Extending ImageNet to Arabic using Arabic WordNet

Abdulkareem Alsudais

Received: 11 September 2020 / Revised: 31 August 2021 / Accepted: 3 January 2022

Abstract

This paper investigates the extension of ImageNet and its millions of English-labeled images to Arabic using Arabic WordNet. The primary finding is the identification of Arabic synsets for 1219 of the 21,841 synsets used in ImageNet, which represents 1.1 million images. By leveraging the parent-child structure of synsets in ImageNet, this dataset is extended to 10,462 synsets (and 7.1 million images) that have an Arabic label, which is either a match or a direct hyponym, and to 17,438 synsets (and 11 million images) when a hyponym of a hyponym is included. Samples evaluated suggest that generating Arabic labels for images in ImageNet using hypernyms does indeed produce meaningful results. The precision values for seven evaluated samples exceeded 90%. Moreover, when all the images in the samples were combined, the precision value equaled 93%. For the entire ImageNet, when all hypernyms for a node are considered, an Arabic synset is found for all but four synsets. This represents the major contribution of this work: a dataset of 14,195,756 images that have Arabic labels. The resulting dataset presents Arabic labels for 99.9% of the images in ImageNet.

Keywords ImageNet · Computer vision · Arabic WordNet · Arabic computer vision · Language and computer vision · Linked data

1 Introduction

ImageNet is a dataset comprised of 14 million images [22, 56]. Each image in the dataset is labeled with a WordNet [47] noun synset representing the identifying object in the image. This integration with an established lexical knowledge base such as WordNet enables researchers to connect images and their labels to additional external resources. The Fall 2011 release of the dataset has a total of 21,841 unique synsets that are used to label images. The dataset is organized by dividing these synsets into several major subtrees. Moreover, ImageNet is
structured in a way that maintains the semantic hierarchical structure of synsets in WordNet, where each image is also linked to branches of hypernyms. One important feature of ImageNet is its fine-grained categories, which refer to similar classes (such as different breeds of dogs) with “subtle visual differences” [67]. ImageNet is often credited as a major influencer for the recent advances in computer vision and deep learning [19, 38, 63, 65], and many of the novel deep learning methods that have been highly impactful in the last few years have leveraged ImageNet [39, 64, 66]. Further, the dataset has been used to solve tasks in the intersection of language and vision research [32, 37, 78].

While computer vision research has seen significant progress in recent years, the focus has been on English. Limited work exists to extend research to other languages, including Arabic. This lack of research may present challenges to scientists, researchers, and practitioners who seek to address problems related to computer vision in Arabic. Furthermore, the unavailability of a large dataset of images labeled in Arabic may prevent the development of solutions that address challenging tasks, such as visual question answering and image classification in Arabic. Therefore, a large dataset of images labeled with Arabic has the potential to progress research in Arabic computer vision. Moreover, scholars studying Arabic natural language processing often develop methods specifically designed for Arabic. Thus, it is possible that similarly unique methods are needed for Arabic computer vision.

The primary objective of this paper is to investigate the effectiveness of extending ImageNet to Arabic using Arabic WordNet (AWN) by searching in AWN for all the synsets used in ImageNet. This paper is an extension of a paper presented at a workshop organized in conjunction with the 2020 annual conference of the association for computational linguistics [11]. AWN was originally developed in 2006 [14]. Since then, several authors have attempted to extend it by improving its coverage or quality [3, 5, 12, 15, 53]. Since AWN uses the same synset IDs used in other WordNets, it can be extended in several ways similar to how researchers have leveraged WordNet to propose novel solutions for research challenges. These solutions include creating bilingual embeddings based on the WordNets of six languages [30], extracting the semantic relationships and similarities between entities [43], and generating Bag-of-Concepts representations for document classification [44]. The possibility of using AWN to extend ImageNet has been experimented with in one paper [10]. In the paper, the author used AWN to find Arabic synsets for a small sample of 100 images from ImageNet and indicated that Arabic synsets were found for only six synsets. However, the author did not attempt to discover if Arabic synsets were available for hypernyms of these synsets when direct matches were not found. Similarly, BabelNet is a resource that links wordnets in the Open Multilingual WordNet to images in ImageNet [48]. However, to the best of my knowledge, it also does not go beyond the few Arabic synsets that can be directly matched with images in ImageNet.

In ImageNet, each synset is connected to a tree of related synsets that include hypernyms and hyponyms. A hypernym is a direct “parent” for the synset. The relationship between a hyponym and a hypernym is often referred to as an “is-a” relationship where the hyponym “is-a” hypernym. As shown in Table 1, there are images in ImageNet for which an Arabic synset for the synset is unavailable in AWN; however, Arabic synsets for the hypernym or the hypernym of the hypernym are available. Therefore, an Arabic synset can be added to an image by processing the branch, or many branches, of hypernyms available for each synset and only stopping when one is found.

In this paper, two approaches are used to determine whether Arabic synsets can be identified for synsets used in ImageNet to label images. First, a direct match for each synset
within ImageNet is searched for in AWN. Second, to overcome the problems of limited availability for direct matches, Arabic synsets are searched for in the branches of hypernyms connected to each synset in ImageNet. Since using hypernyms to label images is a novel approach, an evaluation is conducted on samples of the data. The objective of this evaluation is to confirm the ability to provide meaningful Arabic labels using hypernyms. Finally, two subsets of ImageNet are tested.

The rest of the paper is organized as follows: Section 2 highlights the research objectives; Section 3 introduces related works; Section 4 describes the methods employed to extend ImageNet to Arabic using AWN and explains the subsets of ImageNet used; Section 5 provides all of the details regarding the results and the evaluation; Section provides a discussion; and Section 6 offers a conclusion.

### 2 Research objectives

The primary motivation of this paper is to discover if AWN synsets can be found for the images used in ImageNet. More specifically, Arabic WordNet is tested to examine its ability to link English synsets used in ImageNet to Arabic synsets in AWN. To the best of my knowledge, there is no large dataset of images labeled with Arabic. This presents challenges to scientists interested in research questions related to Arabic computer vision as well as practitioners eager to develop new solutions that require large datasets of images labeled in Arabic. This also represents a major motivation for this work: create a new and large dataset of images labeled in Arabic. The second major objective is to experiment with using AWN hypernyms to labels images. If such a method succeeds, the number of images in ImageNet with Arabic labels can be increased. Moreover, the same method can be used to extend ImageNet to other languages.

In summary, this paper has three key contributions:

- It investigates the possibility of extending ImageNet and its millions of images to Arabic using Arabic WordNet (AWN).
- It tests and evaluates using synsets’ branches of hypernyms to find AWN synsets when direct matches are not available.
Most importantly, it yields a new, large dataset of images with Arabic labels that were found directly or using hypernyms.

