B-AIS: An Automated Process for Black-box Evaluation of Visual Perception in AI-enabled Software against Domain Semantics

Hamed Barzamini, Mona Rahimi
Northern Illinois University
DeKalb, USA
{hbarzamini,mrahimi1}@niu.edu

ABSTRACT

AI-enabled software systems (AIS) are prevalent in a wide range of applications, such as visual tasks of autonomous systems, extensively deployed in automotive, aerial, and naval domains. Hence, it is crucial for humans to evaluate the model's intelligence before AIS is deployed to safety-critical environments, such as public roads.

In this paper, we assess AIS visual intelligence through measuring the completeness of its perception of primary concepts in a domain and the concept variants. For instance, is the visual perception of an autonomous detector mature enough to recognize the instances of *pedestrian* (an automotive domain's concept) in Halloween customes? An AIS will be more reliable once the model's ability to perceive a concept is displayed in a human-understandable language. For instance, is the pedestrian in *wheelchair* mistakenly recognized as a pedestrian on *bike*, since the domain concepts bike and wheelchair, both associate with a mutual feature *wheel*?

We answer the above-type questions by implementing a generic process within a framework, called B-AIS, which systematically evaluates AIS perception against the semantic specifications of a domain, while treating the model as a black-box. Semantics is the meaning and understanding of words in a language, and therefore, is more comprehensible for humans' brains than the AIS pixel-level visual information. B-AIS processes the heterogeneous artifacts to be comparable, and leverages the comparison's results to reveal AIS weaknesses in a human-understandable language. The evaluations of B-AIS for the two vision tasks of pedestrian and aircraft detection showed a F_2 measure of 95% and 85% as well as 45% and 72% respectively in the dataset and model for the detection of pedestrian and aircraft variants.

CCS CONCEPTS

• **Software and its engineering → Requirements analysis**; • **Computing methodologies → Machine learning**.

KEYWORDS

AI-enabled Software, Domain Semantics, Multimodal Learning

ACM Reference Format:
Hamed Barzamini, Mona Rahimi. 2022. B-AIS: An Automated Process for Black-box Evaluation of Visual Perception in AI-enabled Software against Domain Semantics. In *37th IEEE/ACM International Conference on Automated Software Engineering (ASE '22), October 10–14, 2022, Rochester, MI, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3551349.3561162

1 INTRODUCTION

The application of data-driven AI (deep learning) has become widespread in a large variety of domains, from computer science to philosophy and ethics. The AI models, driven by sampling data, do not conventionally implement pre-defined requirements specifications [5, 6]. Instead, their inductive nature tends to learn the specifications from training samples.

This characteristic of deep learning (DL) models is desirable for engineering software systems operating in domains, which contain concepts that are difficult for humans to specify, and hence, are difficult to be programmed (*i.e., hard-to-specify* domain concepts). For instance, what is a comprehensive definition of a potential pedestrian? Specifying, and hence programming, the characteristics of pedestrians versus non-pedestrians is difficult for humans. Note that specification refers to providing a general description of both types, which is precise for the existing samples but additionally is comprehensive to cover a large variety of unseen (will be seen in future) instances. In such cases, DL models, such as deep convolutional neural networks, are trained on large amounts of historical data from diverse pedestrians so as to learn the visual characteristics of pedestrians and non-pedestrians, and distinguish them from each other. After training, the models' inductive perception (learnings) of the domain concept and its variants (*e.g.,* the pedestrian variants) will be then generalized to unseen instances of the AIS operating environment.

While often successful, generalizing the models perception from seen instances to not-yet-seen samples may lead to misclassification of non-pedestrians as potential pedestrians (false positives), or discarding the actual pedestrians (false negatives). The problem exacerbates, once types and their features' variance significantly differ between the two sub-populations of training and testing. Misclassifications cause severe hazards, especially in domains with safety applications, such as automotive, in which the mistypes of visual perception tasks in the autonomous systems (AS) are hazardous and safety-related [50]. In such scenarios, confidence in AIS perception of domain concept variants (instances with various appearance), which AIS is initially designed and trained to perceive, is essential for safety assurance purposes [67].

In this regard, several approaches attempt to reason about the AIS decisions and explain their class predictions. For instance, the research area of eXplainable Artificial Intelligence (XAI) attempts

ASE '22, October 10–14, 2022, Rochester, MI, USA

ACM ISBN 978-1-4503-9475-8/22/10…$15.00
https://doi.org/10.1145/3551349.3561162

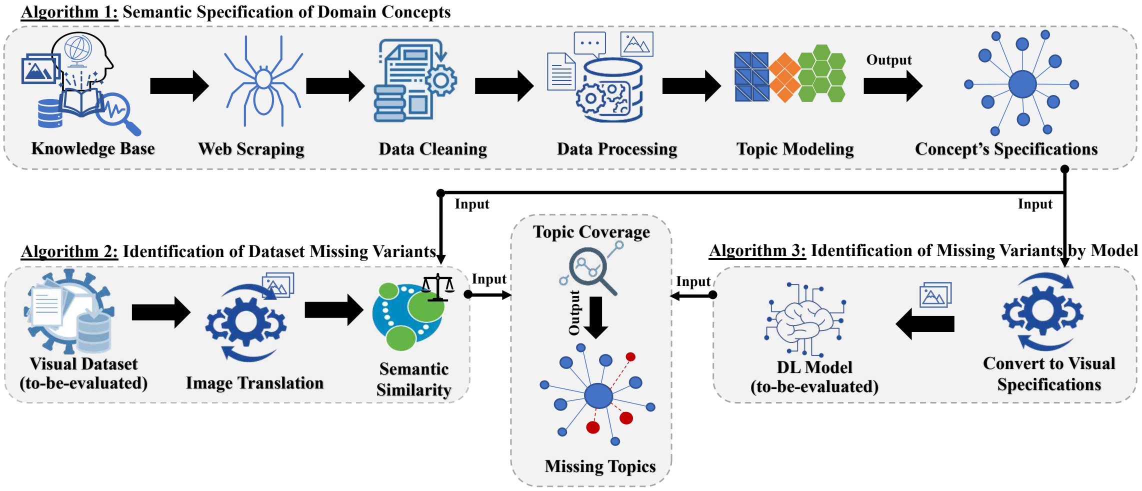
Algorithm 1: Semantic Specification of Domain Concepts
Knowledge Base
Web Scraping
Data Cleaning
Data Processing
Topic Modeling
Output
Concept's Specifications
Input
Input
Algorithm 2: Identification of Dataset Missing Variants
Visual Dataset (to-be-evaluated)
Image Translation
Semantic Similarity
Input
Topic Coverage
Output
Missing Topics
Input
Algorithm 3: Identification of Missing Variants by Model
DL Model (to-be-evaluated)
Convert to Visual Specifications

Figure 1: B-AIS: An Automated Framework for Concept Augmentation and Deep Learning Models Evaluation.

to generate human-understandable explanations for decisions made by AIS, with the ultimate purpose of correcting and improving the models inference of concepts, for which the model was initially trained. Our research has a small common ground with XAI, since both focus on the power of evaluating AI models decisions for humans [48]. However, this work **differs from XAI**, since the aim here is not to explain why and how a correct or incorrect prediction is made, rather our concern is to evaluate AIS prediction of the concepts variants, which are shown to be semantically important to the humans' perception.

