Context-Aware Knowledge Extraction from Legal Documents Through Zero-Shot Classification

Context-Aware Knowledge Extraction from Legal Documents through Zero-Shot Classification

Alfio Ferrara[1][0000−0002−4991−4984], Sergio Picascia[1], and Davide Riva[1]

Università degli Studi di Milano
Department of Computer Science
via Celoria, 18 - 20133, Milano, Italy
name.surname@unimi.it

Abstract. The extraction of conceptual and terminological knowledge from legal documents is a crucial task in the legal domain. In this paper we propose ASKE (Automated System for Knowledge Extraction), a system for the extraction of knowledge that exploits contextual embedding and zero-shot learning techniques in order to retrieve relevant conceptual and terminological knowledge from legal documents. Moreover, in the paper we discuss some preliminary experimental results on a real dataset consisting of a corpus of Illinois State Courts' decisions taken from the Caselaw Access Project (CAP).

Keywords: legal knowledge extraction · legal document retrieval · zero-shot learning

1 Introduction

The extraction of knowledge from large textual corpora of legal documents is a crucial task for providing useful and relevant suggestions to legal actors, such as judges and lawyers, in handling new incoming cases. One of the main challenges in this context is that legal documents use a peculiar language and terminology, which makes it difficult to correctly classify and retrieve documents with unsupervised techniques based on standard information retrieval technologies. On the other hand, exploiting supervised learning techniques is also difficult due to the absence of sufficiently large annotated corpora.

In this paper we propose ASKE (Automated System for Knowledge Extraction), a system for the extraction of knowledge that exploits contextual embedding techniques to manage the meaning of legal concepts and terms taking into account the particular context in which they are used [1]. Moreover, ASKE operates cyclically in a completely unsupervised environment, exploiting zero-shot learning techniques [4]. At each cycle, ASKE classifies documents chunks according to a set of initial concepts, and retrieves relevant conceptual and terminological knowledge from them. This information will finally contribute to the enrichment of the ASKE Conceptual Graph (ACG).

In order to evaluate the ASKE performances, we discuss some preliminary experimental results on a real dataset consisting of a corpus of Illinois State Courts' decisions taken from the Caselaw Access Project (CAP). Finally, the paper presents an example of how ASKE can be exploited as a support tool for jurists, helping them in retrieving relevant legal precedents.

The paper is organized as follows. In Section 2, we present the ASKE methodology. In Section 3, we describe the conceptual model of the ASKE Conceptual Graph. In Section 4, we present some preliminary results obtained by evaluating ASKE on a corpus of real legal documents. In Section 5, we present the related work. In Section 6, we discuss our conclusing remarks.

2 The ASKE methodology

Knowledge extraction in ASKE is performed through a sequence of operations that constitutes a cycle. Each cycle can be repeated for a predefined number of iterations, called generations. The extracted knowledge populates the ASKE Conceptual Graph (ACG), which is continuously updated at each cycle, allowing the model to incrementally classify documents, retrieve terminology and form new concepts, generation after generation. The four main phases of the ASKE cycle are shown in Figure 1.

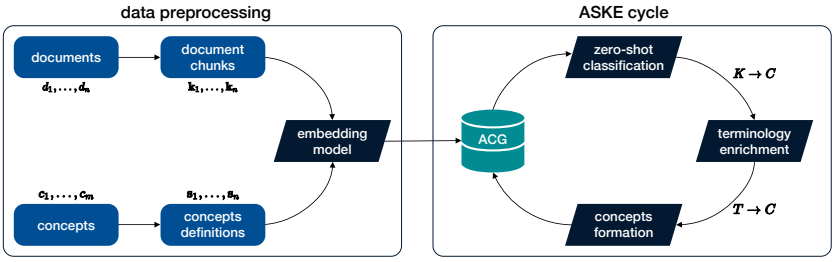

Fig. 1: Data preprocessing and ASKE cycle.