3 Related work

3.1 Language and computer vision

A large dataset of images labeled in a selected language can be of great significance for language and computer vision research. Such research represents an emergent area of study that addresses several challenging tasks [18, 33, 69]. One example of work in this area involves the generation of captions for images [23, 45]. The objective is to automatically interpret the contents of a given image in order to create captions. Similarly, there is work to create new images by processing sentences written in English [52]. In the paper, the authors created a method that relies on understanding the details provided in the input sentence, such as colors or sizes of birds, for instance. Another common task is visual question answering [28, 57]. The objective is to process a user’s question about an image and then return an answer. For example, if a user asks, “How many children are wearing black?” then the provided answer will contain the correct number of children based on 1) processing the image to understand its content and 2) accurately interpreting the question to comprehend its purpose. Another related task is video question answering, which focuses on answering users’ inquiries based on the contents of reference videos [74]. Some recent papers proposed new frameworks to classify the sentiments of contents that have both text and images [34, 40, 72]. A similar area pertains to visual dialog [76]. The goal, in this case, is to enable an “AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content” [21]. A task related to this field entails linking entities in the textual content of a document to the corresponding entities in the visual content of the same document [24]. Finally, recent work also targets the modification of images based on instructions provided in English [70]. Additionally, this work is motivated by recent research that focus on extending tagged images using semantics [41, 42]. In summary, these papers provide examples of research challenges at the intersection of language and computer vision. One common feature of these papers is their focus on novel solutions for English. Therefore, in order to develop similar solutions for tasks related to Arabic computer vision—such as Arabic question answering, visual dialog, and image captioning—a crucial initial step is to create a large dataset that can then be customized for these individual tasks.

3.2 Computer vision for other languages

Several authors have worked on computer vision research in languages other than English. In one paper, the authors extended a sample of images in ImageNet to German using human subjects [54]. Their dataset included 1,305,602 images. Another work in German focused on the creation of a new dataset that extends Flickr30K (which has images labeled in English) using crowdsourcing and professional translators [25]. This dataset has been used as a benchmark in shared tasks in which the objective is to process an image with a label in one language and generate a description in other languages. There are two other noteworthy large datasets. The first is comprised of 164,062 images and 820,310 captions in Japanese [75]. The second is entitled “VATEX” and consists of over 41,250 videos and 825,000 English or
Chinese captions [71]. Finally, “How2” is a dataset of 80,000 clips with English subtitles for their Portuguese translations [60]. These papers highlight the potential impacts of new datasets of images and videos labeled with languages other than English and how they, similar to the data presented in this paper, can be used to study several challenging tasks related to language and computer vision research.

### 3.3 Arabic computer vision

Limited research exists related to Arabic computer vision. Several authors have generated Arabic captions for images [6, 36], while others have created new datasets related to Arabic computer vision. In one paper, the researchers constructed a dataset of 3000 Arabic clips that they classified with one of the following emotional labels: “happy,” “sad,” “angry,” “surprise,” “disgust,” and “neutral” [62]. Another paper presented a new dataset that consists of 1100 videos of speakers who read 10 different Arabic words [27]. Other research in Arabic computer vision includes digitalization of historical Arabic text using methods that rely on processing these documents as images [35, 51]. In the same research domain, the authors of another paper discovered that typical optical character recognition software (OCR) does not perform well when processing documents in Arabic due to OCR’s inability to “understand the layout of an Arabic document” [58]. In the only other closely related paper, the author investigated the possibility of generating Arabic labels for images in ImageNet using an online translator [10]. In the paper, the author targeted a sample of 1895 images from ImageNet and used an online translator to generate Arabic labels for the synsets. A human judge then evaluated the accuracy of the translations. The results indicated that the translations were accurate for 65% of the images, which represented 1643 unique synsets and 1,910,935 images. This suggests that solely using a translator to translate labels of images in ImageNet may not produce highly accurate results.

Many research areas at the intersection of language and computer vision may indeed need solutions specifically designed for Arabic, however, this may not always be the case. Visual sentiment analysis (which is also referred to as image sentiment classification [77])—that is, the task of identifying the sentiment expressed in images—[17] may be one example where Arabic-specific design is not necessary. It can be argued that existing solutions can be applied to Arabic by simply translating the identified sentiment such as “positive,” “negative,” or “neutral.” Since there are a limited number of classes, it should be straightforward to simply provide the Arabic translations for each class. However, in one paper, the authors argued that emotional facial expressions may be different depending on the cultural context [62]. The focus on specific cultural features present in images was also researched by other scholars. In one paper, the authors collected images of Chinese dishes in order to help develop a recognition system for Chinese food [20]. In another, the authors focused on premodern Japanese paintings and created a dataset of faces present in these works of art [68]. The dataset includes classes of faces that are unique to these types of images, such as faces of “warriors,” “commoners,” and “nobles.” Therefore, when addressing any challenges at the intersection of language and computer vision research, it is important to first determine if existing methods developed in English can be applied to another language with minimal modifications.
3.4 Arabic natural language processing

This work also relates to Arabic natural language processing, which has seen progress in recent years. Scholars have developed several solutions specifically tailored to Arabic, including work in named entity recognition [9, 31], sentiment analysis [4, 7], and question answering in Arabic [8, 55]. The interest in question answering in Arabic suggests that there might be a subsequent interest in methods that focus on visual question answering in Arabic. Research in Arabic natural language processing has also focused on the creation of new datasets that can be used to benchmark new methods or be incorporated in novel techniques. Examples of such datasets include a dataset of Arabic news articles labeled with a single label (called SANAD), a dataset of Arabic news articles labeled with multiple labels (called NADiaA) [26], a dataset of tweets from several cities classified with the Arabic dialect of the tweets [2], and a dataset of ironic Arabic tweets [1]. Another example is a dataset of 309 K sentences written in three different Arabic dialects as well as Modern Standard Arabic [29]. These papers indicate that constructing new datasets for Arabic may be necessary in order to investigate unexplored research areas. Additionally, the impacts of some of these new datasets suggest that they can be of great significance to the research community.

4 Methodology

In ImageNet, each synset has a name and an ID. To begin exploring the possibility of finding Arabic synsets and labels for images in ImageNet using AWN, all the synsets’ IDs are retrieved from ImageNet. There are several releases for ImageNet. In this paper, the Fall 2011 is used. This release includes 21,841 unique WordNet synsets, and each is linked to one or many images. For example, the synset ID “n07873807” includes 1296 images of “pizza.” There are also 1186 images for “dish,” the direct hyponym for “pizza.” Moreover, all the images labeled with “pizza” can also be labeled with “dish.” Since “dish” is a hyponym for several other synsets used in ImageNet, such as “sushi” and “curry,” the number of images for “dish” extends to all images with a synset that is a hyponym (child) for “dish.”

ImageNet data are downloaded directly from the ImageNet website.1 ImageNet provides the URLs of the images; however, many of the URLs are no longer active and therefore inaccessible [10]. The valid URLs are viewed in order to access the images. These images were viewed only to learn more about the dataset and collect images needed for the evaluation process. The results for synsets can be applied to the entire dataset regardless of whether they are currently available online. Due to ongoing developments related to the removal of problematic images present in the “person” subtree, ImageNet no longer provides a method to download the full dataset directly [73]. The dataset has a total of 14,197,122 images. Each synset has an average of 944 images directly assigned to the synset. The minimum number of images is 1 image, and the maximum is 2382 images for a synset.

