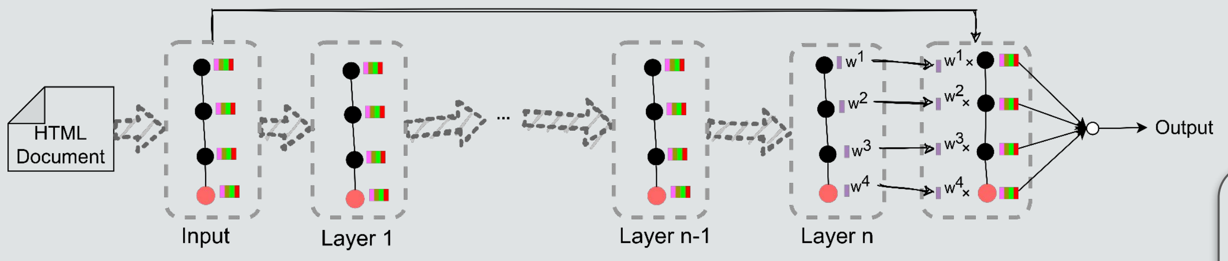


Fig 3: weight-GCN model architecture

Experiments

Method	train	validation	test
random	0.78	0.779	0.779
title	0.834	0.835	0.833
text after image	0.671	0.672	0.67
wGCN	**0.381**	**0.386**	**0.381**
oracle	0.293	0.297	0.293

Table1: average cosine similarity loss between the context and the reference text [split dataset by webpages]

Method	train	validation	test
random	0.792	0.736	0.800
title	0.834	0.861	0.814
text after image	0.701	0.571	0.705
wGCN	**0.415**	**0.404**	**0.441**
oracle	0.334	0.264	0.259

Table2: average cosine similarity loss between the context and the reference text [split dataset by websites]

Our model significantly outperforms WICE based on heuristics, and can work directly with HTML, making large-scale WICE tractable.

Chen DANG: chdangg@gmail.com