Integration of Computer Vision with Analogical Reasoning for Characterizing Unknowns

Kara Combs
Applied Research Solutions, USA
Kcombs@appliedres.com

Trevor J. Bihl
Air Force Research Laboratory, USA
Trevor.Bihl.2@us.af.mil

Subhashini Ganapathy
Wright State University, USA
Subhashini.Ganapathy@wright.edu

Abstract

Current state-of-the-art artificial intelligence struggles with accurate interpretation of out-of-library (OOL) objects. One method proposed remedy is analogical reasoning (AR), which utilizes abductive reasoning to draw inferences on an unfamiliar scenario given knowledge about a similar familiar scenario. Currently, applications of visual AR gravitate toward analogy-formatted image problems rather than to computer vision data sets. The Image Recognition Through Analogical Reasoning Algorithm (IRTARA) approach described herein shows how AR can be leveraged to improve computer vision in OOL situations. IRTARA produces a word-based term frequency list that characterizes the OOL object of interest. To evaluate the quality of the results of IRTARA, both quantitative and qualitative assessments are used, including a baseline to compare the automated methods with human-generated results. Fifteen OOL objects were tested using IRTARA, which showed consistent results across all three evaluation methods on the objects that performed exceptionally well or poorly overall.

1. Introduction

Culturally, artificial intelligence (AI) is associated with computers that can completely mimic human thought processes; however, the case in real-world AI applications is considerably brittle (McCarthy, 2004). The vast majority of today's AI is classified as "weak," meaning it is limited to the tasks and datasets that the AI was originally trained to do (IBM, 2021). When considering new data, i.e., unexpected queries from outside the original training scope, poor results are often the result. However, the ability of AI to evaluate these unknown instances is generally termed "out-of-library" (OOL), since they are outside the scope of the training data (Situ, Friend, Bauer, & Bihl, 2016). Necessarily, appropriately handling OOLs is a critical step in the direction of "strong" AI, which can provide generalization in perception and cognition (IBM, 2021).

AI is a broad domain and includes many applications and methods, including natural language processing and computer vision (CV) (McCarthy, 2004). Due to the rise in popularity of applications such as handwriting recognition, depth perception, and augmented reality, the ability to accurately identify and describe images is of great importance (Google, 2021). In this context, OOL handling would be considered images of objects the AI has not been previously trained on such as explored in zero-shot learning (Socher, Ganjoo, Manning, & Ng, 2013). One way to assist the transition to "strong" AI for CV is through integrating typical CV methods with other popular methods such as artificial neural networks.

The primary research question is how OOL objects can be understood by AI without overly computationally expensive deep learning methods and then, evaluate the success of these proposed methods. This paper addresses this question by proposing the Image Recognition Through Analogical Reasoning Algorithm (IRTARA) methodology which leverages the advantages of analogical reasoning in an imaged-based OOL scenario. Several metrics were further created to assess the quality of IRTARA's results based on automated methods as well as human-based judgment.

This paper aims to address the following research questions:
RQ1: What is the current state of research on CV, analogical reasoning, and their intersection?
RQ2: How can analogical reasoning be leveraged in an OOL-CV scenario via an automated, repeatable process?
RQ1 is addressed through a review of the background of CV methods and analogical reasoning algorithms and a discussion of their general capabilities (functional and algorithmic). Limited prior work exists in the image-based and image-to-text AR (Lu, Liu, Ichien, & Holyoak, 2019) (Sadeghi, Zitnick, & Farhadi, 2015) (Doumas & Hummel, 2010) (Reed, Zhang, Yuting, & Lee, 2015) (Hwang, Grauman, & Sha, 2013), with the

URI: https://hdl.handle.net/10125/102748
978-0-9981331-6-4

vast majority of literature focusing on text-based AR (Gentner, 1983) (Holyoak & Thagard, 1989) (Mikolov, Tomas, Yih, & Zweig, 2013) (Pennington, Socher, & Manning, 2014) (Bojanowski, Grave, Joulin, & Mikolov, 2017) (Hummel & Holyoak, 1997) (Wilson, Halford, Gray, & Phillips, 2001). The former research in AR applied to images has been limited to analogy-formatted data (e.g., *A is to B as C is to D*), thus this paper aims to address RQ2 through the proposed IRTARA methodology. Additionally, an analysis of IRTARA on classification image data with results and their interpretation through qualitative and quantitative measures is then presented and followed by conclusions.

2. Background

Presently, considerable AI utility has been seen in image data through convolutional neural networks (CNN), a form of an artificial neural network (see (LeCun, et al., 1989); (He, Zhang, Ren, & Sun, Deep residual learning for image recognition, 2016); (Liu, et al., 2018)). Even the most advanced deep CNN, unless integrated with another process, can only produce results that it was pre-trained on and aware of, i.e., not OOL objects. One method with proven success in extrapolating new information is analogical reasoning which has seen limited image-based applications.