1 http://www.image-net.org/download.php
4.1 Direct Arabic synsets for synsets in ImageNet

The first step is to investigate if Arabic synsets are available for each of the 21,841 synsets used in ImageNet. To complete this, all the synsets’ IDs are processed. For each synset,AWN is searched to find a direct match. There are several versions of AWN and several methods to access it currently exist. To gain additional knowledge on WordNet and AWN, and to determine a reliable method to access it, the online interfaces for both the Open Multilingual WordNet (OMW)\(^2\) [16] and the Princeton WordNet\(^3\) [50] are tested. Additionally, the WordNet interface in the python library NLTK\(^4\) [13] is tested. The interface includes the OMW, which has AWN [3, 14]. This version of AWN has 9916 Arabic synsets, which is less than the number of synsets used in ImageNet. This is the first indicator that it may not be possible to find direct matches for all synsets. Still, it is not clear if it is possible for a synset in ImageNet to be directly linked to several synsets in AWN. Based on this experimental phase, the NLTK interface is selected to access Arabic synsets in AWN. ImageNet uses WordNet 3.0, which has synsets IDs that are different than the ones used in WordNet 3.1. Therefore, the results in this paper are necessarily achieved by using WordNet 3.0.

4.2 Arabic synsets for hypernyms

Since ImageNet structures synsets based on the semantic structure of synsets in WordNet, a synset in ImageNet is essentially a node that is connected to a branch or several branches of hypernyms. Table 2 shows an example of how the AWN synset of a hypernym can be used to label images of a synset. The objective of this step is to discover if an Arabic synset in AWN is available for the list of hypernyms linked to a synset. To accomplish this, the parent-child (or hypernym-hyponym) pairs in ImageNet are downloaded from a webpage in ImageNet’s website.\(^5\) Algorithm 1 includes details of the steps followed in order to find direct matches as well as Arabic hypernyms for synsets.

---

**Algorithm 1:** Finding direct AWN synsets for ImageNet synsets as well as AWN synsets for all the hypernyms linked to ImageNet synsets.

| Input: | List of all synsets in ImageNet. “Synset” below refers to the synset from ImageNet. |
|---|---|
| Output: | 1) Direct_AWN: list of Arabic synsets from AWN that are directly linked to a synset in ImageNet, and 2) Hyper_AWN: a list of Arabic synsets from AWN that are linked to a hypernym of a synset in ImageNet. |

1: For synset in ImageNet:
2: If find_in_AWN (synset) is True then
3: Direct_AWN.add (synset, AWN_synset)
4: Find_Hyprs (synset, synset, 1)
5: func Find_Hyprs (synset, hyper_synset, level):
6: Hyper+=get_hyprs (hyper_synset)
7: If length (Hyperms) == 0:
8: Stop #no more hypernyms to process
9: For hyper_synset in Hyperms:
10: If find_in_AWN (hyper_synset) is True then
11: Hyper_AWN.add (synset, hyper_AWN_synset, level)
12: Find_Hyprs (synset, hyper_synset, level+1)

---

The algorithm relies on the use of a recursive function that looks for an Arabic synset for all the hypernyms connected to the synset at all levels. The stopping condition for this recursive

---

\(^2\) http://compling.hss.ntu.edu.sg/omw/

\(^3\) http://wordnet-rdf.princeton.edu/

\(^4\) http://www.nltk.org/howto/wordnet.html

\(^5\) http://www.image-net.org/archive/wordnet.is_a.txt
function is when all possible hypernyms are processed. To complete this process, all hypernyms of a synset are retrieved. In most cases, a synset that is a leaf node has one or two direct hypernyms. Each hypernym is searched for in AWN in order to discover if an Arabic synset for the hypernym exists in AWN. If one is found, it is added to the set of Arabic synsets for the primary synset. The level of the hypernym is also saved. For example, since an Arabic synset is available for the synset “dish,” which happens to be a hypernym for “pizza,” the Arabic synset for “dish” is saved for the “pizza” synset. Additionally, the number “one” is saved because “dish” is a direct hypernym for “pizza.” Similarity, the number “two” is saved for “nutriment,” which is the hypernym for “dish.” If a synset has two direct hypernyms, they are both saved as appearing in level one. Hypernyms are only saved when an Arabic synset is found.

This step results in a dataset of all AWN synsets in ImageNet, as well as the Arabic synsets available at each level of their hypernyms. Since it is possible for a synset to have two hypernyms, the objective of searching hypernyms is to indicate if any Arabic synsets exist for any of the hypernyms. After running the algorithm, all the synsets are processed again to label them with the first discovered synset (whether it was found directly or using hypernyms). However, released datasets of this paper will have all the synsets found in case some researchers want to develop certain methods that utilize all. Finally, it is unclear if using hypernyms to label images in Arabic will produce images with acceptable and meaningful labels.

Because the production of adequate Arabic labels from hypernyms is uncertain, an additional step is added to evaluate the results. Such an evaluation is completed by studying a sample of the Arabic synsets found in each level of hypernyms. For example, at level one (direct hypernyms), all the found Arabic synsets are processed. A representative random sample of synsets is then selected from all the synsets at level one that have Arabic synsets. The sizes for these samples are determined using standard methods to select representative sample sizes. The samples selected for each level include only synsets in which Arabic synsets are not found at any previous level. For example, synsets considered for the sample at the second level include only Arabic synsets that are not found directly or by using the direct hypernym (first level). Then, for each synset in each sample, all URLs of images linked to the synset are accessed until an image that is still available online is found. Since images collected from the website Flickr are more likely to remain available online [10], only images collected from Flickr are processed. If no online images from Flickr are found for the synset, all the other images collected from other websites are accessed. Discovered images are then added to an online spreadsheet. Each image is evaluated manually by the author to determine if the found Arabic synset accurately describes the content of the image. For each image in the sheet, the Arabic synset generated is classified as either “accurate” or “inaccurate.” The original English synsets are not viewed during the evaluation. Since some of the Arabic synsets consisted of uncommon words, an Arabic dictionary is used in order to determine the definitions of the words. Finally, the standard precision measure is used to evaluate the results for each sample.
4.3 The “person” subtree

The “person” subtree is one of the major categories in ImageNet, representing 8.3% of the images in the total dataset [73]. The tree includes 2832 unique synsets. Some of these synsets are located higher in the tree, while others represent leaf nodes with no hyponyms. Examples of synsets in this tree include ones related to occupations such as “doctor” and “entertainer,” nationalities such as “Senegali” and “Syrian,” and social ties such as “grandfather” and “friend.” This subtree includes several synsets that have been criticized for representational issues, including biases and offensive images [46, 61]. Recently, some of the scientists behind ImageNet addressed concerns regarding issues in the “person” subtree and indicated that an upgrade of ImageNet will be released with two changes: 1) only up to 158 of the 2832 synsets in the “person” subtree will be kept, and 2) attention will be given to representation biases in images in the synsets that remain [73]. In anticipation of these updates, the results section provides the results obtained when the 2832 synsets in the “person” subtree are not included. By omitting synsets in the “person” subtree, the total number of unique synsets decreases from 21,841 to 19,009.