The current research in XAI, fails to provide a benchmark or point of reference, against which the models' perception of the domains concepts could be evaluated [95]. This is often up to the end users to decide, whether the models' learnings is *enough* to cover the concept variants. This seems relevant in the scenarios, for which specific domain-knowledge is required to interpret the models' predictions (*e.g.,* classifying different types of birds). Yet, in the majority of domains which adopt AI for visual perception, the domain concepts are common knowledge, such as specifications of a pedestrian. The concepts' variability, learned by the model, requires to be compared against an external, reliable and comprehensive knowledge base. While the existing approaches are helpful in reasoning about the models' decisions, they fail to identify what the model has not yet learned, with respect to the specifications of the concepts potentially-existing variants.

To this end, this research aims to fill the gap between the specification (conceptualization) of a targeted domain concept and its visualization in a dataset, and what a model learns of the concept variants. We exploit the semantic specifications of domain concepts, as a reference point, to evaluate the relative completeness and accuracy of the variability of a domain concept in AIS training data, as well as in the models' perception of the concept.

The proposed process initially specifies the variants of a concept, which are important to the accuracy of a model's perception of the concept, while AIS is in operation. Referring to the derived specifications, diversity and relative completeness of the concept's instances, expected to be recognized by the model, are evaluated once in the dataset, and once again in the model. The model is treated as a black-box during the assessment, since we solely focus on the model's final prediction of each variant, while excluding attention to the process of decision making. The evaluation results showed that a systematic assessment of AIS, with respect to the semantics of a domain, enables us to independently determine the missing concept variants in the dataset and in the model. Once dataset augmentation and model re-training is based on a benchmark, which represents the semantics of a targeted concept, rather than on an adhoc basis and in the ignorance of the actually missing variability of a concept, the resultant AIS will be more reliable.

In this document, we specifically focus on supervised training (solution-based dataset) for the visual perception tasks in data-driven AIS (deep neural networks), and not other machine learning paradigms. This paper, in particular seeks to answer the research questions below:

- ***RQ*$_1$** (Functionality): Can semantic specifications of domain concepts serve as a benchmark to evaluate AIS visual perception of the concepts?
- ***RQ*$_2$** (Usability): How useful is this evaluation for the improvement of AIS faulty perception of the concepts variants?

Figure 1 illustrates an overview of the B-AIS modules and the interaction of the three primary algorithms within the framework, discussed in more details in the following sections.

2 SEMANTIC SPECIFICATION OF DOMAIN CONCEPTS (ALGORITHM 1)

For the evaluation of deep learning models, a reliable reference point in a human-understandable language is required. This section lays a semantic ground truth for the specifications of domain concepts.

AIS domain concepts are often socially specified, as their variants are large and unpredictable. Due to the concepts intuitive nature, humans have a common knowledge of what they mean, yet delineating the concepts in natural language is inherently difficult for humans. In fact, the concepts indescribable nature is the major motivation for adopting AI to specify them rather than programming their perception. For instance, for *pedestrian* no relatively complete domain document exist. Although there are a few general domain semantic webs that include a limited specification of the term pedestrian, such as WordNet [49], theses sources fail to adequately capture all varying instances of the concept in sufficient details. For example, WordNet defines a pedestrian as a “person who travels by foot” and associates the word with the terms *walker* and *footer*. This definition is limited given that it excludes, for example, pedestrians riding a bike, roller-skating, or using a wheelchair. It also fails to describe a pedestrian’s appearance in terms of attributes, such as clothing and posture.

To build a semantic benchmark, the primary function of algorithm one, $S_c \leftarrow Specify(c)$, receives a concept as an input (*e.g.* pedestrian) and returns a set of partial specifications. To tackle the challenge of specifying intuitive domain concepts, this function extracts domain knowledge from a large set of online knowledge-bases, such as online books, articles, encyclopedia, dictionaries, semantic webs, legal documents, social media, news feeds, as well as data in the form of image and video frames in public repositories. To specify the partial specifications, the online sources are automatically searched for both linguistic and visual information.

In cases that domain knowledge is stored in the textual format, a wide range of Natural Language Processing (NLP) techniques [9, 14, 89], such as topic modeling [89, 97], text classification [9, 97], and summarization [22, 31, 82] are adopted and adapted to automatically mine, retrieve and and process the textual information for *important* accompanying features of the concept. The importance of the features are determined based on a combination of multiple metrics, including semantic, lexical, syntactic characteristics, such as cosine similarity [59], frequency of co-occurrence [16], and grammatical importance [87].

Further to process visual sources of domain knowledge, such as available video and image sets, a variety of image processing techniques and convolutional neural networks (CNNs) [55, 80, 94], are adopted. For instance, scene graph generation (SGG) techniques [80, 92, 96], which translate the pixel-level visual data, such as video and image information, to natural language.

The process iteratively and incrementally specifies the concepts variants by identifying a list of terms (features) which repeatedly (1) *accompanying* the concept, such as terms which most frequently appeared before and after the term *pedestrian* in highly-ranked corpora (e.g. *careless* pedestrian and pedestrian *appearance*); (2) *co-occurring* with the concept, such as terms that frequently appear with the concept in the same phrase, paragraph, or page (e.g. *wheelchair*); (3) semantically are closely related to the given concept according to a large corpora (e.g. *sidewalk*).

Later, adopting a part-of-speech tagging algorithm[4], the most frequent relations between the concept and the extracted terms are specified. For instance, pedestrian *is* careless; *has* an appearance; *walks* on sidewalk; and *is* in wheelchair. As such, the final presentation of partial specifications are in the form of triplets of *“subject (concept variant)-predicates (the potential relations)-object (important features)”*, such as *pedestrian-is-man*, *man-has-bag*, *woman-walking-crosswalk*, and *handicap-sitting-wheelchair*.

Note that while implementation details of this function is further discussed in Section 5.1.1, a wide variety of techniques for building domain knowledge, such as semantic webs, are researched and proposed in the requirements engineering domain[1, 12, 21, 35, 47], and therefore, the implementation of this step is not limited to a certain approach. Furthermore, this work focuses on the exploitation of domain semantics, for assessing AIS visual perception, as the authors previously disseminated several systematic and pragmatic processes for the construction of semantic webs for hard-to-specify domain concepts (citations are removed due to the double-blind review process and will be added if accepted).

Due to heterogeneous information, namely linguistic and visual, the retrieved multimodal information is individually processed for each modality. As such, each source is independently processed and translated into a mutual and human-understandable language. The combined features from both sources are then organized into a universal structured format. To structure and re-use the extracted information, a series of machine learning (ML) techniques, such as classification and clustering [36, 88], are applied to meaningfully organize the information in a more human and machine-readable format, namely semantic webs [2, 7, 72]. Once built, the benchmark is leveraged for the purpose of explanation, assessment, and augmentation of AIS.