Learning by analogies, as in analogical reasoning, is based on using information from the familiar “base” and extending this information onto an unfamiliar “target” (Gentner & Maravilla, 2018). The success of analogical reasoning in solving analogy problems has been proven in both the visual/pictorial (Polya, 1990; Zhang, Gao, Baoxiong, Zhu, & Song-Chun, 2019) and text/verbal space (French, 2002; Rogers, Drozd, & Li, 2017). Considerable emphasis has been on the development of analogical reasoning for text-based analogies with many algorithms developed to address the wide range of text-based analogy problems (Combs, Bihl, Ganapathy, & Staples, 2022). These text-based problems range from novel word problems (e.g., king:queen::man:woman) to mapping sentence elements (e.g., “She is growing like a weed”) to drawing parallels between stories (Ichien, Lu, & Holyoak, 2020). Initially, analogical reasoning started as psychologically-based algorithms (see (Gentner, 1983); (Holyoak & Thagard, 1989); (Hofstadter & Mitchell, 1995)) but recently, with the rise of natural language processing, vector space models and artificial neural network approaches have increased in popularity (Combs, Bihl, Ganapathy, & Staples, 2022). To date, the most prominent vector space models include Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Mikolov, Tomas, Yih, & Zweig, 2013), Global Vectors (GloVe) (Pennington, Socher, & (Hofstadter & Mitchell, 1995) (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)
Manning, 2014), and fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017). Within the artificial neural network scope, models include Learning and Inference with Schemas and Analogies (Hummel & Holyoak, 1997) and Structured Tensor Analogical Reasoning 2 (Wilson, Halford, Gray, & Phillips, 2001). A handful of these algorithms were selected for an apples-to-apples comparison which showed the advantages and disadvantages of each (Combs, Bihl, Ganapathy, & Staples, 2022).

The exploitation of these approaches in the image space has been limited to simply drawing visual analogies rather than applying analogical reasoning as a methodology. ANALOGY was arguably the first analogical reasoning algorithm that was designed to solve geometric analogy problems. The visual analogical reasoning space has largely been dominated by similar geometric-based problems such as Raven’s Progressive Matrices (see (Raven & Court, 1938)). Of the remaining visual analogy algorithms, only a handful applies analogical reasoning to a CV-like problem. One example is the Visalogy, which can solve visual analogy problems that would be phrased as “A red car is to a blue car as a red bike is to what?” (Sadeghi, Zitnick, & Farhadi, 2015). Though successful in its application, Visalogy is limited in regards to most analogies centered around action, attribute, or repositioning of an object (Sadeghi, Zitnick, & Farhadi, 2015). Another visual analogy application is demonstrated by utilizing the semantic and visual aspects of an image from a visual analogy data set to solve *A:B::C:D*-like problems (Lu, Liu, Ichien, & Holyoak, 2019). When evaluating text- and image-based AR methods, the former is a very thoroughly explored field compared to the latter. Challenges the latter faces include significant computational resources, additional processing and interpretation, and the tendency to be catered to an analogy-formatted image data set (meaning where the problem(s) to be solved are stated such that “Image A is to Image B as Image C is to what?”). A clear gap is the lack of image-based analogical reasoning applications using a general image data set, but also integrating it with a textual analogy to alleviate the computational and processing expectations typically associated with CV applications.

3. Methodology

Proposed to remedy this gap is the Image Recognition through Analogical Reasoning Algorithm (IRTARA), which integrates a CV algorithm that outputs declarations based on known classes, and an analogical reasoning algorithm that takes these

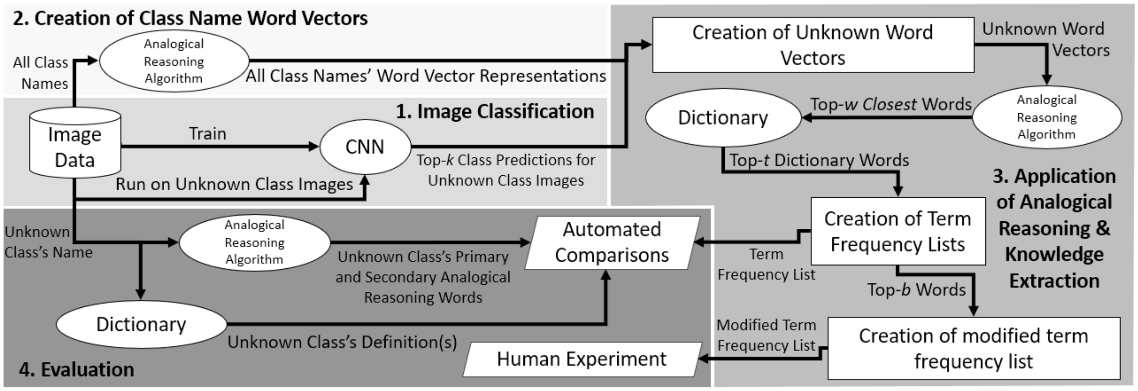

2. Creation of Class Name Word Vectors
All Class Names
Analogical Reasoning Algorithm
All Class Names' Word Vector Representations
1. Image Classification
Image Data
Train
CNN
Top-k Class Predictions for Unknown Class Images
Run on Unknown Class Images
Creation of Unknown Word Vectors
Unknown Word Vectors
Dictionary
Top-w Closest Words
Analogical Reasoning Algorithm
Top-t Dictionary Words
3. Application of Analogical Reasoning & Knowledge Extraction
Creation of Term Frequency Lists
Term Frequency List
Top-b Words
Modified Term Frequency List
Creation of modified term frequency list
Unknown Class's Name
Analogical Reasoning Algorithm
Unknown Class's Primary and Secondary Analogical Reasoning Words
Automated Comparisons
Dictionary
Unknown Class's Definition(s)
Human Experiment
4. Evaluation

Figure 1. IRTARA framework

declarations and searches for the meaning of unknowns. In addition to filling a gap for analogical reasoning applied to OOL CV scenarios, IRTARA had an additional advantage over former image-based methods by leveraging the matured text-based analogical reasoning application(s).