Data Preprocessing. The two main inputs of our model are a corpus of legal documents and a set of initial concepts. In particular, we consider a corpus of text documents $D = d_1, \ldots, d_n$, such as case law decisions or legal codes. Each document is preprocessed using a tokenizer that splits the text in a set of document chunks k and provides the lemmatized version of the terms t that

constitute the documents. The set of initial concepts $C = c_1, \ldots, c_m$ can be provided by the user according to three different strategies: (i) the user provides just one or more terms t to use as initial concepts; (ii) the user provides one or more definitions s for each initial concept in form of a short text, such as a sentence; (iii) the user selects one or more concepts from an existing knowledge base.. In case (ii) and (iii), each concept is already associated with one or more definitions s, given by the user in (ii) or provided by the knowledge base in (iii). In the first case, instead, we retrieve the definitions of each term associated to a concept from an external knowledge base, e.g. WordNet, by taking one definition s for each of the possible senses associated to a term. The last step of data preprocessing is to transform document chunks k and concept definitions s in their corresponding vector representation, projecting them in the same semantic space. This is achieved using Sentence-BERT [1], a modification of the original BERT model, which exploits siamese and triplets networks, being able to derive semantically meaningful sentence embeddings. The inputs for the embedding model will therefore be the set of document chunks K and the set of concept definitions S. Since a concept c_i may be associated to multiple definitions $s_{i1}, \ldots, s_{ij}$, we define its position in the embedding space as the centroid $\mathbf{c}_i$ of $\mathbf{s}_{i1}, \ldots \mathbf{s}_{ij}$. Document chunks and the set of initial concepts constitute together the first version of the ACG.

Zero-Shot Classification. Given the initial version of the ACG, ASKE is ready to perform the zero-shot multi label classification. In this phase, document chunks are assigned to none or multiple concepts, $f : K \to C$. A similarity measure σ, e.g. cosine similarity, between the embedding vector $\mathbf{k}$ of each document chunk and the embedding vector $\mathbf{c}$ of each concept is computed and, eventually, the two are associated if this similarity is higher than a predefined threshold α:

$$f(K, C) = \{c_i : \sigma(\mathbf{k}_j, \mathbf{c}_i) \geq \alpha\}$$

The hyperparameter α is crucial since it may remarkably affect the classification output: for example, choosing a high value of α will result in a high value of precision in the classification but potentially may find only a small set of document chunks for each concept.

Terminology Enrichment. For each concept c_i, ASKE retrieves the set of terms T_i appearing in the document chunks K_i associated to it. Then, these terms are placed in the embedding space computing the vector representation of their definition(s) s_{t_i} retrieved from an external knowledge base. In the case of polysemic terms, only the definition whose embedding is closest to the concept $\mathbf{c}_i$ is maintained. For each candidate terms the similarity σ between their embedding vector and the ones of the concept and of the document chunks is computed. The terms whose similarity is greater than the parameter β are taken into account and become candidates for enriching the terminology of the concept.

$$g(K, C, S_t) = \{t : \sigma(\mathbf{s}_{t_i}, \mathbf{c}_i) + \sigma(\mathbf{s}_{t_i}, \mathbf{k}_i) \geq \beta\}$$

The set of candidate terms is then sorted in descending order according to the similarity score. In addition, a learning rate γ can also be defined, representing the maximum number of terms which will be associated to a certain concept at each generation. Applying an upper bound γ and lower bound β ensures that, at each cycle, the process of terminology enrichment will include only a small set of terms that are supposed to be meaningful with respect to the concept at hand.

Concepts Formation. As a last phase, ASKE may introduce new concepts in the ACG by a process of concept formation. This process consists in applying, for each concept c_i, a clustering algorithm, such as affinity propagation [2], over the embedding vectors S_i of the terms t_i belonging to it. Among the resulting clusters, the one including the original definition of the concept c_i will continue to represent the original concept, preserving its existence in the ACG. For each remaining cluster, a new concept c_j will be generated, and its label will be represented by the term closest to the center of the cluster itself. Moreover, a derive from relation between c_i and c_j is defined in ACG.

3 The ASKE Conceptual Graph

The knowledge extracted during the ASKE execution cycles is stored in the ASKE Conceptual Graph (ACG), that is organized as shown in Figure 2.