Algorithm 1 : Semantic Specification of Domain Concepts

```
Require: Domain concepts C against which AIS is evaluated.
$Array\ D_c \leftarrow C$                                ▷ Domain Concepts
function SPECIFY(c)
    for each $c \in C\ in\ D_c$ do
        Array $S_c \leftarrow Specify(c)$;
    end for
     return $S_c$
end function
```

3 IDENTIFICATION OF DATASET MISSING VARIANTS (ALGORITHM 2)

This section describes an algorithm, which takes the outcome of the first algorithm (partial specifications) as an input and assesses a given AIS dataset against the semantically important features of the domain concept and its variants.

The lack of diversity in AIS unsystematically- and arbitrarily-gathered training instances will lead to a gap between the specifications of a domain’s concept and what the model will learn as the targeted concept [34, 69]. To tackle this problem and improve AIS generalizability, a large body of research propose to apply dataset certification and augmentation methods [28, 33, 39, 40, 74, 91].

Data augmentation are techniques used to increase the amount of data by adding slightly modified copies of already existing data or newly created synthetic data from existing data[73]. These techniques leverage the existing instances of a domain concept in the data as the base for generating the additional samples, while leaving out variances of the concept which are initially missing from the original dataset. While data augmentation is sometimes effective, the process heavily relies on the assumption that more information can be extracted from the original dataset through augmentation. The dataset missing instances of a domain concept may contain a different set of features, entirely unfamiliar to the model, but similar to the actual input data during the AIS operation. For instance, the inspection of a commonly used pedestrian dataset revealed the lack of images of pedestrians in wheelchairs in several widely-used pedestrian datasets [57].

AIS commonly-used training datasets are often collected in unsystematic manners and therefore, are generally comprised of samples, limited in number and diversity [27, 61]. For instance, the established datasets in the context of autonomous driving, such as Caltech [19], KITTI [23], CityPersons [99], and EuroCityPerson (ECP) [11], are collected by a vehicle-mounted camera aimlessly navigating rural roads [27].

3.1 Image Conversion: $imgTranslate$ $(Image)$

In AIS visual perception tasks, the training datasets contain images and video frames of domain concepts, which AIS is expected to recognize during its operation. For instance, to train an object detector for recognizing pedestrians in the automotive domain, the model is trained with datasets, which contain video frames of different-looking instances of pedestrian, while *Algorithm 1*'s establishes specifications of domain concepts, which are representing humans' semantic knowledge, in the form of natural language. For the purpose of dataset assessment, the dataset's visual language is therefore translated into natural language.

Note that the XAI explainability differs from generating descriptions from images which we propose here. Descriptions report the visual information in an image, whereas in XAI explanations, the attempt is to understand why a certain class is appropriate for a piece of given visual information[29].

In the computer vision domain, a wide range of advanced R-CNNs, such as Faster R-CNNs [64], are proposed and developed for the performance of visual perception tasks (*i.e.* image classification and object detection) [65, 80, 81, 92]. A majority of these networks are previously trained on large-size datasets and are able to recognize various generic objects with a high accuracy.

In addition, a pre-trained model can be re-trained for a specific domain to improve the model's perception of varying instances of the domain concepts. For instance, models are trained for obstacle detection in a particular domain, since for example, potential obstacles are differently described in the automotive and naval domains (*e.g.* car vs. a light house). The trained neural models will then identify the potential regions of an image for the presence of objects, specify the objects extension (bounding boxes), and provide each object with a label with domain-specific terms [41].

Furthermore, as mentioned earlier, one area of work in computer vision focuses on developing SGG, generating natural language descriptions of multiple scenes (objects and their relations) in an image [80, 81, 92, 96]. Once trained, the neural model is able to identify and describe the features, associated with the domain concepts.

A more recent class of work took advantage of the encoder-decoder architecture in neural networks to train transformers, enabled to infer associations between different artifact types, such as between lingual (text) and visual (image) artifacts ViLBERT[46], VisualBERT [43], VLP[100], OSCAR[44, 98], DALL-E[60], and CLIP[56]. More details in this regard is provided in related work (Section 6).

Applying and repeating the above-mentioned methods for the primary concepts of a domain provide a human-understandable description of the concepts variants as they are represented in a dataset. Please note that the non-identified primary features, previously marked as important by algorithm 1, may not necessarily represent a missing concept variant in the visual dataset. For instance, if the above-mentioned models are not well-trained for the identification of a domain's features, then failure to identify them in a given dataset will be mistakenly counted as the dataset weakness, while the problem is in fact an object detection or a classification concern and not a sign of an incomplete data. To minimize such scenarios, we propose to initially train domain-specific object detectors for sensitive feature concepts in domains, for which safety is a primary concern (i.e., safety-critical domains).

Algorithm 2 : Identification of Missing Variants in Data

```
Require: $S_c$ and AIS Dataset to be evaluated.
function EVALUATE(Dataset)
    for each $Image \in \mathcal{D}ataset$ do
        Array $Trn_{img} \leftarrow imgTranslate\ (Image)$;
        Array $Sim_c \leftarrow simMeasure\ (S_c, T_{img})$;
    end for
    $report \leftarrow$ Evaluate $(S_c)$;
return $report$
end function
```

3.2 Semantic Identification: $simMeasure$ (S_c, T_{img})

Once the visual content of a dataset is translated into natural language, the content is comparable to the semantic specifications of the domain, extracted from humans sources of knowledge (output of algorithm 1). This mapping provides an insight about the missing and under-represented variants of domain concepts in the data. For the purpose of comparison, we propose to exploit the semantic similarity between the humans specifications and dataset descriptions as a metric to identify the associations. In addition, measuring semantic similarity compensates for the possible terminology differences in the domain and dataset specifications. The magnitude of the similarity score reflects the level of confidence that a primary variant of a domain concept is present in the dataset.

Please note the extent to which the missing variants of a concept is revealed in this step depends on the relative completeness of the semantic benchmark, automatically created in the first algorithm.

4 IDENTIFICATION OF MISSING VARIANTS BY MODEL (ALGORITHMS 3)

A comprehensive dataset does not guarantee that a model fully captures the primary features of domain concepts. Yet, laying a foundation with a systemically-enriched dataset, while the concepts specifications is referenced, is necessary for AIS success. After semantic enrichment of the dataset, to evaluate AIS comprehension of the concepts variants, the model is relatively evaluated against the established standard. For this, unlike XAI approaches, explaining the model's decision, we adopt a black-box testing method to identify the concept variants which the model fails to recognize.

Regardless of CNN notable success in the recent years, the neural models have long been known as "black-box" [26]. This title is given to the complex neural models since explaining their prediction is difficult for the end users due to the model's complex structure. In image processing, the majority of ML models operate on image features, such as pixel values, which do not correspond to high-level concepts that humans could easily understand. In a safety context in particular, delegating decisions to black-boxes without a clear explanation of the models perception skills, we risk the severe consequences of a possible incorrect recognition.

Table 1: High-level topics of *pedestrian* domain concept.