In operation, IRTARA operates as conceptualized in Figure 2. As presented in Figure 2, first (1) image data is considered through a convolutional neural network (CNN) to classify the data, then (2) class name word vectors are created, next (3) the application of a selected analogical reasoning algorithm occurs with knowledge extraction being the result from associations with a selected dictionary, and finally, (4) evaluation of the results occurs.

3.1. Image Classification

In developmental practice, IRTARA involves taking an image data set with class labels and training a CNN as would typically happen in a CV problem. Later this CNN is used on an unknown class of images, which can be from the data set (by removing one of the classes as such demonstrated later in this application) or which can be sourced from an external data set. IRTARA is modular and can utilize any CNN architecture that can produce probabilities associated with how likely that particular unknown class image belongs in the number of classes in which it is trained, *p*. IRTARA is only interested in the top-k classes where k can range from 1 to *p*.

3.2. Creation of Class Name Word Vectors

The pre-existing class names are converted into their respective word embedding vectors via the analogical reasoning algorithm. IRTARA sends the class names to the analogical reasoning algorithm, which retrieves the pre-trained word embedding vectors. In the case where the class name is not within the analogical reasoning algorithm's vocabulary, it may need to be altered into a "representative" version that is recognizable by the algorithm. These class name word vector representations are used with the image classification's class predictions in the next process.

3.3. Application of Analogical Reasoning & Knowledge Extraction

An "unknown word vector" needs to be created for each unknown class image to ideally "represent" the unknown class. If the probability of the unknown class image belonging to the given class is greater than the threshold, α, its word vector representation is retrieved (from the immediate previous process), and it influences the unknown word vector. The class name word vector representation is multiplied by the probability of the image belonging

to that class for the top-k classes if their probability is greater than α shown as

$$\text{Unknown Word Vector} = (\text{Class 1 Word Embedding Vector}) * (\text{Probability of Class 1}) + \ldots + (\text{Class } k \text{ Word Embedding Vector}) * (\text{Probability of Class } k), \tag{1}$$

where the "Class 1 Word Embedding Vector" is a vector with 300 dimensions ranging from 0 to 1. For example, if the CNN classified the given object as "lightning" with a probability of 0.356 and as a "comet" with a probability of 0.178, the resulting unknown word vector would be calculated as

$$\text{Example Unknown Word Vector} = \text{WV}[\text{‘lightning’}] * 0.36 + \text{WV}[\text{‘comet’}] * 0.18, \tag{2}$$

where WV is the word vector representation of the given word. Given the scenario where Class 1's probability is not greater than α, it will be used as the unknown word vector's only influence. The unknown word vector is sent back to the analogical reasoning algorithm, which is tasked with identifying the closest-w words whose word embeddings best match the unknown word vector. IRTARA is interested in the closest-w words, where w can range from one word to the entire vocabulary of the VSM, w_{max}.

Given the closest-w words, they are each sent to the selected dictionary, which retrieves each word's definition(s). In the case that the word identified by the analogical reasoning algorithm does not exist in the dictionary's vocabulary, it is skipped. The definitions are then modified to remove any "stop words" (as defined per (Bird, Loper, & Klein, 2009)) in addition to other words that lacked significant semantic meaning which is at the discretion of the experimenter. The remaining words, dubbed "definition words," are compiled. Starting with the creation of the unknown word vector, this entire process is repeated for each image within the unknown class. This yields a large list of words, which are filtered down to the top-t words based on how frequently they occur. Theoretically, t can range from 1 to the total number of unique words identified; however, t has a direct relationship with computation time. This list of the top-t words is the term frequency list and is the final product of IRTARA.

3.4. Evaluation Process

Two modeling approaches were implemented to predict the quality of the term frequency list, which in return were compared with human judgment. The definition evaluation automatically identifies words that are directly related to the unknown class; whereas, the analogical reasoning evaluation automatically detects associated words to the unknown class. The goals of the definition and analogical reasoning evaluation methods are to quantitatively evaluate the term frequency list in an automated fashion. The human experiment is to provide a qualitative human-created benchmark used to determine how closely the automated methods reflect a human's assessment

3.4.1. Definition Method. This evaluation method compared the term frequency list to the words found in the unknown class's definition(s). This analysis was able to determine directly related words used to describe the unknown class. This method utilized two different metrics due to many words having multiple meanings.

To establish a baseline, each unknown class's definition was retrieved from the same dictionary used in the previous process. Similar to the definition words described earlier, the same stop words and words deemed to lack semantic meaning were also unknown from each unknown class's definition(s). The remaining words are called "all words" since they are the words found in all the definitions for the unknown class. If multiple definitions, the true definition is identified and the words found within are called "true words." In the case there is only one definition, the all and true word lists are identical.

The term frequency list is compared against the unknown class's true and all words. The "best-case scenario" is for the term frequency list to overlap with the true words because these accurately describe the unknown class. However, the "all words" are also considered in case IRTARA can pick up other meanings of the unknown class (if applicable). Since each unknown class's definition has varying word length, the two metrics from this method are expressed as percentages, namely the true word percentage, TW%, and the all words percentage, AW%.