ACG is composed by two main entities, namely Concept and Term. A concept is associated with a label that represents its meaning in a synthetic and understandable way to a human user. This label is selected from the terms associated with the concept and corresponds to the term whose vector is closest to the mean of the vectors of all the terms associated with the concept. Concepts are also linked together by a relationship that describes how they were generated. In particular, the relationship derive from between two concepts c_i and c_j represents the fact that the concept c_j derives from the concept c_i, that is, it was formed starting from an aggregation of terms that were initially associated with c_i. As in all the ACG relations, also the relationship of derivation between two concepts is associated with the knowledge extraction cycle (generation) in which a concept was formed and the degree of similarity that associates a concept with the concepts derived from it. The concepts are also associated with the document chunks by the classification relationship. Note that ASKE performs a multi-label classification process, whereby a document chunk can be associated with multiple concepts. Finally, a concept consists of a collection of terms. The terms are represented by the term entity which is uniquely identified by the sense, that is a reference to a possible specific meaning of the term, and by the definition associated with that sense. The belongs to relationship stores the information about the set of terms associated with a concept in a certain cycle of execution of ASKE (generation) and with a certain degree of similarity (similarity). During the knowledge extraction process, each term acquired by ASKE is derived from a concept. This particular relationship is represented by the derive from relationship. Note that, even if the term is then associated with a different concept,

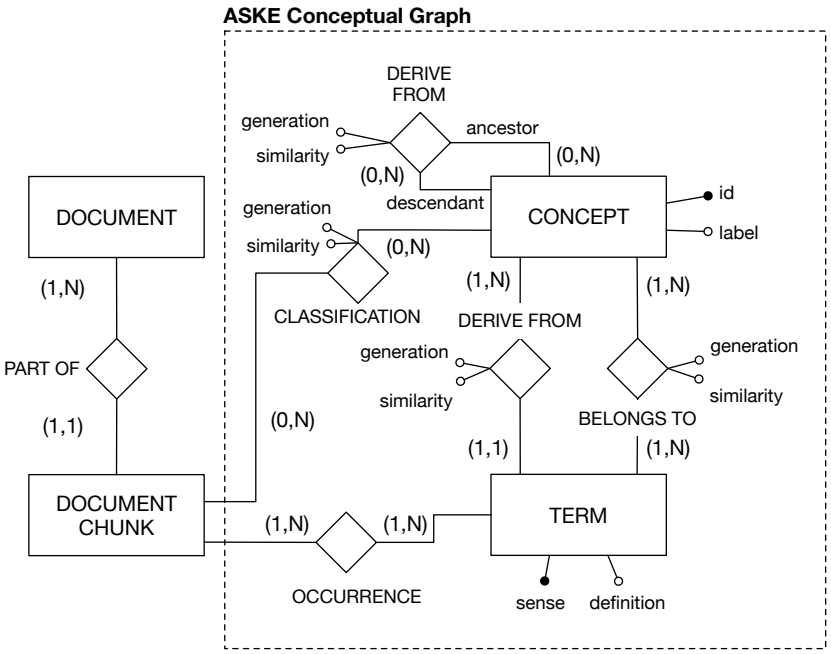

Fig. 2: Conceptual model of the ASKE conceptual graph

the derive from relation still allows us to reconstruct the concept that led to the discovery of the term at hand. Terms also have a relationship occurrence with the document chunks. This relationship represents the presence of a term in one or more specific document chunk.

An example of how ACG is used to store the knowledge extracted from ASKE from a corpus of legal documents is shown in Figure 3.

In the example, the concept C7, labeled Drug, was the initial concept. During the first generation (i.e., the first knowledge extraction cycle), C7 is associated, among the others, with the document chunk K1 with a similarity of 0.58 between the vector representation of K1 and the vector corresponding to C7. Through this association between C7 and the document chunks, new terms are extracted from the documents and associated with C7. In particular, a term associated with C7 is addition_0. Note that, in this example, where the term definitions are taken from WordNet, addition_0 corresponds to a particular synset of WordNet, whose definition is reported in the ACG. The knowledge extraction process is then repeated in subsequent generations, until, in the second generation, a group of terms initially associated with C7 and including addiction_0 is promoted to a new concept C14. One of the terms, namely drug abuse, is used to label the new concept and addiction_0 is now associated with it. However, we keep the derives from relation between addiction_0 and C7 to keep memory of the conceptual origin of addiction_0. In the subsequent generation 3, a new set of document chunks