T.	Pedestrian Topic Variants
a	**Transportation Modes:** passenger, vehicle, subway...
b	**Train Transport:** train, railway, rail, line, subway, station...
c	**Road Types:** alley, bridleway, path, footpath, boulevard...
d	**Accidents:** license, speed, fine, offense, violation, reckless...
e	**Road Types:** freeway, interstate, turnpike, route...
f	**Racial Protest:** constable, arrest, police, protest...
g	**Pedestrians in Customs:** custom, disney, snoopy, mickey...
h	**Auto-Detect Challenges:** detection, sift, ocr, pixel, statue...
i	**Walking Disabilities:** wheelchair, disable, leg, ataxia...
j	**Safety:** safety, nhtsa, barrier, hazard, bumper, guardrail...
k	**Children:** infant, harness, baby, rider, bicycle, stroller...
l	**CarTech Faults:** collision, vehicle, fatality, brake, v2x, v2v...
m	**Background:** building, wall, concrete, apartment, house...
n	**Careless Ped.:** random, walker, jaywalking, magazine...
o	**Campers:** camping, hikers, footpath, trail...

A well-balanced dataset does not provide assurance that a model fairly learns the primary features of concept variants. For example, a model can be biased towards detecting *bicycles* instead of *wheelchairs* as both contain the feature *wheel*. Specifying what a model has learned helps to improve the models interpretation of the mistaken concept variants through re-balancing the dataset, re-configuring the model, or re-training a different-type model. For instance, one could identify the missing variants to augment the training dataset with potentially under-specified concepts which have not been properly learned by the ML model. For example, if a feature is identified in the benchmark (*e.g.* wheelchair) but analyzing the model reveals that the model fails to recognize several associating concept variants (*e.g.* pedestrians in wheelchair), augmenting the dataset with instances of this variant offers the model a chance to better learn the concept variants.

Table 2: High-level topics of *aircraft* domain concept.

T.	Aviation Topic Variants
a	**UAV:** uav, drone, dji, aerial...
b	**Early Aircraft:** flyingboat, glider, kite, floatplane...
c	**Aviation:** aircraft, aerodynamic, spacecraft, balloon...
d	**Aircraft Type:** bomber, fighter, beechcraft, prototype...
e	**Aircraft Mission:** squadron, airfield, fighter, dh60, hornet...
f	**Boeing:** 737s, 737ng, bombardier, airline, ATI...
g	**Aerospace Engineer:** aeronautics, optimization, mechanical, monoplane, navy, superjet...
h	**Airline:** aeroflot, lufthansa, fleet, airport, boeing, airbus...
i	**Flight:** plane, hijack, crash, flight, anxiety, radar, officer...
j	**Pilot:** pilot, flew, aerial, corporate...

Our proposed framework evaluates the relative completeness of models perception of domain concepts and the extent to which, the model was able to capture diversity of the concepts instances during the training. For this, the model is evaluated with the specification of each instance, as is collected in the semantic benchmark. To pass each concept variant to the model as an input, each specification is required to initially be transformed into a set of images which fairly display the concept variant for a given description. Once the textual specification of each variant is converted into visual data, the variant's instances are then fed to the AIS model. The expected response of the model is obviously the classification of the entire variants as the main concept. For instance, once wheelchair appears as a primary feature of handicapped variant of pedestrians, the specification is converted into several visual instances of this variant, and is passed to the model. The model's prediction against the verdict, which here is the prediction of pedestrian, is evaluated for each instance of a pedestrian in wheelchair.

4.1 Image Creation: $imgCreate\ (s)$

The pixel-level format of the AIS-comprehensible input requires non-trivial transformations of the benchmark specifications to visual data (*e.g.* images or video frames). The converted instance needs to fairly reflect each concept variant.

Since the specifications of concept instances are in natural language, any of the commonly-known search engines could be leveraged for the purpose of retrieving variant-relevant images [38]. In addition to the text-based engines which return image results, image search engines are in particular designed for seeking visual data, such as ImageRover [71], WebSeek [75], Diogenes [38], and Atlas WISE [37]. They primarily differ in design and implementation of sub-tasks, such as the methods for data gathering and digestion, indexing and query specification, which will impact their retrieval ability. The appropriate engine could be evaluated according to its image similarity, Web coverage, and performance efficiency [84].

Additionally, several image alignment algorithms are developed, allowing to discover the correspondence relationships among images with varying degrees of overlap. The produced alignments are leveraged by various *image stitching* algorithms to solve the limitation of image/video availability, blending the images in a seamless manner to create new ones [3, 52, 70]. These technologies could provide visual translations of concept variants specifications.

4.2 Model Detection: $modelDetection\ (S_{img})$

Once each specification is converted to pixel-level information, AIS performance in detecting each concept variant demonstrates the extent to which the model learned the given variant during training. When the model fails to recognize a variant in the majority of its given instances, the model is potentially not yet well-trained for that particular variant of the concept. For instance, processing a variety of images and video frames of diverse *pedestrians in wheelchair*, we expect the majority of the classification decisions to be *pedestrian* in a binary classifier. Verifying that the model is able to detect a concept's primary variants, contained in an internet-scale semantic benchmark, may not particularly guarantee the same behavior during the operation (due to the environment-related uncertainties) but it brings us closer to ensure that a concept is well-learned by a model with respect to humans semantic perception.

5 EVALUATION

In this section we conducted four experiments to assess the functionality and usability of our framework.

We selected the **automotive domain** for several reasons. First, the aforementioned domain suffers remarkably from lack of specifications [58, 68, 76, 86]. Second, AIS misperception of surroundings in this domain leads to significant loss of life or damage to properties and environment. Third, this domain is of our and our industrial collaborators' interest, who provide us with domain-specific knowledge, once required. In this domain, we selected *pedestrian* as the targeted domain concept due to its criticality for the primary vision perception tasks in safety-relevant applications, such as in *pedestrian detection.*

We repeated the experiments in the **aviation domain** for the primary concept of *aircraft* to assess the generalizability of the B-AIS. We observed a similar behavior of B-AIS, as well as similar promising results in the both domains.

5.1 RQ_1: Functionality Assessment

To assess the functionality, we initially constructed a semantic benchmark for the domain concept (experiment 1). The automatically-established benchmark is then leveraged to identify the weaknesses present in a dataset (experiment 2) and further the faulty detections by the model (experiment 3).