4.4 The 1000 categories subset

The ImageNet Large Scale Visual Recognition Challenge (ILSVRC) is a popular competition where scientists in the computer vision research community compete to solve several research challenges [56]. For many of these challenges, a subset of 1000 categories or synsets from ImageNet are used instead of the entire dataset. According to Yang et al. (2020), this subset, rather than the entire dataset, is what most researchers use even outside of the ILSVRC. For this reason, the present study processed the synsets in this subset of 1000 with the goal of providing additional details about the Arabic synsets found both directly and using hypernyms. The list of synsets in this subset is retrieved from a page in the 2017 edition of ILSVRC. All but one synset have images. The synset with the WordNet ID “n04399382” has no available images in either the original ImageNet dataset or online. This subset has 1,290,129 images. Each synset in the subset has an average of 1291 images. The maximum number of images for a synset is 3047 and the minimum is 272 images.

5 Results

This section explains the extension of ImageNet to Arabic using AWN. Three classes of results are presented based on the dataset used: 1) results obtained when the entire ImageNet dataset is processed (Sections 5.1., 5.2., and 5.3.), 2) when synsets in the “person” subtree are omitted (Sections 5.4.), and 3) when the 1000 categories subset is used (Sections 5.5). For each of these three classes, the results for Arabic synsets found directly in AWN, as well as ones discovered using hyponym-hypernym relationships, are provided.

6 http://image-net.org/challenges/LSVRC/2017/browse-synsets
7 http://imagenet.stanford.edu/api/text/imagenet.synset.geturls?wnid=n04399382
5.1 Direct Arabic synsets

Direct matches were found for 1219 of the 21,841 synsets used in ImageNet. Since each synset is linked to many images, the dataset of 1219 synsets was extended to 1,150,651 images, which is 8.1% of ImageNet’s total number of images. This dataset represents a major contribution of the paper, as it can be used in several tasks related to Arabic computer vision. Table 3 includes ImageNet synsets, AWN synsets, and examples of images of mammals for synsets where an Arabic synset was found in AWN. These mammalian examples illuminate how several of the AWN synsets pertain to higher-level categories while others are more fine-grained. In this case, “carnivore” represents a more general category, while “vaulting horse” is more specific. Some of the synsets in the table are related. For example, “cheetah” is a hyponym of “carnivore” and “mammal” because the branch of hypernyms for “cheetah” is cheetah - > big cat - > feline - > carnivore - > placental - > mammal.

The quality of these labels is likely high, because all of the labels pertain to direct matches. Additional information on these results is provided by analyzing the textual structure of the synsets from ImageNet with a corresponding AWN synset. Three classes were created based on the number of words in each synset: “unigrams,” “bigrams,” and “Ngrams.” Unigrams refer to ImageNet synsets that include only one word, bigrams refer to synsets that include only two words, and Ngrams refer to synsets that include three or more words. Results indicate that most of the Arabic synsets found were for unigrams, while the percentage for Ngrams was the lowest. For the 1219 Arabic synsets found, 81.8% were of unigrams even though only 56% of the synsets in ImageNet are unigrams. Table 4 includes summary statistics of the number of Arabic synsets found for each class.

5.2 Arabic synsets for hypernyms

To further expand the dataset, hypernyms of synsets used in ImageNet were searched for in AWN. Consequently, Arabic synsets were identified for all but four synsets used in ImageNet. These four synsets include only 1366 images. This indicates that there are only 1366 images in ImageNet without Arabic synsets in AWN for the synset or one of its hypernyms in its branch of hypernyms. Table 5 includes samples from the first levels of synsets and their images. The “hypernym found” column lists all the hypernyms for each synset. The first Arabic hypernym found in AWN is in bold. For example, “seat” is the first AWN synset found for “chesterfield.”

| Images | ImageNet Synset ID | ImageNet Synset | Direct AWN Synset |
|--------|--------------------|----------------|------------------|
| ![Image](image1.png) | p02130308 | Cheetah (n.01) | فهد |
| ![Image](image2.png) | p01861778 | Mammal (n.01) | حيوان ثنائي القذأب |
| ![Image](image3.png) | p02062017 | Aquatic mammal (n.01) | حيوانات مائية |
| ![Image](image4.png) | p04524142 | Vaulting horse (n.01) | جصان وآب |
| ![Image](image5.png) | p02075296 | Carnivore (n.01) | لاجم، أكل اللحم |

Table 3 Examples of images from ImageNet with their synsets and Arabic synsets as found in AWN
Table 4 The number of synsets found for all, unigrams, bigrams, and Ngrams. The percentages represent the synsets found compared to total available.

| Images      | ImageNet Synsets | AWN Synsets Found |
|-------------|------------------|-------------------|
| All         | 21,841           | 1219 (5.59%)      |
| Unigrams    | 12,040           | 998 (8.28%)       |
| Bigrams     | 8940             | 207 (2.31%)       |
| Ngrams      | 816              | 14 (1.62%)        |

and thus all images that belong to this synsets were labeled with the Arabic synset for “seat.” The images in the table are a sample of the images labeled with the synset in ImageNet.

A detailed summary of the results is presented in Table 6. In the table, “AWN synsets” refers to the number of Arabic synsets found for a synset in ImageNet at each level regardless of what was found in previous levels. The “AWN synset + previous” refers to the total number of Arabic synsets that were identified when the synsets found at a level and the previous levels are combined. For example, for first level hypernyms (Row # and Column #3 in Table 6), the number found was 10,462, which represents synsets found directly (1219) and synsets found using the first level hypernyms that were not available directly. Put differently, synsets found directly and using first level hypernyms are not double counted. The “Images in ImageNet” refers to the total number of images found for each Arabic synset at each level. When only the first and second level hypernyms were considered, the dataset included Arabic synsets for 79.8% of the synsets and 81.2% of the images in ImageNet. This represents a large dataset of 11,533,525 images, all of which are labeled with an Arabic synset that is either the direct match for the synset used in ImageNet, the Arabic synset for the hypernym, or the Arabic synset for the hypernym of a hypernym. Although a synset in this subset (Row #4 in Table 6) was found for 17,438 of the synsets used in ImageNet, many of the identified Arabic synsets were used more than once since the total number of synsets in AWN is only 9916. For example, the Arabic synset for “candy (n_01)” was the first hypernym found for 118 different synsets. Similarly, the Arabic synset for “edible fruit (n_01)” was used as the Arabic label for all the images that are linked to 142 synsets. Finally, it is important to note that as the level of the hypernym increases, the hypernyms become more abstract and general. For example, some of the 7th-level hypernyms include “entity,” “act,” and “event.” Therefore, the usability of Arabic synsets at higher levels requires additional investigation.