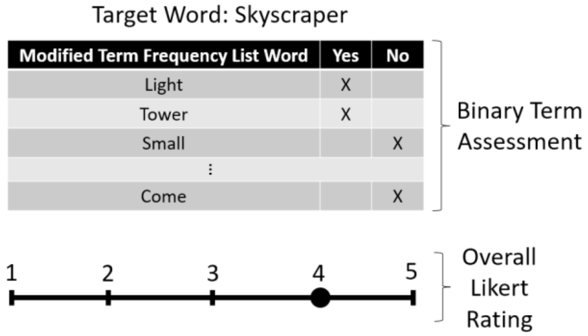
3.4.2. Analogical Reasoning Method. The analogical reasoning evaluation method seeks to identify associated words to the unknown class that may not appear directly in the definition. This evaluation method has two metrics looking at the overlap between the term frequency list and the unknown class's primary and secondary words as determined by the analogical reasoning algorithm.

Primary words are the top-u words closest to the unknown class based on cosine similarity. Furthermore, the top-v words closest to the top-u primary words based on cosine similarity are considered the number of secondary-per-primary words. There can be up to uv_{max} secondary words; however, after excluding word variations and duplicates, it is usually less than uv_{max}.

The term frequency list ideally overlaps with primary words; however, secondary words are also related to the unknown class to a lesser degree. The metrics that use this information are the percentage of primary words, PW%, and the percentage of secondary words, SW%, which looks at how many words in the term frequency list are also primary or secondary word, respectively.

3.4.3. Human Experiment. The final and only non-automated method looks at how a human judges the quality of the term frequency list. The human experiment only considers a modified term frequency list, which consists of the top-b words of the original term frequency list. The human experiment yields two metrics: descriptive words, DW, the number of words the majority of respondents deems descriptive for an unknown class (based on the "binary term assessment" portion of the survey), and a quality score, QS, which reflects how well the descriptive words describe the unknown class (based on the "overall Likert rating" portion of the survey). Figure 2 is an example of how the questions for both portions were presented to the respondents.

The Binary Term Assessment determines how many of the modified term frequency list words "describe" the target word, which is the class man. Looking at the example in Figure 2's Binary Term Assessment section, the respondent is presented with the target word, "skyscraper," and words from the modified term frequency list. The respondent would go through the latter words one-by-one and answer "yes" or "no" to whether, "by itself, in combination of another listed word or its characteristics, [would the given modified term frequency list word] describe or could be associated with [the unknown class]?" Returning to the example, the respondent would ask the aforementioned question for each word "light," "tower," etc., and mark their respective answers in the column like what's shown in Figure 2. The number of "Yes's" was compiled for each of the *b*-word(s) and then, the number of modified term frequency list words where a majority (defined as 50% + 1) of respondents said "Yes" were summed as the unknown class's descriptive words. This process was dubbed the "binary term assessment."


Target Word: Skyscraper

Modified Term Frequency List Word	Yes	No
Light	X	
Tower	X	
Small		X
⋮		
Come		X

Figure 2. Example survey section

The quality score had the respondents look at the descriptive words identified in the previous step and give a ranking between 1-5 on a Likert scale regarding how well they, as a whole, described the unknown class. As shown in Figure 2, the respondents were given a slider that could accept values between 1 and 5 inclusive. This value is averaged across all respondents and the standard deviation was also calculated. This process was called the "overall Likert rating."

4. IRTARA Module and Parameter Selection

IRTARA requires four modules which are the data set, CNN, analogical reasoning algorithm, and dictionary. Six IRTARA parameters were selected by the authors in addition to one determined by the data set selected.

4.1. Module Selection

"Module" is a term referring to any portion of IRTARA that can be replaced with another data set, algorithm, or application. Briefly, and described below, IRTARA was applied to the Caltech-256 data set (Griffin, Holub, & Perona, 2007), IRTARA was incorporated with a shallow 11-layer CNN, the GloVe (Pennington, Socher, & Manning, 2014) analogical reasoning algorithm, and the PyDictionary (Bora, 2020) and Lexico (Oxford University Press, 2021) dictionaries.

4.1.1. Data Set Selection. Though not explicitly part of the process shown in Figure 1, before the running of IRTARA, an appropriate data set needed to be selected. Ideal characteristics of a data set include 1) variety and scope of classes, 2) focus on single objects in an image, and 3) availability of baseline results. Based on these various needs, Caltech-256 (see (Griffin, Holub, & Perona, 2007)) was selected to test this framework based on its variety of concrete classes and depth of samples per class. Caltech-256 has 257 classes, but for this study, the 257th "clutter" class was not considered, leaving 256 classes. In each iteration tested, the unknown class is taken from Caltech-256 classes so it is "known" to the experimenter, but "unknown" to IRTARA. This means that the CNN is trained on 255 classes, and then attempts to classify the remaining "unknown" class accordingly. For example, if the unknown class in the first iteration is "coffee mug," the coffee mug images would be set aside and the CNN would train on the remaining 255 classes. In the next iteration, if the unknown class was "American flag," the CNN would train on the remaining 255 classes, coffee mug included, and then attempt to classify the American flag images. Caltech-256 also included images of varying sizes and thus all were resized within the IRTARA algorithm to 128 x 128 pixels grayscale images.

