Neural Networks 76 (2016) 65–96

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Learning contextualized semantics from co-occurring terms via a Siamese architecture

Ubai Sandouk, Ke Chen *

School of Computer Science, University of Manchester, Manchester, M13 9PL, UK

ARTICLE INFO

Article history:
Received 11 August 2015
Received in revised form 28 October 2015
Accepted 13 January 2016
Available online 22 January 2016

Keywords:
Contextualized semantics
Descriptive terms
Siamese architecture
Out of vocabulary
Semantic priming
Representation learning

ABSTRACT

One of the biggest challenges in Multimedia information retrieval and understanding is to bridge the semantic gap by properly modeling concept semantics in context. The presence of out of vocabulary (OOV) concepts exacerbates this difficulty. To address the semantic gap issues, we formulate a problem on learning contextualized semantics from descriptive terms and propose a novel Siamese architecture to model the contextualized semantics from descriptive terms. By means of pattern aggregation and probabilistic topic models, our Siamese architecture captures contextualized semantics from the co-occurring descriptive terms via unsupervised learning, which leads to a concept embedding space of the terms in context. Furthermore, the co-occurring OOV concepts can be easily represented in the learnt concept embedding space. The main properties of the concept embedding space are demonstrated via visualization. Using various settings in semantic priming, we have carried out a thorough evaluation by comparing our approach to a number of state-of-the-art methods on six annotation corpora in different domains, i.e., MagTag5K, CAL500 and Million Song Dataset in the music domain as well as Corel5K, LabelMe and SUNDatabase in the image domain. Experimental results on semantic priming suggest that our approach outperforms those state-of-the-art methods considerably in various aspects.

1. Introduction

Multimedia information retrieval (MMIR) is a collective terminology referring to a number of tasks involving indexing, comparison and retrieval of multimedia objects (Jaimes, Christel, Gilles, Sarukkai, & Ma, 2005). As media content is created at an exponential rate, it has become increasingly difficult to manage even personal repositories of multimedia so as to make MMIR more and more demanding. Moreover, users expect certain levels of MMIR services from web service providers such as YouTube and Flickr. In addition, information processing tasks in fields such as medicine (Müller, Michoux, Bandon, & Geissbuhler, 2004) and education (Chang, Eleftheriadis, & Mcclintock, 1998) benefit enormously from advances in MMIR. In general, the most challenging problem in MMIR is the so-called *semantic gap* (Smeulders, Worring, Santini, Gupta, & Jain, 2000), which stems from the difficulty in linking low-level media representation, e.g., computationally extractable features, to high-level semantic concepts describing the media content, e.g., human-like understanding. Bridging this gap has motivated a number of approaches including feature extraction (Lew, Sebe, Djeraba, & Jain, 2006), user-inclusive design (Schedl, Flexer, & Urbano, 2013), and high-level context modeling (Marques, Barenholtz, & Charvillat, 2011). By modeling concepts, the use of *semantics*, i.e., the representation of high-level concepts and their interactions, leads to improvements in MMIR applications as well as the interpretability of the retrieved results (Kaminskas & Ricci, 2012). As a result, semantics acquisition and representation are critical in bridging the semantic gap. The richness, meaningfulness and applicability of semantics rely primarily on the sources of concept-level relatedness information. Examples of such sources include manually constructed knowledge graphs or ontologies (Kim, Scerri, Breslin, Decker, & Kim, 2008), automatically analyzed media content (Torralba, 2003) or well-explored collections of crowd-sourced descriptive terms or tags (Miotto & Lanckriet, 2012).

As one of the information sources, *descriptive terms*, including keywords, labels and other textual descriptions of media, have also been used in capturing the term-based semantics underlying co-occurring descriptive terms. Such semantics provides direct concept-level knowledge regarding the concerned multimedia objects. Typical applications include music crowd tagging services

* Corresponding author.
E-mail addresses: ubai.sandouk@cs.manchester.ac.uk (U. Sandouk), chen@cs.manchester.ac.uk (K. Chen).

http://dx.doi.org/10.1016/j.neunet.2016.01.004
0893-6080/

Nomenclature

Symbol	Definition
$x[i]$	The ith element of vector $\boldsymbol{x}$
$\|X\|$	The cardinality of the set X
Υ	Document-term relatedness matrix
τ	A single descriptive term
Γ	The collection of training descriptive terms
δ	A single document consisting of m descriptive terms
Δ	The collection of training documents
ϕ	A single topic produced from LDA analysis
Φ	The set of topics produced from LDA analysis of the training dataset
$\boldsymbol{t}(\tau)$	A representation of the descriptive term τ
$l(\tau\|\delta)$	A representation of the local context of term τ
$\boldsymbol{BoW}(\delta)$	The binary Bag of Words representation of document δ
$\overline{\boldsymbol{BoW}}(\delta)$	The binary complement of $\boldsymbol{BoW}(\delta)$
$\boldsymbol{x}(\tau, \delta)$	The collective representation of a term in context, i.e. $(\boldsymbol{t}(\tau), \boldsymbol{l}(\tau\|\delta))$
h	Index of layer in the neural network
$W_h, \boldsymbol{b}_h$	Weight matrix and biases pertaining to layer h
$\boldsymbol{CE}(\tau\|\delta)$	Contextualized embedding representation of a term in context
$\boldsymbol{E}(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$	Euclidean distance between two terms' CE representations
$\mathbb{E}$	The abbreviated notation of CE Euclidean distance
$\mathbb{d}$	The abbreviated notation of KL-divergence
$\mathbb{S}$	KL-divergence based similarity metric for local context
τ_{oov}	Out of vocabulary (OOV) descriptive term
$\boldsymbol{t}(\tau_{oov})$	OOV descriptive term representation
δ_{iv}	In-vocabulary terms in a document containing an OOV term
$\boldsymbol{CE}(\boldsymbol{x}(\tau_{oov}, \delta_{iv}))$	Feature-based semantic representation of an OOV term
$\boldsymbol{CE}(\tau_{oov}\|\delta_{iv})$	Concept-based semantic representation of an OOV term