Table 5 Images from ImageNet and their synsets with the first Arabic synset for a hypernym as found in AWN.

| Images       | ImageNet Synset | Level | Hypernym Found            | AWN Synset |
|--------------|-----------------|-------|---------------------------|------------|
| Electric fan | Electric fan> Fan (n_04) | 1     | مروحة                     |            |
| Travel iron  | Travel iron> Iron> home appliance (n_01) | 2     | جهاز مزدّن |            |
| Yagi         | Yagi> directional antenna> antenna> electrical device (n_01) | 3     | أداة كهربائية, جهاز كهربائي |            |
| Chesterfield | Chesterfield> davenport> convertible> Sofa> Seat (n_02) | 4     | كرسي, مفتوح |            |
| Blue shark   | Blue shark> requiem shark> shark> elasmobranch> Fish (n_01) | 5     | سمكة |            |
5.3 Evaluation for hypernyms

While this dataset of Arabic synsets for images in ImageNet is likely reliable since it is based on the utilization of previously existing and evaluated datasets, certain characteristics of the Arabic language and naming decisions in AWN suggest the potential need for proper evaluation of the dataset’s accuracy. For example, one of the synsets found in AWN was “nurse.” Unlike English, Arabic nouns are gendered. Accordingly, in AWN the Arabic synset for “nurse” is “male nurse.” Therefore, an automated image classification system that relies on this synset may suggest “male nurse” for images of female nurses. Additionally, labelling images using the Arabic synsets for hypernyms may not produce satisfactory or meaningful results. Therefore, further evaluation, as fully explained in Section 4.2., is needed in order to provide quantitative measurements for the accuracy of labels found at each level.

For each of the seven levels, the values for the precision measure were 90.4% or above. These findings suggest that generating Arabic labels for images in ImageNet using hypernyms does indeed produce meaningful results. Table 7 includes the results as well as the number of true positive and false positive instances. For sample sizes with at least 300 synsets, the highest results occurred for hypernyms at the fifth level. This evaluation was completed using binary classes. For each image, the classes available were either “true (correct)” or “false (incorrect).” Therefore, the evaluation did not measure the quality of the generated labels. It was observed that some images in the true positive column had high-quality labels. Others, however, had acceptable labels that were, nevertheless, of lesser quality. This is one limitation of the evaluation.

Several observations are drawn from this evaluation process. First, many of the images that were classified as inaccurate were of synsets in the “person” subtree. Therefore, it is possible that the values for the precision measure are higher if synsets in the “person” subtree are

Table 6 Summary of the number of synsets and number of images found

| Level               | AWN Synsets | AWN Synsets+Previous | Images in ImageNet | Images+Previous |
|--------------------|-------------|----------------------|--------------------|-----------------|
| Direct             | 1219 (5.59%)| –                    | 1,150,651 (8.1%)   | –               |
| First level Hypernyms | 10,400 (47.6%) | 10,462 (47.9%)   | 7,113,853 (50.1%) | 7,177,537 (50.5%) |
| 2nd level Hypernyms  | 16,837 (77.0%) | 17,438 (79.8%)   | 11,088,519 (78.1%) | 11,533,525 (81.2%) |
| 3rd level Hypernyms  | 19,490 (89.2%) | 20,267 (93.7%)   | 12,697,838 (89.4%) | 13,266,956 (93.4%) |
| 4th level Hypernyms  | 20,671 (94.6%) | 21,397 (97.9%)   | 13,365,182 (94.1%) | 13,916,531 (98.0%) |
| 5th level Hypernyms  | 20,816 (95.3%) | 21,751 (99.5%)   | 13,483,006 (94.9%) | 14,148,669 (99.6%) |
| 6th level Hypernyms  | 19,910 (91.1%) | 21,830 (99.94%)  | 12,865,510 (90.6%) | 14,195,325 (99.8%) |
| 7th level Hypernyms  | 18,191 (83.2%) | 21,837 (99.98%)  | 11,729,718 (82.6%) | 14,195,756 (99.9%) |

Table 7 Summary of the evaluation for Arabic synsets found using hypernyms

| Level      | Sample size | True positive | False positive | Precision   |
|------------|-------------|---------------|----------------|-------------|
| First level| 378         | 348           | 30             | 92.06%      |
| 2nd level  | 372         | 347           | 25             | 93.28%      |
| 3rd level  | 354         | 320           | 34             | 90.4%       |
| 4th level  | 357         | 327           | 30             | 91.6%       |
| 5th level  | 323         | 316           | 7              | 97.83%      |
| 6th level  | 69          | 66            | 3              | 95.65%      |
| 7th level  | 6           | 6             | 0              | 100%        |
| All        | 1730        | 1859          | 129            | 93.06%      |
omitted. Second, many of the Arabic synsets found are used in several levels. For example, the Arabic synsets for “electrical_devices_n_01” are used for multiple synsets at levels 1, 2, 3, 4, and 5. Third, some synsets have two different Arabic hypernyms in the same level. For example, no direct AWN synset is available for “rock climbing.” However, the Arabic synsets for the two direct hypernyms, “sport” and “climb,” are available. All of these observations present some of the opportunities for future research.

5.4 Results after omitting the “person” subtree

This section includes the results obtained when synsets in the “person” subtree were not included. The “person” subtree has 2832 synsets. After their removal, the number of synsets in the dataset was reduced from 21,841 to 19,009 and the number of images was reduced from 14,197,122 to 13,019,837 images. The IDs for synsets in the tree were identified by following the branch of hypernyms for each synset. Doing so allowed for the classification of synsets where the synset “person.n.01” is a hypernym in any level. The complete results are presented in Table 8. When the percentages of Arabic synsets found in the two sets (ImageNet with and without the “person” subtree) were compared, the results were slightly lower for the latter. Moreover, the percentages of synsets found in each level only decreased marginally. New and previously unreported issues pertaining to images in the “person” subtree were discovered. For instance, it was observed that images in the “Iraqi” synset were mostly war related. This could be problematic for several applications that utilize this synset. For example, predictive models that classify images based on learning from synsets in the “person” subtree may classify images that pertain to war as ones of Iraqis. Additional issues were discovered for images in the “Syrian” synset. For instance, it can be argued that some individuals presented in this synset may not actually be Syrian.