4.1.2. Image Classification Method. The CNN is the primary concern of the image classification section (enumerated process 1) of Figure 1. To avoid computational demands required in applying deep CNN architectures, a 11-layer CNN that balanced accuracy and computational performance were developed for IRTARA demonstration purposes. The CNN had the following architecture:

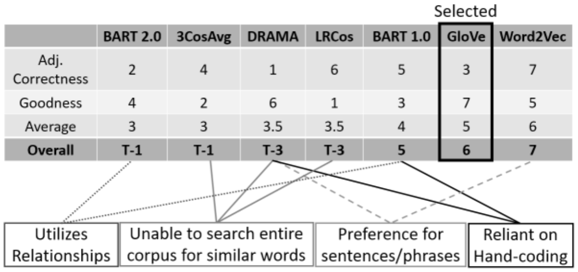
128x128-1C128-MP2-16C63-MP2-32C30- MP2-64C14- MP2-128C6-128N-255N (3)

using the Cireşan-CNN-representation from (Bihl, Schoenbeck, Steeneck, & Jordan, 2020).

When trained on all 256 classes, this architecture had an average of 22.5% classification accuracy across 10 runs (compared to 38% for (Griffin, Holub, & Perona, 2007) across 40 runs), where *optimizer* = `Adam()`, *batch_size* = 32, *epochs* = 10, and *validation_split* = 0.1. Despite a CNN with higher accuracy, e.g., ResNet (He, Zhang, Ren, & Sun, 2016) or VGGNet (Simonyan & Zisserman, 2015), being likely to yield better results, such CNNs are computationally costly and the algorithm of (3) can be rapidly retrained to assess IRTARA. Thus, the development and this demonstration of IRTARA focused on the whole process and used this simple CNN which trained quickly on standard desktop hardware.

4.1.3. Analogical Reasoning Method. To select the analogical reasoning algorithm (involved in processes 2, 3, and 4 of Figure 1), the review of AR algorithms from (Combs, Bihl, Ganapathy, & Staples, 2022) was used and the following methods were considered: Bayesian Analogy with Relational Transformations (BART) 1.0 (Lu, Chen, & Holyoak, 2012) and 2.0 (Lu, Wu, & Holyoak, 2019), 3 Cosine Average (3CosAvg) (Drozd, Gladkova, & Matsuoka, 2016), Distributed Representation Analogy MApper (DRAMA) (Eliasmith & Thagard, 2001), Linear Regression Cosine (LRCos) (Drozd, Gladkova, & Matsuoka, 2016), GloVe (Pennington, Socher, & Manning, 2014), and Word2Vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) (Mikolov, Tomas, Yih, & Zweig, 2013). To select an AR method for IRTARA, the adjusted correctness (based selection of the correct answer) and goodness (how close to an "ideal" analogy the correct answer is according to the algorithm) (Combs, Bihl, Ganapathy, & Staples, 2022). This produced the ranking of methods seen in Figure 3. From this process and other constraints, GloVe was selected based on that it does not utilize analogy relationships, it does use singular-word embeddings (as opposed to multi-word phrases), and its ease of implementation is visually shown in Figure 3. Specifically, the Glove-wiki-gigaword-300 model was used, which was pre-trained on 2014 Wikipedia and Gigaword 5 textual data to create the word vector for 400,000 words, each with 300 dimensions (Pennington, Socher, & Manning, 2014). In certain cases, GloVe was unaware of the original class name, so a substitute representation was used. These representations typically were a simplification (e.g., "American flag" to "flag") or a merge between two-word vectors representing different words (e.g., "baseball bat" to "baseball" & "bat") so as long they would not be confused with another class. This verification and revision process would need to be repeated whenever different word embeddings are used as well as when a different image data set is used.

4.1.4. Dictionary. The dictionary is important in process 3 of Figure 1, application of analogical reasoning and knowledge extraction. The primary dictionary used was the external Python library, PyDictionary, based on ease of integration with IRTARA and standardized definition format. This dictionary was used in processes 3 and 4 of Figure 1. PyDictionary uses WordNet (see (Princeton University, 2010)) for its definitions which were created in 1995. Considering how language has evolved and changed since then, some of the definitions came from an alternative dictionary, Lexico, which affected 16 classes. In addition to using an alternative dictionary, some of the definitions were created by simplifying the original class name (e.g., "self-propelled lawn mower" to "mower") and merging individual words' definitions (e.g., "cowboy hat"). Like with the GloVe word vector representations, the definitions would be verified and modified accordingly whenever a new dictionary is used and/or a different image data set.



	BART 2.0	3CosAvg	DRAMA	LRCos	BART 1.0	GloVe	Word2Vec
Adj. Correctness	2	4	1	6	5	3	7
Goodness	4	2	6	1	3	7	5
Average	3	3	3.5	3.5	4	5	6
Overall	**T-1**	**T-1**	**T-3**	**T-3**	**5**	**6**	**7**

Selected
Utilizes Relationships
Unable to search entire corpus for similar words
Preference for sentences/phrases
Reliant on Hand-coding

Figure 3. AR Justification, from (Combs K. L., 2021)

4.2. Parameter Selection

IRTARA requires several parameters to be determined before running, which are described and shown in Table 1. Only the number of CNN classes, *p*, is dictated outside of the user's control since it's reliant on the input image data set. The remaining parameters were selected based on brief, informal experimentation on a range of varying values.