(Law, Settles, & Mitchell, 2010) and multi-object image dataset analysis (Rabinovich, Vedaldi, Galleguillos, Wiewiora, & Belongie, 2007). Thanks to crowd-sourced annotation (Turnbull, Barrington, & Lanckriet, 2008) and game-based tags collection (Law, Ahn, Dannenberg, & Crawford, 2007), large collections of descriptive terms are now available. Those term collections can be analyzed for occurring patterns to reveal concept-level relatedness and similarity. Term-based semantics is expected to be transferable since it is acquired from high-level concepts independent of any specific MMIR tasks. It is worth stating that term-based semantics is different from linguistic semantics. First of all, descriptive terms are not only words but also symbols, abbreviations and complete sentences, e.g., "r'n'b" (musical style), "90s" (musical type), "stack of books" (visual concept), and so on. Next, descriptive terms may have a domain specific meaning different from their common linguistic meaning, e.g., "rock" is genre in music (not an earth substance) and "horn" is an instrument in music but is also a visual concept in images. Finally, the vocabulary used for descriptive terms is subject to change in time and cannot be fixed to represent a closed set of concepts. Those distinctions limit the usability of available linguistic resources such as linguistic dictionaries and generic word embedding from capturing term-based semantics. Therefore, we believe that the rich semantics conveyed in descriptive terms should be better explored and exploited to bridge the semantic gap.

By close investigation of various descriptive terms collections, we observe that terms could be used differently to represent various types of semantics and relatedness: (a) a term may have multiple meanings and the intended meaning cannot be decided unless the term co-occurs with other coherent terms, e.g., the term "guitar" can refer to an acoustic guitar when it co-occurs with terms like "strings", "classical", and so on, or to an electric guitar when it co-occurs with terms such as "metal", "rock", and so on; (b) different terms may intend the exact same meaning regardless of context, e.g., "drums" and "drumset"; (c) different terms may have either similar or different meaning depending on context, e.g., "trees" and "forest" convey similar concepts and have similar meaning in context of natural scene (conveying a concept of many trees) but "tree" is by no means similar to "forest" when used in description of an urban scene; (d) different terms may share partial meaning but have different connotations, e.g., "house" and "building" convey some similar concepts but "building" has a wider connotation; and (e) co-occurring terms may not have their meanings in singularity or in pair but in group only, e.g., {"wing", "tail", "metallic"} together define a concept of an airplane while {"leg", "cat", "tail",etc.} collectively present a concept of a cat and its body parts. The observations described above indicate the complexity and the necessity of taking the context into account in semantic learning from terms. Obviously, simply counting co-occurrence (Rabinovich et al., 2007) is insufficient in modeling various types of semantics and relatedness in descriptive terms to capture accurate concepts, and more sophisticated techniques are required so that we can capture all the intended semantics, or concepts and their relatedness, in descriptive terms accurately.

In general, a set of m terms, $\delta = \{\tau_i\}_{i=1}^{m}$, is often used collectively to describe the semantics underlying a single multimedia object where τ_i is a descriptive term and δ is the collective notation of the m terms, named *document* hereinafter. Furthermore, all m terms appearing in a document δ are dubbed as *accompany* terms. Our observation reveals that for a specific term τ_i in a document δ, the accompany terms jointly create its contextual niche, named *local context*, that helps inferring the accurate intended meaning of τ_i in that situation. In other words, the term along with its local context uniquely defines a concept of the accurate meaning. By taking such local contexts into account, we would learn a new type of relatedness between terms, named *contextualized relatedness*, by exploring terms' co-occurrence in different documents in a collection. Unlike the *global relatedness* where relatedness of terms is fixed irrespective of their local contexts, the contextualized relatedness of two terms is subject to change in the presence of different local contexts. In order to represent such contextualized semantics, we would embed all terms in a concept representation space that reflects the contextualized relatedness of terms. Formally, this emerging problem is formulated as follows: given a term τ and its accompany terms in δ, we would establish a mapping: $(\boldsymbol{t}(\tau), \boldsymbol{l}(\tau|\delta)) \rightarrow \boldsymbol{CE}(\tau|\delta)$, where $\boldsymbol{t}(\tau)$ and $\boldsymbol{l}(\tau|\delta)$ are the feature vectors of the term τ and its local context in δ and $\boldsymbol{CE}(\tau|\delta)$ is a concept embedding representation of τ given its local context in δ, so that the contextualized semantic similarity of terms be properly reflected via a distance metric in the concept embedding representation space. This is a challenging problem due to the actual facts as follows: (a) terms get their meaning in groups rather than in singularity or in pair; (b) it is unclear how to capture intrinsic context in terms; and (c) terms that are not seen in training may appear in application runtime and hence may confuse a semantic learning model, this issue is known as *out-of-vocabulary* (OOV) issue in literature. Nevertheless, solving this problem brings us closer to bridging the semantic gap as a solution to this problem not only provides a term-level contextualized semantic representation, named *concept embedding* (CE) hereinafter, for a term to grasp an accurate concept as well as contextualized concept relatedness but also the representations of co-occurring terms in a document collectively form a novel document-level representation precisely modeling the concepts in groups