A new machine learning strategy that could enhance computer vision

July 16 2018, by Ingrid Fadelli

The model is capable of learning features that encode well the semantic content of the images. Given an image query (image on the left), the model is able to retrieve images which are semantically similar (depict the same type of object), although they might be visually dissimilar (different colours, backgrounds or compositions). Credit: arXiv:1807.02110 [cs.CV]

Researchers from the Universitat Autonoma de Barcelona, Carnegie Mellon University and International Institute of Information Technology, Hyderabad, India, have developed a technique that could allow deep learning algorithms to learn the visual features of images in a self-supervised fashion, without the need for annotations by human researchers.

To achieve remarkable results in computer vision tasks, deep learning algorithms need to be trained on large-scale annotated datasets that include extensive information about every image. However, collecting and manually annotating these images requires huge amounts of time, resources, and human effort.

"We aim to give computers the capability to read and understand textual information in any type of image in the real-world," says Dimosthenis Karatzas, one of the researchers who carried out the study, in an interview with *Tech Xplore*.

Humans use textual information to interpret all situations presented to them, as well as to describe what is happening around them or in a particular image. Researchers are now trying to give similar capabilities to machines, as this would vastly reduce the amount of resources spent on annotating large datasets.

In their study, Karatzas and his colleagues designed computational models that join textual information about images with the visual information contained within them, using data from Wikipedia or other online platforms. They then used these models to train deep-learning algorithms on how to select good visual features that semantically describe images.

As in other models based on convolutional neural networks (CNNs), features are learned end-to-end, with different layers automatically learning to focus on different things, ranging from pixel level details in the first layers to more abstract features in the last ones.

The model developed by Karatzas and his colleagues, however, does not require specific annotations for each image. Instead, the textual context where the image is found (e.g. a Wikipedia article) acts as the supervisory signal.

In other words, the new technique created by this team of researchers provides an alternative to fully unsupervised algorithms, which uses non-visual elements in correlation with the images, acting as a source for self-supervised training.

"This turns to be a very efficient way to learn how to represent images in a computer, without requiring any explicit annotations – labels about the content of the images – which take a lot of time and manual effort to generate," explains Karatzas. "These new image representations, learnt in a self-supervised way, are discriminatory enough to be used in a range of typical computer vision tasks, such as image classification and object detection."

The methodology developed by the researchers allows the use of text as the supervisory signal to learn useful image features. This could open up new possibilities for deep learning, allowing algorithms to learn good quality image features without the need for annotations, simply by analysing textual and visual sources that are readily available online.

By training their algorithms using images from the internet, the researchers highlighted the value of content that is readily available online.

"Our study demonstrated that the Web can be exploited as a pool of noisy data to learn useful representations about image content," says Karatzas. "We are not the first, nor the only ones that hinted towards this direction, but our work has demonstrated a specific way to do so, making use of Wikipedia articles as the data to learn from."

In future studies, Karatzas and his colleagues will try to identify the best ways to use image-embedded textual information to automatically describe and answer questions about image content.

"We will continue our work on the joint-embedding of textual and visual information, looking for novel ways to perform semantic retrieval by tapping on noisy information available in the Web and Social Media," adds Karatzas.

More information: TextTopicNet - Self-Supervised Learning of Visual Features Through Embedding Images on Semantic Text Spaces, arXiv:1807.02110 [cs.CV] arxiv.org/abs/1807.02110

© 2018 Tech Xplore

Citation: A new machine learning strategy that could enhance computer vision (2018, July 16) retrieved 23 April 2024 from https://techxplore.com/news/2018-07-machine-strategy-vision.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.