Table 1. IRTARA parameters

Parameter		Val.
# of CNN classes	p	256
# of top classes per image	k	5
Thres. for influencing unknown word vector	α	0.05
# of closest words per unknown word vector	w	5
# words in the term frequency list	t	100
# of primary words	u	20
# of secondary words	v	10
# of words in the modified term frequency list	b	21

5. Results

Fifteen Caltech-256 classes were chosen at random to be the unknown class, which yielded a wide range of results. Tables 2-5 show the classes and their top 5 term frequency words in order of their overall rankings found in Table 6 (see (Combs K. L., 2021) for full results). Broadly looking at the lists many words repeat across lists, which may be due to the dictionary's tendency to frequently use those words in its definitions.

Each evaluation method yielded two metrics and a rank assigned to each class ranging from 1 to 15. In case of a tie in rank, the average of the places was used, i.e., if two classes were tied for third and fourth place, they would both receive a score of 3.5. Each evaluation method's rank contributed equally to the overall rank.

5.1. Definition Evaluation

The definition evaluation produced two metrics that looked at the percentage of words in the term frequency list that also appears "true" definition, TW%, and those that appear in any definition of the unknown class, AW%. The percentage of true words ranges from 0-50%; whereas, the percentage of all words range was slightly lower, between 0% to 33.3%. The top-three performing classes were skyscraper, t-shirt, and iguanas. The classes were ordered from highest to lowest based on the true words and all words percentage and assigned a rank between 1 and 15. An average of these rankings was taken, ordered, and ranked again for the definition evaluation rank shown in Table 6.

5.2. Analogical Reasoning Evaluation

The analogical reasoning evaluation also produced two metrics that looked at the percentage of words in the term frequency list that also appeared as primary or secondary words, PW% and SW%, respectively. Around the board, most classes had low scores for both metrics; however, the top 3 in both were Mars (15%; 5.4%), galaxy (15%; 3.8%), and skyscraper (15%; 2.7%). Similarly, to the definition evaluation rank, there was a rank assigned to each class based on its percentage of primary and secondary words. An average of these ranks was used to order and rank the classes for the analogical reasoning rank shown in Table 6.

5.3. Human Experiment

These results were derived from a homework assignment given to a mix of 25 undergraduate/graduate students enrolled in a Midwestern university's introductory human factors engineering class. The class consisted of 10 graduate and 15 undergraduate students all of whom were pursuing a degree within the biomedical, industrial, and human factors engineering department. Of the 25 subjects, 7 were male and 18 were female. The results reflect a 26th respondent, which are the opinions of the first author who was also a female graduate student in the department.

The descriptive words metric ("DW" column in Table 6), looking at the number of words at least (50% + 1) respondents thought were relevant, ranged from 4-18 words (out of 21 total words in the modified term frequency list). This metric showed the top classes to be galaxy (18), fireworks (14), and iguanas (13). The second metric from the human experiment, the quality score, QS, ranged from 1 to 5, which was averaged across all 26 respondents' responses. Most of the classes are statistically similar to one another, with the exception being the galaxy class with a score between 3.69-5.

A rank was calculated for both metrics like in the previous methods. The quality score rank was based only on the average quality score. These rankings were averaged and ranked against to calculate the human experiment rank for this method. The top three classes saw galaxy rank first followed by the tied classes, fireworks and t-shirt.

5.4. Overall Rank

The top-three classes, t-shirt, skyscraper, and iguanas, consistently scored in the top 50% of the rankings for all three evaluation ranks. Whereas, at the bottom, floppy disk (11.5), sheet music (13), chandelier (14), and AK-47 (15) consistently ranked in the bottom 50% for all three evaluation methods. The best example with varying results is galaxy, which performed poorly on the definition evaluation (14), but well on the analogical reasoning evaluation (2) and the human experiment (1). These results suggest that IRTARA performs consistently at the extreme ends, but to a lesser extent with mid-range results. Spearman's rank coefficient was calculated for both automated methods in comparison with the human experiment. The definition-human evaluation yielded a $\rho = 0.168$ and $p\text{-}value = 0.549$ and the analogical-reasoning-human evaluation yielded a $\rho = 0.434$ and $p\text{-}value = 0.082$.

6. Conclusions

Of significant interest to artificial intelligence (AI) research is to accurately interpret and describe out-of-library (OOL) objects (Situ, Friend, Bauer, & Bihl, 2016). Analogical reasoning has been proposed to assist with this end goal in both textual and visual scenarios; however, there has been limited research conducted regarding its application in computer vision (CV) problems. This paper describes the Image Recognition Through Analogical Reasoning Algorithm (IRTARA), which integrates standard image classification methods from CV with the semantic meaning and interpretation from an analogical reasoning algorithm and dictionary.

IRTARA consists of four processes. Image classification is the first process that involves training the CNN on the *p* classes and then, running the image data from the unknown class through it. The second process creates and assigns a word vector representation (from the selected analogical reasoning algorithm) to each class name within the dataset. The third process applies analogical reasoning by creating an unknown word vector for each of the unknown class images (by taking the top-k predicted class's word vector representation multiplied by the image's probability of belonging in said predicted class if the probability is greater than the minimum threshold, α). The analogical reasoning algorithm uses the unknown word vector to identify the closest-*w* word vector found within its vocabulary. The definition(s) of these words are pulled from the selected dictionary and the words with semantic meaning are called "definition words." The top-t most frequently occurring definition words for an unknown class are compiled in its term frequency list. In the final stage of IRTARA, evaluation, the quality of the term frequency list was measured through two automated methods, the definition and analogical reasoning evaluations, to be compared to the results from the human experiment. The definition evaluation method identifies directly-related words as found in the unknown class's definition; whereas, the analogical reasoning considers associated words based on the top-*u* primary words and up to the top-*uv* secondary words. The human evaluation exists to create a baseline for how a human might judge the term frequency list compared to the two automated methods.