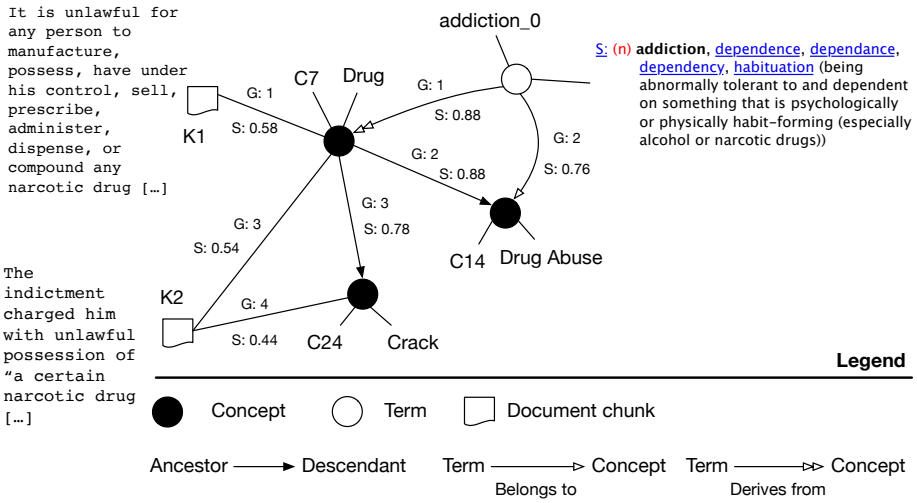

Fig. 3: Example of concepts, terms, and document chunks extracted by ASKE from a corpus of legal documents

is associated with C7, including K2. This new document chunks lead to the discovery of new terms and to the generation of new concepts, including the new C24 (Crack) concept. Then, in the subsequent generation 4, C24 is also associated with K2. Thus, K2 will be associated finally to more than one concept, namely, C7 (Drug) and, more specifically, C24 (Crack).

4 A Case Study on the Illinois Caselaw Corpus

Our aim here is to evaluate ASKE according to its ability to generate new terminological and conceptual knowledge starting from a small set of concepts and a possibly large corpus of documents. In the ASKE Conceptual Graph, we validate the concept derivation relationships, which link a concept to another that derived from it in any iteration, and the conceptual clusters of terms, i.e. the sets of terms that represent each concept, by submitting them to human evaluators.

Dataset. ASKE is applied to a subset of the Illinois Caselaw Corpus, a dataset from the CaseLaw Access Project[1]. Our dataset contains about 57,000 Case Law Decisions (CLDs) in the Illinois jurisdiction from 1771 to 2010. The CLDs were split into 9,5 million chunks, which constitute the input of ASKE together with a list of initial concepts, in the form of *(label; definition)* pairs. This procedure corresponds to the option (ii) described in Section 2, in which the definitions

[1] https://case.law

are provided by the users, here relying on the "Crime In Illinois 2020" report[2], issued by the Illinois State Police. The concepts correspond to 11 crime categories listed in the report, namely: homicide, rape, robbery, assault (also present with label battery), burglary, theft, vehicle theft, arson, human trafficking for commercial sex purposes (here labeled as prostitution) and human trafficking (for servitude purposes). Drug-related crimes, treated separately in the report, were added under the label drug, whose definition is taken from an American legal dictionary.[3]

Evaluation process and results. The evaluation process is twofold: in the ASKE Conceptual Graph, we want to i) validate concept derivation relationships and to ii) assess the coherence of conceptual clusters of terms. For an initial assessment, we relied on a restricted crowdsourcing campaign consisting of two classes of tasks and involving 8 workers.