Algorithm 3 : Model's Black-box Evaluation

```
Require: $S_c$ and AIS Model to be evaluated.
function EVALUATE(Model)
    for each $s \in \mathcal{S}_c$ do
        Array $Img_s \leftarrow imgCreate\ (s)$;
        Array $Vrdct_s \leftarrow modelDetection\ (Img_s)$;
    end for
    $report \leftarrow$ Evaluate $(Vrdct_s)$;
return $report$;
end function
```

5.1.1 ***Experiment 1:*** *Identification of Domain Concepts (Algorithm 1).*

To build a semantic ground truth, we developed a process which initially searches through a large set of knowledge-bases for any term, contextually related to the input, creating an initial domain-specific search query. This process is computationally expensive since a large set of knowledge bases, such as Google n-gram[51] Onelook [53] and, RelatedWords [63] are thoroughly searched for each query. Google n-gram is an online search engine that provides a search for 155 billion words from American English and 34 billion words from British English and provides high-frequency terms associated with a given term as a search query. The RelatedWords is an open-source project that runs several algorithms, such as word embedding, to convert words into multidimensional real-valued vectors representing their meanings. The generated vectors of the words are then mapped in a space of pre-computed vectors according to a set of existing corpora. The similarity of the words is then calculated according to the cosine similarity of the angle between their associating vectors, which is representative of the vectors distance in the space. RelatedWords also uses ConceptNet [77] to retrieve words that have meaningful relationships to our query. Onelook indexes over a thousand online dictionaries and encyclopedias to return the words related to a search query. In addition to dictionaries and encyclopedias, Onelook internally works on Datamuse API to search various data sources.

We implemented a two-phase process to retrieve the most related terms to our initial seed, pedestrian. First, the Google n-gram knowledge base is searched for accompanying and co-occurring terms with each concept. The database will return all terms that more frequently occurred within a given short distance (up to four terms before and after) of the initial term. Yet to identify the related terms that did not appear within our identified range, the RelatedWords and OneLook are searched for semantically related terms to the input. This process resulted in retrieving about 412 pedestrian-related terms, as well as 700 aviation-related terms.

We then applied lemmatization to the retrieved words, resulting in 358 and 518 related terms respectively for pedestrian and aircraft [8]. We decided to use lemmatization rather than stemming since both reduce the inflectional forms of terms, while lemmatization preserves the derivationally related words, such as those starting or ending with un-, dis-, mis-, -ness, -ish, -ism, -ful, and -less. This is accomplished by specifying the words' part-of-speech tags (grammatical roles).

Given the expanded list of domain-specific terms, to further improve the quality of the search, each term was automatically searched in additional sources, possibly including detailed specifications of the concept-related term, such as online dictionaries and documents. For this purpose, two different online encyclopedias, namely Britannica and Wikipedia, were first searched for each term in the extended list.

The Google search engine was secondly utilized for each *term*, being replaced in a search query as "What is the *term*?". The documents related to the first 100 returned links were retrieved for each query. We performed level one web scraping for each document, meaning that we only extracted the textual information and not the additional links within each page. This phase retrieved a large set of documents related to each augmented term.

As we used publicly available services, we faced Google search engine rate limit of 5 requests per 20 minutes and 20-30 requests per minute on Wikipedia and Britannica. Given the 358, and 518 terms linked to pedestrians and aircraft respectively and the rate restriction, the search process took a total of about 47.7 and 24.2 hours, retrieving 51,963 and 26,344 documents for pedestrian and aircraft respectively. The average length of the retrieved documents were 52 lines for pedestrian and 480 lines for aircraft as the entries about aircraft were noticeably more lengthy compared to the pedestrian concept.

Each set of documents is then organized into a meaningful hierarchy of topics. Although topic models, such as LDA [9] and NMF [42], have shown promises for topic modeling, tuning their hyper-parameters is often challenging. For this reason, to identify dominant topics of relevance to the domain concept, we adopted a transformer-based topic modeling technique [25] shown to produce highly cohesive clusters [83]. Table 1 partially represents the output of the first algorithm. Due to space limitations, the rest of topics are available in our repository [1]. Since the validity of the semantic ground truth was the topic of the authors previous works, here we focused on developing methods to leverage the semantic

[1] https://github.com/AI-EnabledSoftwareEngineering-AISE/B-AIS

benchmark for AIS evaluation with the aim of improving their reliability. For this reason we will not discuss the completeness of the benchmark in this document.

5.1.2 ***Experiment 2:*** *Identification of Missing Variants in Dataset (Algorithm 2).* During this experiment, we referred to the specifications of the concept to semantically evaluate the relative completeness of a commonly-used dataset for visual perception tasks. For this assessment, we chose the Wikipedia-based Image Text Dataset (WIT) [2], which consists of 11.5 million image-text samples in 108 languages, containing 37.6 million distinct entities [78].

For evaluation purposes, only entries with English captions, 5,411,978 image-text entries, were selected from the dataset. Further, as our interest is only in the images related to each domain concept, we automatically selected the images in the upper quartile (Q_3), whose *caption* similarity score to the concept was within the highest 25% of the entire set, resulting in 1,355,838 text-image entities.

Each WIT image is accompanied with multiple descriptive textual bodies, providing information about the image and page, from which the image is retrieved. The textual body contains a *reference*, which is the image caption, visible on the wiki page directly below the image; *attribution*, which is the text on the Wikimedia page of the image; *alt. description*, which is not visible in general, but commonly used for accessibility and screen readers [78]; and *page description* containing the first paragraph of the wiki page context.

Table 3: Similarity measures of the topics' specifications to the associating images in the WIT Dataset. The aqua cells represent the *missing* topics in the dataset.

	Pedestrian Concept					Ground Truth			
T.	#	μ	σ	Q_3	max	μ	σ	Q_3	max
a	544	.51	.04	.54	.67	.19	.16	.30	.78
b	680	.50	.06	.54	.74	.17	.16	.30	.70
c	544	.52	.03	.53	.68	.15	.19	.22	1.00
d	544	.51	.07	.56	.76	.22	.24	.36	1.00
e	815	.53	.04	.55	.66	.21	.20	.30	.88
f	407	.48	.08	.53	.64	.19	.18	.30	.75
g	272	**.45**	.48	.57	.78	**.04**	.00	.10	.33
h	408	**.42**	.09	.53	.66	**.12**	.11	.20	.70
i	407	.48	.11	.60	.71	.16	.17	.29	.86
j	136	.47	.04	.49	.67	.16	.17	.29	.86
k	272	.45	.04	.46	.58	.21	.20	.33	.89
l	408	**.37**	.34	.36	.51	**.07**	.11	.10	.60
m	407	.48	.05	.52	.63	.19	.19	.30	.80
n	408	**.43**	.05	.46	.61	**.08**	.12	.14	.57
o	544	.52	.05	.56	.71	.14	.17	.25	.88

According to the authors' evaluation of this information for training zero-shot learning models [78], we decided to concatenate and consider the reference and attribution descriptions of each image as a description for each image, referring to it as *'caption'* in this document. The captions were then searched for the relevance of the image to each variance of the domain concept in the benchmark. The page descriptions, containing more detailed information were reserved for the evaluation of the dataset missing topics, which B-AIS detected. In this paper, we refer to the descriptions as *'context'*.

Table 4: Similarity measures of the topics' specifications to the associating images in the WIT Dataset. The aqua cells represent the *missing* topics in the dataset.