Table 2. Top-5 words of select term frequency lists for t-shirt, skyscraper, iguana, & galaxy

Rank	1. T-shirt		2. Skyscraper		3. Iguana		4. Galaxy	
	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
1	Ball	383	Light	81	Long	118	Planet	153
2	Light	289	Tower	54	Small	106	Sun	135
3	Game	270	Small	54	Large	85	Mythology	92
4	Face	253	Building	49	Coat	65	Th	80
5	Small	243	Little	48	Genus	61	Small	71

Table 3. Top-5 words of select term frequency lists for fireworks, mars, frog, & rainbow

Rank	5. Fireworks		6. Mars		7. Frog		8. Rainbow	
	Word	Freq	Word	Freq	Word	Freq.	Word	Freq.
1	Large	81	Brain	263	Large	114	Light	604
2	Small	54	Skull	227	Body	97	Little	321
3	Long	54	Nervous	226	Fungi	92	Illumination	299
4	Cloud	49	Ability	226	Small	80	Fire	242
5	Light	48	Planet	224	Edible	72	United	219

Table 4. Top-5 words of select term frequency lists for people, Swiss army knife, floppy disk, & waterfall

Rank	9. People		10. Swiss Army Knife		11.5. Floppy Disk		11.5. Waterfall	
	Word	Freq.	Word	Freq.	Word	Freq.	Word	Freq.
1	Large	186	Small	93	Small	60	Fungi	157
2	Body	160	Ball	76	Ball	49	Large	133
3	Ball	148	Instrument	71	Body	48	Fleshy	117
4	Move	143	Body	71	Long	47	Body	103
5	Small	138	Device	60	Device	44	Edible	93

Table 5. Top-5 words of select term frequency lists for sheet music, chandelier, & AK-47

Rank	13. Sheet Music		14. Chandelier		15. AK-47	
	Word	Freq.	Word	Freq.	Word	Freq.
1	Small	137	Small	91	Long	67
2	Rectangular	88	Observe	68	Small	61
3	Area	86	Person	60	Move	59
4	Glass	85	Determine	53	Person	48
5	Box	76	Light	49	Played	42

Table 6. Summary of results

Unknown Class	Definition Evaluation			Analogical Reasoning Evaluation			Human Experiment			Overall Rank
	TW%	AW%	Rank	PW %	SW%	Rank	DW	QS (avg ± stdev)	Ran k	
Ak-47	0%	0%	14	0%	0%	14	7	2.04 ± 0.77	14.5	15
Chandelier	0%	0%	14	0%	0%	14	5	2.52 ± 0.95	13	14
Fireworks	12.5%	12.5%	7	0%	1.7%	9	14	3.62 ± 0.85	2	5
Floppy Disk	0%	6.7%	11.5	0%	0.9%	11	9	2.81 ± 1.02	10	11.5
Frog	10%	15%	8	0%	2.4%	6	11	2.92 ± 0.9	7.5	7
Galaxy	0%	0%	14	15%	3.8%	2	18	4.5 ± 0.81	1	4
Iguanas	31.3%	31.3%	3	0%	2.1%	7	13	3.15 ± 0.89	4	3
Mars	23.5%	20.6%	5	15%	5.2%	1	4	2.58 ± 1.14	11	6
People	0%	16.7%	10	5%	2.2%	4	9	2.69 ± 1.05	7.5	9
Rainbow	11.1%	9.1%	9	0%	1.8%	8	9	3.42 ± 0.81	5	8
Sheet Music	0%	6.7%	11.5	0%	1%	10	7	2.08 ± 1.06	15	13
Skyscraper	50%	50%	1	15%	2.7%	3	8	3.35 ± 0.89	6	2
Swiss Army Knife	16.7%	16.7%	6	0%	0%	14	9	2.65 ± 0.85	9	10
T-shirt	40%	33.3%	2	5%	1.5%	5	12	3.27 ± 1.12	3	1
Waterfall	25%	25%	4	0%	0.8%	12	5	2.27 ± 0.96	12	11.5
Average	13%	14%		3%	2%		9.24	2.91 ± 0.93		

To test IRTARA's methodology, evaluated against the Caltech-256 dataset (Griffin, Holub, & Perona, 2007), with parameters consistent with those found in Table 1. Overall, the three evaluation methods show consistency with unknown classes that have performed on the extreme ends (exceptionally well or poorly). Namely, when acting as the unknown class, t-shirt, skyscraper, and iguana consistently ranked in the top half across all three evaluations. On the other spectrum, floppy disk, sheet music, chandelier, and AK-47 consistently ranked in the bottom half across the evaluation methods. However, there was a significant amount of ambiguity for those that rank in-between. Using the Spearman rank coefficient to compare how well the automated methods match the human experiment ranks, it was determined that the analogical reasoning evaluation ranks had a higher correlation (ρ = 0.43; *p-value* = 0.08) compared to the definition evaluation ranks (ρ = 0.17; *p-value* = 0.55).