4.1 Concept derivation relationships

For the evaluation of concept derivation relationships, we asked the workers to choose, in a set of 4 concepts given with their definition, the one that most closely relates to a concept in the question (i.e. *"Among the following, what is a concept that closely relates to concept* HOMICIDE *defined as* THE WILLFUL (NON-NEGLIGENT) KILLING OF ONE HUMAN BEING BY ANOTHER?"). There were 84 such tasks, of which 56 included a concept that ASKE derived directly from the one in the question, another one that was derived indirectly, and two other random concepts or *"None of the above"*. Other 28 out of 84 did not include the directly derived concept. Each task was executed by a minimum of one worker up to a maximum of 5. The workers choice is the ground truth of our evaluation GA. To the end of the evaluation, ASKE performs well on a task when the concept derived by ASKE from the concept in the question (either directly or indirectly) is included in the ground truth, i.e., the concept(s) chosen by the workers. This is formally represented by the evaluation function $e(t_i)$ that returns 1 if the concept retrieved by ASKE is in GA and 0 otherwise. It is easy to see that for ASKE the task was more difficult when there is a consensus among the workers, because the number GA of ground-truth answers is lower. To take into account this different value of the tasks, we computed a task score θ_i for each task as $\theta_i = 1 - (\mid GA_i \mid /4)$, where $\mid GA \mid$ is the size of the ground truth for the task i. Θ_T is the sum of θ_i for all the tasks T. The overall performance of ASKE over the set T of tasks is then evaluated as:

$$ASKE = \frac{1}{\Theta_T} \sum_{i=1}^{|T|} e(t_i)\theta_i$$

As a baseline, we computed the score of a hypothetical systems that randomly chooses one concept for each task. Since the probability of selecting one of the

[2] https://isp.illinois.gov/CrimeReporting/CrimeInIllinoisReports
[3] https://dictionary.law.com/Default.aspx?selected=343

ground truth concepts is equal to $\mid GA \mid /4$, the performances of the baseline on a task i is evaluated as $(\mid GA \mid /4)\theta_i$. Moreover, we also computed the performances of ASKE by taking into account only the concepts directly derived from the concept in the task question (this last measure is called $ASKE_d$). The result of this evaluation is reported in Figure 4.

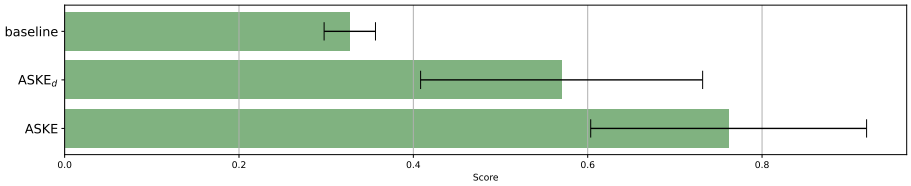

Fig. 4: Evaluation of ASKE concept derivation relationships

The evaluation was run on the Illinois Caselaw Corpus with the BERT model all-mpnet-base-v2[4] and 30 cycles. From this preliminary evaluation, we observe that ASKE may derive general or legal concepts by an accuracy of almost 80%. The $ASKE$ score is particularly affected by 11 tasks that received 4 out of 4 or 5 out of 5 answers not including the ASKE concept. Of such answers, 35 are related to the input concept Robbery, for which no worker identified legal concepts like Hearing, and Decree, nor generic concepts like Unreasonable as directly derived concepts.

4.2 Term clusters

For term cluster evaluation, we asked the workers to assess if two terms, and their corresponding definitions in WordNet, were semantically related one to the other. In this evaluation, we have a correct result if two terms belong to the same concept in the ACG and they are considered semantically related by the workers (True Positive). If instead two terms belong to the same ACG concept but are not related according to the crowd we have a False Positive. Finally, we have a False Negative if two terms are considered similar but do not belong to the same concept in ACG. By exploiting this criterion, we computed the weighted Precision and Recall of ASKE according to different levels of worker consensus with the decision of ASKE, as shown in Figure 5.

We note that in most of the tasks that have only one respondent the worker agrees with ASKE. The set of false positives is almost completely made of pairs of terms indicating components of the judicial process, the most common ones being: (action; trial), (trial; judgment), (trial; objection), (action; review). This suggests that ASKE is effective in grouping terms belonging to the judicial domain

[4] https://www.sbert.net/docs/pretrained_models.html

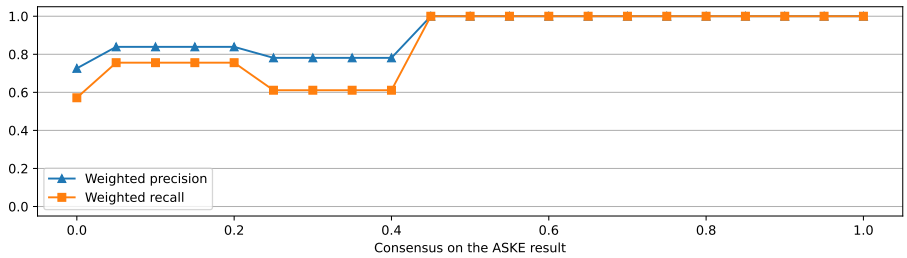

Fig. 5: Weighted precision and recall of ASKE according to the crowd consensus

together, although the process of forming new concepts can be improved by increasing the system's ability to recognize sub-concepts within a more general concept.