5.5 Results for the 1000 categories subset

Arabic synsets were found for only 77 of the 1000 synsets. When the first and second level hypernyms were used, the number of synsets found were 523 and 801 synsets, respectively. These two sets include 681,588 images when first-level hypernyms were used and 1,080,157 images when second-level hypernyms were used. A label was found for all the images in the 1000 categories when all the hypernyms were processed. The hypernyms in the first six levels were examined in order to find at least one Arabic hypernym. Figure 1 displays the subsets and

Table 8 Summary of the number of synsets and number of images found when synsets in the “person” subtree are omitted

| Level            | AWN Synsets | AWN Synsets+ Previous | Images in ImageNet | Images+Previous |
|------------------|-------------|-----------------------|--------------------|-----------------|
| Direct           | 993 (5.2%)  | –                     | 990,189 (7.6%)     | –               |
| First level Hypnyms | 8846 (46.5%) | 8890 (46.7%)          | 6,419,390 (49.3%)  | 6,469,802 (49.6%) |
| 2nd level Hypnyms | 14,541 (76.4%) | 15,064 (79.2%)       | 10,099,744 (77.5%) | 10,508,392 (80.7%) |
| 3rd level Hypnyms | 16,870 (88.7%) | 17,571 (92.4%)       | 11,588,355 (89%)   | 12,128,876 (93.1%) |
| 4th level Hypnyms | 17,900 (94.1%) | 18,589 (97%)         | 12,204,832 (93.7%) | 12,746,289 (97.8%) |
| 5th level Hypnyms | 17,990 (94.6%) | 18,924 (99.5%)       | 12,307,766 (94.5%) | 12,973,354 (99.6%) |
| 6th level Hypnyms | 17,254 (90.7%) | 18,998 (99.94%)      | 11,766,924 (90.3%) | 13,018,040 (99.98%) |
| 7th level Hypnyms | 15,828 (83.2%) | 19,005 (99.97%)      | 10,739,341 (82.4%) | 13,018,471 (99.989%) |
result summaries for 1) all of ImageNet, 2) ImageNet without the “person” subtree and, 3) the 1000 categories set. The percentage associated with each bar represents the percentage of coverage when the results are compared to the total number of images in each subset.

6 Discussion

The results demonstrate the ability of the approach proposed in this paper to generate datasets of images labeled in Arabic using AWN both directly and using hypernyms. For the latter, our evaluation suggests that the majority of labels discovered were accurate. One important question to ask is regarding the reasons why the method and evaluation were successful. Since this work relies on Arabic WordNet, one main explanation is the availability and quality of synsets in AWN. Put differently, if these labels were not available in AWN, our method would not have been able to find them. A second reason for the results found is the fact that ImageNet only uses nouns. It is possible that the matches discovered would not be this high if verbs were also used. Third, due to 1) the hierarchical nature of labels in ImageNet and 2) how our method to discover Arabic labels relies on hypernyms, many of the same Arabic synsets were the ones discovered for several different synsets in ImageNet.

To improve the results found, several options are available. One option is to use the Extended Open Multilingual Wordnet (1.2), which enhances Arabic WordNet by utilizing Wiktionary and Unicode Common Locale Data Repository [15]. This automatic extension of AWN increases the total number of unique synsets available in AWN to 14,650 synsets. By using this improved version of AWN, the number of Arabic synsets found for ones used in ImageNet may increase. However, since this extension relied on automated methods, it is possible that the quality of the results will decrease. Another method to increase the number of Arabic synsets found is to propose new techniques that will allow for a direct discovery of
Arabic synsets using the information available about the synsets used in ImageNet such as their full definitions as available in WordNet. Such automated techniques could leverage information available in ImageNet and WordNet in order to find Arabic words for the synsets used when that particular synset is not available in AWN.

Several directions for future work exist. One extension is to use this dataset to create new datasets that may be useful for a variety of research challenges related to Arabic computer vision. For instance, the AWN labels that were found can be used to create a dataset of Arabic captions for images. Another avenue for further research would involve applying the same methods employed in this paper to extend ImageNet to other languages using the WordNets for these languages, such as the French WordNet [59] and the Danish WordNet [49].

7 Conclusion

This paper has explored the possibility of extending ImageNet to Arabic. The following discoveries were made: 1) an Arabic synset in AWN exists for 1219 of the synsets used in ImageNet, which represents 1,150,651 images, and 2) Arabic synsets in AWN exist for 99.9% of the images in ImageNet when the branches of hypernyms for synsets were considered. Two subsets of ImageNet were also processed and showed similar results. This work presents several implications. First, this large dataset of images labeled with Arabic could be used to enhance a variety of information systems related to Arabic computer vision. For example, an application can utilize these images to teach Arabic words to children who suffer from learning disabilities by showing them images from the dataset. Second, the dataset can be incorporated in predictive AI models that aim to classify images in Arabic. Finally, the dataset can be used to tackle several challenging research questions related to Arabic computer vision.

Acknowledgements

This project was supported by the Deanship of Scientific Research at Prince Sattam bin Abdulaziz University, Al-Kharj, Saudi Arabia.