	Aircraft Concept					Ground Truth			
T.	#	μ	σ	Q_3	max	μ	σ	Q_3	max
a	136	.41	.08	.46	.62	**.12**	.14	.20	.50
b	264	.42	.08	.47	.64	.31	.20	.43	.86
c	546	.44	.11	.50	.71	.18	.16	.25	.88
d	720	.41	.08	.47	.62	.24	.19	.40	.80
e	707	.40	.07	.44	.64	.30	.23	.44	.89
f	451	.41	.06	.45	.59	.20	.15	.27	.67
g	518	.43	.10	.49	.69	**.11**	.11	.18	.50
h	952	**.37**	.08	.43	.61	**.20**	.14	.30	.88
i	927	**.37**	.10	.44	.69	.24	.21	.40	.90
j	395	.42	.09	.48	.68	.21	.19	.33	.78

To measure semantic similarity, we initially evaluated and compared three state-of-the-art universal sentence encoders, namely SBERT [62], USE [13], and InferSent [15] on 1% of our data (54,119 images). While the same data was fed to the three models, their performance significantly differed. The models nearly generated similar scores, however the execution time was significantly shorter for SBERT. We believe this occurred, since SBERT executes both the query and corpus embedding processes on the same GPU.

SBERT is a pre-trained network, deriving embedding vectors of given sentences, to identify semantically-relevant sentences by comparing the cosine similarity of between-vector angles. Adopting SBERT, the image captions, as well as the specifications of each topic (from the previous experiment) were embedded in a single vector space and the cosine similarities were calculated between the embedding of the specification topics and captions.

Given the image-topic similarity scores, for each topic we selected the images with scores within the highest percentile (99.99%). The count of the remaining images, as well as the mean, standard deviation, Q_3, and maximum similarity are respectively shown under the first tab of Tables 3 and 4, *'Pedestrian or Aircraft Concept'*.

Finally, we flagged a variant as not-covered in the dataset, if the average similarity of the associated images was lower than the average of the entire samples in the dataset (0.47 for pedestrian and 0.40 for aircraft) as shown in bold font and aqua background respectively in Tables 3 and 4.

Ground Truth: The B-AIS output here is therefore, the missing variances of the concepts in the selected dataset. For the evaluation purposes and since manually going through the images was not feasible, we decided to build an approximate ground truth for the topics, missing and present in the dataset.

A script was implemented so that for each topic, the entire topic's associated terms were individually searched in the context (the page descriptions) of pedestrian- and aircraft-related images in the dataset. As such, we measured the appearance frequency of topics relevant terms in the image descriptions. This showed whether or not the comprehensive description of an image mentions any of the relevant terms to the topic or to the augmented descriptions of the topic variants. We marked the topics whose associated terms on average appeared less frequently than the average of the entire terms (0.153 and 0.211 respectively for pedestrian and aircraft),

[2] https://github.com/google-research-datasets/wit

Table 5: Similarity of the *'Pedestrian'* topics to the images, retrieved for AIS semantic black-box evaluation.

	Caption Similarity					Image Coverage			
T.	#	μ	σ	Q_1	Q_3	μ	σ	Q_1	Q_3
a	38	.40	.11	.34	.51	.60	.15	.50	.71
b	36	.40	.12	.36	.49	.66	.22	.68	.71
c	5	.45	.05	.42	.45	.80	.14	.80	.80
d	30	.50	.14	.40	.58	.71	.22	.62	.85
e	50	.43	.12	.36	.53	.75	.21	.57	1.0
f	21	.43	.08	.39	.49	.81	.25	.64	1.0
g	3	.20	.17	.10	.25	.57	.06	.55	.60
h	0	0	0	0	0	0	0	0	0
i	22	.32	.17	.18	.38	.66	.43	.30	.96
j	0	0	0	0	0	0	0	0	0
k	19	.39	.13	.33	.47	.91	.11	.80	1.0
l	10	.45	.02	.44	.46	.57	.20	.50	.64
m	26	.33	.10	.30	.42	.59	.24	.38	.70
n	11	.36	.13	.31	.46	.49	.02	.50	.50
o	40	.43	.12	.37	.51	.73	.17	.58	.83

as not fully covered. The remaining variants were considered as relatively covered in the dataset with respect to the semantically-relevant concept variants. The mean, standard deviation, Q_3, and maximum similarity of each topic are respectively reported under the second tab of Tables 3 and 4, *'Ground Truth'*. The first column contains corresponding topic ids from Tables 1 and 2.

Table 6: Similarity of the *'Aircraft'* topics to the images, retrieved for AIS semantic black-box evaluation.

	Caption Similarity					Image Coverage			
T.	#	μ	σ	Q_1	Q_3	μ	σ	Q_1	Q_3
a	7	.46	.07	.45	.50	.49	.31	.20	.75
b	13	.54	.05	.50	.57	.45	.21	.43	.43
c	31	.52	.07	.47	.58	.29	.26	.12	.33
d	35	.54	.07	.50	.71	.43	.20	.30	.58
e	27	.61	.0.7	.59	.64	.51	.28	.29	.67
f	23	.54	.08	.47	.59	.47	.21	.33	.55
g	25	.49	.05	.45	.54	.34	.21	.18	.36
h	47	.56	.08	.52	.60	.47	.19	.30	.60
i	46	.51	.08	.46	.56	.36	.26	.12	.60
j	19	.51	.08	.43	.58	.45	.30	.22	.67

Results Discussion: As shown in Table 3, B-AIS was able to semantically identify five out of the six missing topics in pedestrian dataset and one out of three in aircraft dataset relative to the established benchmarks respectively with 100% recall and precision of 80% (95% F_2) and 33% recall and precision of 50% (45% F_2).

F-score measures the accuracy of a test using precision and recall, while favors recall over precision. Since F_2-measure puts more attention on minimizing false negatives than minimizing false positives, reporting F_2 seems more relevant in this research. B-AIS additionally recognized *children* variant of *pedestrian* (topic i) as potentially missing in the dataset. Yet searching the description of the images showed the children-related terms occurred with a frequency of 0.21 which was lower than the average, and therefore was considered as not-sufficiently-covered by B-AIS.

Note that for the purpose of the experiments and the safety-related topics of the selected domain, we implemented B-AIS to be pessimistic, as we chose to only consider images within the highest similarity percentile (99.99%) to be related. Further, we decided to consider all concept variants to be of the same importance, and therefore, we assigned the same threshold to all the concept variants. This means, we accepted the risk of false positives, over the false negatives, appropriate for this domain. However, depending on the criticality and importance of each variant, one could decide to be more conservative (select higher threshold) only for specific types, such as children (topic k) or pedestrians with walking disabilities (topic i), and less concerned with other variants, such as pedestrians on the sidewalk (topic c).

Table 7: Statistics of model (OFA) black-box testing of *Pedestrian* variants specifications, before & after re-training.