5 Related Work

The main topic of interest for our work has been the zero-shot learning (ZSL) approach. ZSL is a problem setup in the field of machine learning, where a classifier is required to predict labels of examples extracted from classes that were never observed in the training phase. It was firstly referred to as *dataless classification* in 2008 [3] and has quickly become a subject of interest, particularly in the field of natural language processing. The great advantage of this approach consists in the resulting classifier being able to operate efficiently in a partially or totally unlabeled environment.

It is possible to classify ZSL techniques according to three different criteria, as explained in [4]: the learning setting, the semantic space and the method. Firstly, ZSL can be applied on a completely unlabeled dataset, as in the original paper [3], or on a partially labeled one, like in [5]; with this last approach, called generalized ZSL, the goal of the classifier shifts to distinguishing between observation from already seen classes, and examples from unseen ones. Secondly, one may discern an engineered semantic space from a learned semantic space: the first is designed by humans and can be constructed upon a set of attributes [6] or a collection of keywords [7], while the second is built on top of the results of a machine learning model, as in the case of a text-embedding space [8]. Finally, ZSL methods can be divided in instance-based [9], whose focus is on obtaining examples for unseen classes, and classifier-based [10], which instead focus on directly building a classifier for unlabeled instances.

ASKE aims at classifying documents and extracting knowledge from them, building a conceptual graph on top of it. This process happens in a completely unsupervised environment, operating in a text-embedding space and applying an instance-based method; in particular, the employed method goes under the category of projection methods, which consist in labeling instances by collocating these examples in the same semantic space with class prototypes. It relies on a

very limited initial knowledge and it is able to address the problem of term disambiguation.

6 Concluding Remarks

We presented ASKE (Automated System for Knowledge Extraction), a system for the extraction of knowledge that exploits contextual embedding and zero-shot learning techniques in order to retrieve relevant conceptual and terminological knowledge from legal documents. The experimental results are still preliminary, but they encourage thinking that ASKE can be effective in identifying relevant concepts and terms in the legal context. Besides the extraction of concepts and terms, ASKE can also be used to retrieve past documents (such as CLDs) that refer to a certain concept of interest for a legal actor. Our future work will be in this direction. In particular, we aim at exploiting an ongoing collaboration with Italian lawyers and judges in order to further evaluate the effectiveness of ASKE as a tool for supporting the legal work.

References

1. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-networks. https://doi.org/10.48550/ARXIV.1908.10084
2. Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. science, 315(5814), 972-976.
3. Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: dataless classification. In *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2 (AAAI'08)*. AAAI Press, 830–835.
4. Wei Wang, Vincent W. Zheng, Han Yu, and Chunyan Miao. 2019. A Survey of Zero-Shot Learning: Settings, Methods, and Applications. ACM Trans. Intell. Syst. Technol. 10, 2, Article 13 (March 2019), 37 pages. https://doi.org/10.1145/3293318
5. Xian, Y., Lampert, C. H., Schiele, B., & Akata, Z. (2017). Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1707.00600
6. C. H. Lampert, H. Nickisch and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 951-958. https://doi.org/10.1109/CVPR.2009.5206594
7. Qiao, R., Liu, L., Shen, C., & Hengel, A. van den. (2016). Less is more: zero-shot learning from online textual documents with noise suppression (Version 1). arXiv. https://doi.org/10.48550/ARXIV.1604.01146
8. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., & Schiele, B. (2016). Latent Embeddings for Zero-shot Classification (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1603.08895
9. Xu, X., Hospedales, T., & Gong, S. (2015). Transductive Zero-Shot Action Recognition by Word-Vector Embedding (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1511.04458
10. Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Ranzato, M. A., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. Advances in neural information processing systems, 26.