References

1. Abbes I, Zaghouani W, El-Hardlo O, Ashour F (2020) DAICT: A Dialectal Arabic Irony Corpus Extracted from Twitter. In: Proc. 12th Lang. Resour. Eval. Conf. European Language Resources Association, Marseille, France, pp 6265–6271. https://www.aclweb.org/anthology/2020.lrec-1.768
2. Abdul-Mageed M, Alhuwali H, Elaraby M You Tweet What You Speak: A City-Level Dataset of Arabic Dialects. In: Proc. Elev. Int. Conf. Lang. Resour. Eval. ({LREC} 2018), vol 2018. European language resources association (ELRA), Miyazaki, Japan. https://www.aclweb.org/anthology/L18-1577
3. Abouenour L, Bouzoubaa K, Rosso P (2013) On the evaluation and improvement of Arabic WordNet coverage and usability. Lang Resour Eval 47:891–917. https://doi.org/10.1007/s10579-013-9237-0
4. Al-Ayyoub M, Khamaiseh AA, Jararweh Y, Al-Kabi MN (2019) A comprehensive survey of arabic sentiment analysis. Inf Process Manag 56:320–342. https://doi.org/10.1016/j.ipm.2018.07.006
5. Alkhalifa M, Rodriguez H (2009) Automatically extending NE coverage of Arabic WordNet using Wikipedia. In: 3rd Int. Conf. Arab. Lang. Process. CITALA2009, Rabat, Morocco, pp 20–36
6. Al-muzaini HA, Al-yahya TN, Benhidour H (2018) Automatic Arabic image captioning using RNN-LSTM-based language model and CNN. Int J Adv Comput Sci Appl 9:67–73
7. Al-smadi M, Al-ayyoub M, Jararweh Y, Qawasmeh O (2019) Enhancing aspect-based sentiment analysis of Arabic hotels’ reviews using morphological, syntactic and semantic features. Inf Process Manag 56:308–319. https://doi.org/10.1016/j.ipm.2018.01.006
8. Al-smadi M, Al-dalabih I, Jararweh Y, Juola P (2019) Leveraging linked open data to automatically answer Arabic questions. IEEE Access. 7:177122–177136. https://doi.org/10.1109/ACCESS.2019.2956233
9. Al-smadi M, Al-zboon S, Jararweh Y, Juola P (2020) Transfer learning for Arabic named entity recognition with deep neural networks. IEEE Access. 8:37736–37745. https://doi.org/10.1109/ACCESS.2020.2973319
10. Alsudais A (2019) Image classification in Arabic: exploring direct English to Arabic translations. IEEE Access 7:122730–122739. https://doi.org/10.1109/ACCESS.2019.2926924
11. Alsudais A (2020) Extending ImageNet to Arabic using Arabic WordNet. In: First Work. Adv. Lang. Vis. Res. The Association for Computational Linguistics
12. Batia MA, Ayadi R, Zrigui M (2019) Reasoning over Arabic WordNet Relations with Neural Tensor Network. Comput. Y Sist 23:935–942. https://doi.org/10.13053/CyS-23-3-3240
13. Bird S, Klein E, Loper E (2009) Natural language processing with Python: analyzing text with the natural language toolkit. O’Reilly Media, Inc.
14. Black W, Elkateb S, Rodriguez H, Alkalifa M, Vossen P, Pease A, Fellbaum C (2006) Introducing the Arabic WordNet Project. In: Proc. Third Int. WordNet Conf, pp 295–299
15. Bond F, Foster R (2013) Linking and Extending an Open Multilingual Wordnet. In: Proc. 51st Annu. Meet. Assoc. Comput. Linguist. Association for Computational Linguistics, Sofia, Bulgaria, pp 1352–1362
16. Bond F, Paik K (2012) A Survey of WordNets and their Licenses. In: Proc. 6th Glob. WordNet Conf. (GWC 2012), pp 64–71
17. Campos V, Jou B, Giró-i-nieto X (2017) From pixels to sentiment: fine-tuning CNNs for visual sentiment prediction. Image Vis Comput 65:15–22. https://doi.org/10.1016/j.imavis.2017.01.011
18. Cavaliere D, Loia V, Saggese A, Senatore S, Vento M (2020) A human-like description of scene events for a proper UAV-based video content analysis. Knowledge-Based Syst 178:163–175. https://doi.org/10.1016/j.knosys.2019.04.026
19. Cetinic E, Lipic T, Grbic S (2018) Fine-tuning convolutional neural networks for fine art classification. Expert Syst Appl 114:107–118. https://doi.org/10.1016/j.eswa.2018.07.026
20. X. Chen, Y. Zhu, H. Zhou, L. Diao, D. Wang. ChineseFoodNet: A Large-scale Image Dataset for Chinese Food Recognition, arXiv. (2017).
21. Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JMF, Parikh D, Batra D (2017) Visual Dialog. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp 326–335
22. Deng J, Dong W, Socher R, Li L, Li K, Fei-fei L (2009) ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conf. Comput. Vis. Pattern Recognit, Miami, FL, pp 248–255. https://doi.org/10.1109/CVPR.2009.5206848
23. Ding S, Qu S, Xi Y, Kumar A, Wan S (2019) Image caption generation with high-level image features. Pattern Recogn Lett 123:89–95. https://doi.org/10.1016/j.patrec.2019.03.021
24. Dost S, Serafini L, Rospocher M, Ballan L, Sperduti A (2020) VTKEL: A resource for Visual-Textual-Knowledge Entity Linking. In: 35th ACM/SIGAPP Symp. Appl. Comput. (SAC 2020), March 30–April 3.
25. Elliot D, Frank S, Sima’an K, Specia L (2016) Multi30K: Multilingual English-German Image Descriptions. In: Proc. 5th Work. Vis. Lang, pp 70–74
26. Elnagar A, Al-debsi R, Einea O (2020) Arabic text classification using deep learning models. Inf Process Manag 57:102121. https://doi.org/10.1016/j.ipm.2019.102121
27. Elrefaei LA, Alhassan TQ, Omar SS (2019) An Arabic visual dataset for visual speech recognition. Procedia Comput Sci 163:400–409. https://doi.org/10.1016/j.procs.2019.12.122
28. Fang Z, Liu J, Li Y, Qiao Y, Lu H (2019) Improving visual question answering using dropout and enhanced question encoder. Pattern Recogn 90:404–414. https://doi.org/10.1016/j.patcog.2019.01.038
29. Farhan W, Talafha B, Abuammar A, Jaikat R, Al-ayyoub M, Tarakji AB, Toma A, Samsung R (2020) Unsupervised dialectal neural machine translation. Inf Process Manag 57:102181. https://doi.org/10.1016/j.ipm.2019.102181
30. Goikoetxea J, Soroa A, Agirre E (2018) Bilingual embeddings with random walks over multilingual wordnets. Knowledge-Based Syst. 150:218–230. https://doi.org/10.1016/j.knosys.2018.03.017
31. Helwe C, Elbassuoni S (2019) Arabic named entity recognition via deep co-learning. Artif Intell Rev 52:197–215. https://doi.org/10.1007/s10462-019-09688-6
32. Huang L, Luo B (2017) Tag refinement of micro-videos by learning from multiple data sources. Multimed Tools Appl 76:20341–20358. https://doi.org/10.1007/s11042-017-4781-z
33. Huang F, Zhang X, Li Z, Zhao Z, He Y (2018) From content to links: social image embedding with deep multimodal. Knowledge-Based Syst. 160:251–264. https://doi.org/10.1016/j.knosys.2018.07.020
34. Huang F, Zhang X, Zhao Z, Xu J, Li Z (2019) Image–text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Syst. 167:26–37. https://doi.org/10.1016/j.knosys.2019.01.019
35. Ibn M, Jmilah H, El-yacoubi MA (2020) Automatic processing of historical Arabic documents: a comprehensive survey histograms of oriented gradients part of Arabic word. Pattern Recogn 100:107144. https://doi.org/10.1016/j.patcog.2019.107144

36. Jindal V (2018) Generating image captions in Arabic using root-word based recurrent neural networks and deep neural networks. In: Proc. NAACL-HLT 2018 Student Res. Work, pp 144–151

37. Kastner MA, Ida I, Kawanishi Y, Hirayama T, Deguchi D, Murase H (2019) Estimating the visual variety of concepts by referring to web popularity. Multimed Tools Appl 78:9463–9488. https://doi.org/10.1007/s11042-018-6528-x

38. Kornblith S, Shlens J, Le QV (2019) Do Better ImageNet Models Transfer Better? In: IEEE Conf. Comput. Vis. Pattern Recognit, vol 2019, pp 2661–2671

39. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Adv. Neural Inf. Process. Syst, pp 1097–1105

40. Kumar A, Srinivasan K, Wen-huang C, Zomaya AY (2020) Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. Inf Process Manag 57. https://doi.org/10.1016/j.ipm.2019.102141

41. Li Z, Tang J (2017) Weakly supervised deep matrix factorization for social image understanding. IEEE Trans Image Process 26:276–288

42. Li Z, Tang J, Member S, Mei T, Member S (2019) Deep collaborative embedding for social image understanding. IEEE Trans Pattern Anal Mach Intell 41:2070–2083

43. Li Q, Li L, Wang W, Li Q, Zhong J (2020) A comprehensive exploration of semantic relation extraction via pre-trained CNNs. Knowledge-Based Syst 194. https://doi.org/10.1016/j.knosys.2020.105488

44. Li P, Mao K, Xu Y, Li Q, Zhang J (2020) Bag-of-concepts representation on automatic knowledge acquisition from probabilistic knowledge base. 193. https://doi.org/10.1016/j.knosys.2019.105436

45. Liu M, Li L, Hu H, Guan W, Tian J (2020) Image caption generation with dual attention mechanism. Inf Process Manag 57:102178. https://doi.org/10.1016/j.ipm.2019.102178

46. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A Survey on Bias and Fairness in Machine Learning, arXiv:1908.09635v2. (2019).