7. Acknowledgements

The views expressed in this paper are those of the authors and do not represent the views of any part of the US Government. This work was cleared for unlimited release under: AFRL-2022-2372.

8. References

Bihl, T., Schoenbeck, J., Steeneck, D., & Jordan, J. (2020). Easy and efficient hyperparameter optimization to address some artificial intelligence “ilities”. *Proceedings of the 53rd Hawaii international conference on system sciences* (pp. 943-952).

Bird, S., Loper, E., & Klein, E. (2009). *Natural Language Processing with Python.* Sebastopol: O'Reilly Media Inc.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics, 5*, 135-146.

Bora, P. (2020). *PyDictionary 2.0.1*. Retrieved from PyPi: https://pypi.org/project/PyDictionary/

Combs, K. L. (2021). *Application of analogical reasoning for use in visual knowledge extraction.* Wright State University. OhioLINK: Electronic Theses and Dissertation Center.

Combs, K., Bihl, T. J., Ganapathy, S., & Staples, D. (2022). Analogical reasoning: An algorithm comparison for natural language processing. *Proceedings of the 55th Hawaii International Conference on System Sciences.*

Doumas, L. A., & Hummel, J. E. (2010). A computational account of the development of the generalization of shape information. *Cognitive Science, 34*, 698-712.

Drozd, A., Gladkova, A., & Matsuoka, S. (2016). Word embeddings, analogies, and machine learning: Beyond king-man+ woman= queen. *Proceedings of coling 2016, the 26th international conference on computational linguistics.*

Eliasmith, C., & Thagard, P. (2001). Integrating structure and meaning: A distributed model of analogical mapping. *Cognitive science, 25*(2), 245-286.

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. *Proceedings of the April 21-23, 1964, spring joint computer conference.*

French, R. M. (2002). The computational modeling of analogy-making. *Trends in cognitive science, 6*(5), 200-205.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science, 10*(3), 277-300.

Gentner, D., & Maravilla, F. (2018). Analogical reasoning. In L. J. Ball, & V. A. Thompson, *International Handbook* *of Thinking & Reasoning* (pp. 186-203). New York: Psychology Press.

Google. (2021, December). *Google Research*. Retrieved from Perception: https://research.google/teams/perception/

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset.*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Hofstadter, D. R., & Mitchell, M. (1995). The copycat project: A model of mental fluidity and analogy-making. *Advances in connectionist and neural computation theory, 2*, 205-267.

Holyoak, K. J., & Thagard, P. (1989). Analogical mapping by constraint satisfaction. *Cognitive science*, 295-355.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological review, 104*(3), 427.

Hwang, S. J., Grauman, K., & Sha, F. (2013). Analogy-preserving semantic embedding for visual object categorization. *Proceedings of the 30th international conference on machine learning.*

IBM. (2021, December). *What is artificial intelligence (AI)?* Retrieved from IBM: https://www.ibm.com/topics/artificial-intelligence

Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior research methods, 52*(5), 1803-1816.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation, 1*(4), 541-551.

Liu, C., Zoph, B., Neumann, M., Shlens, J., Hua, W., Li, L.-J., . . . Murphy, K. (2018). Progressive neural architecture search. *Proceedings of the European conference on computer vision (ECCV)* (pp. 19-34).

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological review, 119*(3).

Lu, H., Liu, Q., Ichien, N. Y., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. *Proceedings of the 41st annual conference of the cognitive science society* (pp. 1-7).

Lu, h., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the national academy of sciences, 116*(10), 4176-4181.

McCarthy, J. (2004). *What is artificial intelligence?* Stanford University, Computer Science Department.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111-3119.

Mikolov, Tomas, Yih, W.-t., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 746-751).

Oxford University Press. (2021). *Lexico*. Retrieved from Lexico: https://www.lexico.com/

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

Polya, G. (1990). *Mathematics and plausible reasoning: Induction and analogy in mathematics* (Vol. 1). Princeton: Princeton University Press.

Princeton University. (2010). *About WordNet*. from https://wordnet.princeton.edu/citing-wordnet

Raven, J. C., & Court, J. H. (1938). *Raven's progressive matrices.* Los Angeles: Western Psychological Services.

Reed, S. E., Zhang, Y., Yuting, Z., & Lee, H. (2015). Deep visual analogy-making. *Advances in neural information processing systems.*

Rogers, A., Drozd, A., & Li, B. (2017). The (too many) problems of analogical reasoning with word vectors. *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)* (pp. 135-148).

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). Visalogy: Answering visual analogy questions. *Advances in neural information processing systems.*

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the third international conference on learning representations.*

Situ, J. X., Friend, M. A., Bauer, K. W., & Bihl, T. J. (2016). Contextual features and Bayesian belief networks for improved synthetic aperture radar. *Military operations research, 21*(1), 89-106.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. *Advances in neural information processing systems, 26*, 935-943.

Wilson, W. H., Halford, G. S., Gray, B., & Phillips, S. (2001). The STAR-2 model for mapping hierarchically structured analogs. In D. Gentner, K. J. Holyoak, & B. N. Kokinov, *The analogical mind* (pp. 125-160). Cambridge: MIT Press.

Zhang, C., Gao, F., Baoxiong, J., Zhu, Y., & Song-Chun, Z. (2019). RAVEN: A dataset for relational and analogical visual reasoning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5317-5327).