	Before Training			After Training	
Top.	C-Ped.	C-Top.	*GT*	C-Ped.	C-Top.
a	0.29	0.39	0.63	0.32	0.41
b	**0.26**	0.35	**0.86**	**0.28**	0.39
c	0.37	0.36	**1.0**	0.49	0.46
d	0.41	0.42	0.7	0.52	0.52
e	0.33	0.36	0.66	0.44	0.45
f	**0.26**	0.36	**1.0**	**0.32**	0.42
g	**0.24**	**0.1**	**1.0**	**0.31**	0.13
h	-	-	-	-	-
i	**0.19**	**0.31**	**0.91**	**0.31**	0.4
j	-	-	-	-	-
k	0.28	**0.31**	**0.95**	0.25	0.41
l	**0.26**	0.35	0.6	0.51	0.37
m	**0.26**	0.45	**0.92**	**0.35**	0.47
n	0.4	**0.30**	**1.0**	0.35	0.41
o	0.27	**0.32**	0.75	0.34	0.38
Avrg.	**0.27**	**0.33**	**0.84**	**0.36**	**0.40**

The rest of the columns respectively report the mean, standard deviation, upper quartiles, and the maximum frequency of the entire words relevant to the topic in column one.

Table 8: Statistics of model (OFA) black-box testing of *Aircraft* variants specifications, before & after re-training.

	Before Training			After Training	
Top.	C-Air.	C-Top.	*GT*	C-Air.	C-Top.
a	0.43	0.29	**0.86**	0.49	0.31
b	**0.32**	0.42	0.38	0.38	0.41
c	**0.40**	**0.25**	**0.81**	**0.50**	**0.33**
d	0.41	**0.27**	**0.83**	**0.49**	0.41
e	0.46	**0.27**	**0.85**	**0.48**	0.34
f	0.44	**0.25**	**0.91**	**0.42**	0.31
g	0.46	**0.23**	**0.84**	**0.53**	**0.28**
h	0.46	0.31	**0.83**	0.46	0.36
i	**0.34**	**0.25**	0.72	0.38	**0.35**
j	**0.29**	**0.23**	0.58	0.36	0.30
Avrg.	**0.41**	**0.28**	**0.78**	**0.45**	**0.34**

5.1.3 ***Experiment 3:*** *Identification of Missing Variants by Model (Algorithm 3).* Referring to the semantic variants, we selected a pre-trained model to identify the variances of the domain concepts, not being recognized by the model. Here, we treated the model as a blackbox, passing each topic variant as an input, observing and comparing the model's response to the expected result.

Due to the pixel-level perception of the model, we required to convert each topic variant to the corresponding visual data. For this purpose, we wrote a script to repeatedly query the Google search engine for each variant, load the returned image page, retrieve the image, scrape the image caption, as well as the first paragraph of the loaded page as the context. The extracted body of text was reserved to verify the relevancy of the retrieved images from the web to the topic. For this, we applied SBERT to measure the similarity of the extracted images descriptions to the corresponding topic and removed those with lower similarity than the average of all samples. The first tab in Tables 5 and 6 respectively represents the final count of the remaining images, the mean of the similarity to the topic, standard deviation, Q_1 and Q_3 for each topic. Moreover, we manually verified the relevancy and removed a couple more images, which we found irrelevant to the topic. The second column of the table is representative of this final count. The second part of the table shows our manual evaluation of each topic's set of images. We scored each image based on the number of the relevant terms contained. Then we measured the total score of the images in each topic and reported the statistics of this similarity in the second part of Tables 5 and 6. As shown, this process resulted in removing all retrieved images for the two pedestrian topics h and j, namely 'pedestrian-related auto-detection challenges', and 'pedestrian safety'.

Later, we adopted a model with image captioning capability, called one-for-all (OFA) [90]. OFA uses a sequence-to-sequence learning framework which combines different modalities of data, here vision and language, for the purpose of performing several vision tasks, such as image generation and captioning. We selected this model, since the WIT was not among the OFA training datasets, as we wanted to evaluate the model, independent of the dataset in the previous experiment.

Further, each set of images was passed to OFA individually and the model-generated captions were then semantically compared, once to the domain concepts pedestrian and aircraft, and once again, to the topic relevant terms. The similarity scores of captions to the domain concept is reported in the second column (C-Ped. ans C-Air.) and to the relevant terms (the concept variant) in the third column (C-Top) of Tables 7 and 8.

In this experiment we only refer to the first section of the table under 'Before Training' tab. The rest of the table will be discussed in the following experiment. Finally, the topics, which similarity measures to the concept or to the variant were lower than the average of the entire set (0.271 & 0.336 for pedestrian and 0.41 & 0.27 for aircraft) were flagged as model's potentially-missed variants.

Ground Truth: To evaluate B-AIS detection, two independent researches, not involved in this project, manually verified each instance of each variant. One point was added for the presence of a pedestrian in each image. The scores were further normalized for each topic, shown in the fourth column of Tables 7 and 8 (GT). Note that this column has a negative association with the second and third columns. This means a higher GT score represents the presence of more pedestrians in the images associated with the topic and therefore, it is expected that the similarity score of the model's output to be higher in the second column. Hence, we marked the topics with higher scores than the average (0.844) in the ground truth. Topics with lower GT scores are the topics, for which the collected images were not representative enough.

Results Discussion: As shown, B-AIS was able to detect seven out of eight missing topics by the model (87% recall), missing one topic, c about the road types. In addition, the two topics of car-pedestrian accidents (topic l) and camper pedestrians (topic o) were marked as not sufficiently covered (F_2 measure of 85%). Similarly, for the aircraft concept B-AIS was able to detect five out of seven missing topics with a F_2 measure of 72%, while false positives include topics b, i, and j.

Note that to be persistent in the experiments, while building the ground truth, we selected the topics above the average coverage to be labeled as covered, yet as mentioned earlier, we implemented B-AIS to be pessimistic for the selection of covered topics. As such topics l with 60% coverage, manually assessed, is marked as covered in the ground truth, while B-AIS marks this topic with as not-properly-covered. This inconsistency is accepted since the attempt here is to minimize the risk of false negatives, while keeping the false positives below an overwhelmingly large number.

5.2 RQ_2: Usability Assessment

To assess B-AIS usefulness, we re-trained OFA with the images of the topics which were initially missed by the model, and re-assessed its performance on the same images in the previous experiment.

5.2.1 ***Experiment 4:*** *Model Re-training with Missing Topics.* The B-AIS selected topics, for which OFA failed to generate a similar-enough caption to the pedestrian (marked in the second column of Table 7) were selected, including b, f, g, i, and m. Referring to the WIT dataset, we retrieved the set of relevant images to each topic, labeled by B-AIS in the second experiment. Further %75 of the images were randomly selected for stage-one training, %15 for stage-two training purposes. The process of retraining OFA-base model with the total 1,502 images for 10 and 5 epochs in stage-one and two of the training process took nearly two hours, using 2 GPUs of 16GB Tesla P100-PCIE.

Once OFA was re-trained with the instances of the variants, missed originally by the model, for the second time we measured the model's perception of the same instances used in the previous experiment. Finally, we re-measured the performance of the model referring to the same ground truth for the comparison purposes. The second tab of Tables 7 and 8, 'After Training', provide the similarity scores of the generated captions by the re-trained model to the concept pedestrian (columns C-Ped. and C-Air), as well as the relevant terms of the concept semantic variant (column C-Top).