47. Miller GA (1995) WordNet: a lexical database for English. Commun ACM 38:39–41

48. Navigli R, Ponzetto SP (2012) BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artif Intell 193:217–250. https://doi.org/10.1016/j.artint.2012.07.001

49. Pedersen BS, Nimb S, Asmussen J, Sørensen NH, Trap-Jensen L, Lorentzen H (2009) DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. Lang Resour Eval 43:269–299. https://doi.org/10.1007/s10579-009-9092-1

50. Princeton University (2010) About WordNet. Princeton University, Princeton, NJ 08544, USA

51. Qaroush A, Jaber B, Mohammad K, Washaha M, Maali E, Nayef N (2019) An efficient, font independent word and character segmentation algorithm for printed Arabic text. J. King Saud Univ - Comput Inf Sci. https://doi.org/10.1016/j.jsuci.2019.08.013

52. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In: Int. Conf. Mach. Learn, New York, NY, USA

53. Regragui Y, Abouenour L, Krieche F, Bouzoubaa K, Rosso P (2016) Arabic WordNet: New content and new applications. Arabic WordNet. In: Proc. Eighth Glob. WordNet Conf, pp 330–338

54. Roller S, Schulte S (2013) A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities. In: Proc. 2013 Conf. Empir. Methods Nat. Lang. Lang, Process, Association for Computational Linguistics, Seattle,Washington, USA, pp 1146–1157

55. Romeo S, Da G, Martino S, Belinkov Y, Barrón-cedeño A, Eldesouki M, Darwish K, Mubarak H, Glass J, Moschitti A (2019) Language processing and learning models for community question answering in Arabic. Inf Process Manag 56:274–290. https://doi.org/10.1016/j.ipm.2017.07.003

56. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252. https://doi.org/10.1007/s11263-015-0816-y

57. Ruwa N, Mao Q, Wang L, Gou J, Dong M (2019) Mood-aware visual question answering. Neurocomputing. 330:305–316. https://doi.org/10.1016/j.neucom.2018.11.049

58. Saad RSM, Elanwar RI, Kader NSA, Mashali S, Betke M (2016) BCE-Arabic-v1 dataset: towards interpreting Arabic document images for people with visual impairments categories and subject descriptors. In: PETRA’16, Corfu Island, Greece

59. Sagot B, Fiser D (2008) Building a free French wordnet from multilingual resources. In: Ontolex 2008

60. Sanabria R, Caglayan O, Palaskar S, Elliott D, Barrault L, Specia L, Metze F (2018) How2: a large-scale dataset for multimodal language understanding. In: NeurIPS 2018 work. Vis. Grounded Interact. Lang
61. Shankar S, Halpern Y, Breck E, Atwood J, Wilson J, Sculley D (2017) No classification without representation: assessing geodiversity issues in open data sets for the developing world. In: NIPS 2017 work. Mach. Learn. Dev. World

62. Shaqra FA, Duwairi R, Al-ayyoub M (2019) The audio-visual Arabic dataset for natural emotions. In: 2019 7th Int. Conf. Futur. Internet Things Cloud, IEEE. https://doi.org/10.1109/FiCloud.2019.00054

63. Shi Z, Hao H, Zhao M, Feng Y, He L, Wang Y, Suzuki K (2019) A deep CNN based transfer learning method for false positive reduction. Multimed Tools Appl 78:1017–1033. https://doi.org/10.1007/s11042-018-6082-6

64. K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv: 1409.1556. (2014).

65. Stock P, Cisse M (2018) ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases. ECCV, pp 504–519

66. Szegedy C, Vanhoucke V, Shlens J (2016) Rethinking the inception architecture for computer vision, 2016 IEEE Conf. Comput Vis Pattern Recognit:2818–2826. https://doi.org/10.1109/CVPR.2016.308

67. Tan M, Zhou J, Peng Z, Yu J, Tang F (2020) Fine-grained image classification with factorized deep user click feature. Inf Process Manag 57:102186. https://doi.org/10.1016/j.ipm.2019.102186

68. Y. Tian, C. Suzuki, T. Clanuwat, M. Bober-irizar, A. Lamb, A. Kitamoto, KaoKore: A Pre-modern Japanese Art Facial Expression Dataset, arXiv. (2020).

69. Unar S, Wang X, Wang C, Wang Y (2019) A decisive content based image retrieval approach for feature fusion in visual and textual images. Knowledge-Based Syst. 179:8–20. https://doi.org/10.1016/j.knosys.2019.05.001

70. Vo N, Jiang L, Sun C, Murphy K, Li L-J, Fei-Fei L. (2019) J. Hays., Composing Text and Image for Image Retrieval - An Empirical Odyssey. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit, pp 6439–6448

71. Wang X, Wu J, Chen J, Li L, Wang Y, Wang WY VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In: 2019 IEEE/CVF Int. Conf Comput Vis, vol 2019, pp 4580–4590

72. Xu J, Huang F, Zhang X, Wang S, Li C, Li Z, He Y (2019) Knowledge-based systems visual-textual sentiment classification with bi-directional multi-level. Knowledge-Based Syst. 178:61–73. https://doi.org/10.1016/j.knosys.2019.04.018

73. Yang K, Qinami K, Fei-Fei L, Deng J, Russakovsky O (2020) Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In: Conf. Fairness, Accountability, Transpar. (FAT* ’20), ACM, Barcelona, Spain, pp 547–558. https://doi.org/10.1145/3351095.3375709

74. Ye Y, Zhang S, Li Y, Qian X, Tang S, Pu S (2020) Video question answering via grounded cross-attention network learning. Inf Process Manag 57:102265. https://doi.org/10.1016/j.ipm.2020.102265

75. Y. Yoshikawa, Y. Shigeto, A. Takeuchi, STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset, in: Proc. 55th Annu. Meet. Assoc. Comput. Linguist. (Short Pap., Association for Computational Linguistics, Vancouver, Canada, July 30–August 4, 2017., 2017: pp. 417–421.

76. Zhang J, Wang Q, Han Y (2020) Multi-modal fusion with multi-level attention for visual dialog. Inf Process Manag 57:102152. https://doi.org/10.1016/j.ipm.2019.102152

77. Zhang J, Chen M, Sun H, Li D, Wang Z (2020) Object semantics sentiment correlation analysis enhanced image sentiment classification. Knowl-Based Syst 191:105245. https://doi.org/10.1016/j.knosys.2019.105245

78. Zhou T, Li Z, Zhang C, Ma H (2020) Classify multi-label images via improved CNN model with adversarial network. Multimed Tools Appl 79:6871–6890. https://doi.org/10.1007/s11042-019-08568-z

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.