Results Discussion: As shown in bold font in the fifth column of Table 7, the similarity scores of the missing topics were improved by a total of 36%. The average similarity of the model-generated captions to the pedestrian concept is as well increased from the average of 0.27 for the entire topics to 0.36, and to the associated topic is improved from 0.33 to 0.40. In case of the aircraft concept, the average similarity changed from 0.41 to 0.45 and from 0.28 to 0.34 respectively for the concept and the concept's topics.

The images 2, 3, and 4 in the automotive, and 5, 6, and 7 in the aviation domains illustrate examples of the model's improved perception in the form of generated captions before and after the retraining process. As captioned below each image, the first sentence is the initial OFA-generated caption, while *'After'* shows the generated caption by the same model after we retrained it for B-AIS-identified missing topics.

5.2.2 Discussion of Cost: The online knowledge discovery, including processing encyclopedias and dictionaries and retrieving the relevant context, largely relies on the performance of the API, adopted for the process. For instance, a commercial API may potentially facilitate the process in comparison to open source APIs. The computational cost of the data cleaning, processing, and topic modeling are influenced by the size of the retrieved documents. For instance, the process retrieved 51,963 documents for the concept of pedestrian and 26,344 for aircraft. As such, the topics were generated within about one hour, using a system with an Intel(R) Xeon(R) CPU E5-2687W v2 @ 3.40GHz and 500GB RAM.

The image processing is computationally more expensive than processing text, and largely benefits from the use of GPU's power. The cost thus directly depends on the quantity and quality of the GPUs or the other alternative, solutions, such as CPUs, or central processing units. Since GPU peak performance and memory bandwidth are often significantly higher than CPUs, GPU image processing in general yields significantly lower computational cost. For instance, in our experiments with a single Tesla P100-PCIE 16GB GPU, calculating semantic similarity between 5,411,978 captions and topics pairs took roughly about two hours for each concept of pedestrian and aircraft. Additionally, fine-tuning the OFA model took about five hours and twenty minutes for the airplane and four hours for the concept pedestrian.

6 RELATED WORK

6.1 Domain Specifications

6.1.1 semantic textual similarity. A previous research provided two methods for generating sentence embeddings which showed a promising transfer of textual data to a variety of different NLP tasks [13]. The authors demonstrated that sentence embeddings can be utilized to achieve accurate results on semantic textual similarity tasks with a surprisingly small amount of task-specific training data. They used unsupervised data from a variety of web sources, including Wikipedia, web news, web question-and-answer pages, and discussion forums, to train their model. Later supervised data is adopted from the Stanford Natural Language Inference (SNLI) corpus [10] to augment the unsupervised learning.

Although BERT [18] received attention for sentence-pair regression tasks, such as semantic textual similarity, the model requires both phrases to be supplied, which results in a processing overhead. Sentence-BERT (SBERT), a pre-trained variant of BERT, employs triplet network architectures to derive semantically relevant sentence embeddings to be compared using cosine-similarity [62].

6.2 Multimodal Deep Neural Networks

Several recent studies have developed models enabled to process artifacts of texual and visual format at the same time.

One research implemented a multi-modal two-stream model, namely ViLBERT, which processes both visual and textual inputs in separate streams. The streams still interact with each other via additional layers called co-attention transformer layers [46]. Another research group introduced a similar model, called VisualBERT, containing a stack of transformer layers that implicitly match bits of an input text and regions in a corresponding input image with self-attention. The attention weights are then used to align words and image regions internally [43].

Despite the success of ViLBERT and VisualBERT in vision tasks, the models are yet required to be distinctly pre-trained for each specific task. Another Vision-Language pre-training (VLP) model, is presented by a different research group, which is a unified encoder-decoder that can be fine-tuned for both caption production and visual question answering (VQA) tasks [100]. VLP surpass BERT-based model as it learns a more universal contextualized vision-language representation by integrating encoder and decoder, allowing models to be fine-tuned for both generation and understanding tasks.

While VLP trains a single model for both generation and comprehension, there are problems with text-to-object semantic alignments in the model. A research group presented a framework, called OSCAR, which employed object tags recognized in photos as an anchor point to considerably simplify learning the alignments [44, 98]. Another approach for transformer-based text-to-image generation is named DALL-E, which auto-regressively model the text and image tokens as a single stream of data [60].

The model uses a two-stage training process, which first trains a discrete variational auto-encoder to compress an image into the image tokens, each element of which is then converted into vector values. Finally CLIP, is a multimodal model which is trained on a large generic dataset, allowing to skip re-training every time the model is applied for a different vision-related task [56].

6.3 eXplainable Artificial Intelligence (XAI)

As AI is used more in people everyday life, as well as safety-critical domains, the AI-based software is faced with a set of regulations [24], known as *right of explanation*, giving individuals the right to obtain an explanation of the inference of automatically produced by a model. The attempt of XAI is to have AIS generate rules which are highly interpretable by humans, meantime saving the accuracy of trained models. The research in the area could be categorized to (1) describing *explainability* notation (2) reviews on explainability methods (3) developing new methods for explainability and (4) evaluating methods explainability [85].

There are several methods adopting graphical representation to explain the model decision on the image set, such as heat-maps [66, 79] and combination of visual and textual explanation, where the CNNs are leveraged for object detection and LSTM for the generation of captions caption[93]. One research work attempted to improve visual explanation by the use of discriminative loss and relevance loss, improving the class and image relevancy of explanations, respectively [20].

Other works go further in improving visual explanations. For instance, one research developed a phrase-critic model which takes in an image and a candidate explanation as input and attempts to determine whether the candidate sentence is relevant to the given image. The model initially grounds the explanation objects in the given image, to further search for a candidate explanation which closely resembles the image [30].

The authors of [32] developed a joint approach, which explicitly model the compositional linguistic structure of referential expressions as well as their end-to-end visual grounding. As it is impossible to achieve a complete transparency in current neural

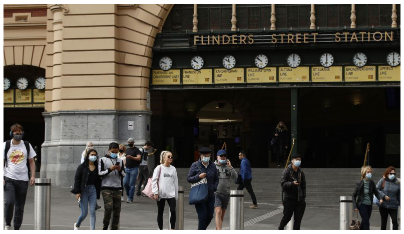

Figure 2: a group of people walking in front of train station. After: people wearing face masks walk in front of train station in London. Topic B: pedestrian, rail, railway, train...

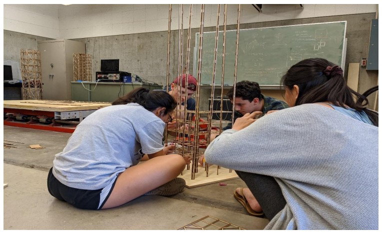

Figure 3: a group of people looking at a man in cage. After: A students working on the scaffolding for the installation. Topic M: construction, pedestrian, structure, building...

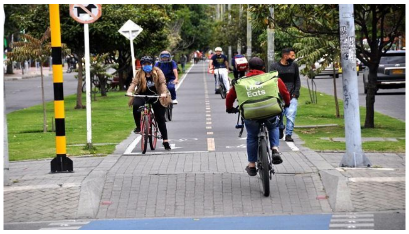

Figure 4: a group of people riding bikes down a street. After: cyclists and pedestrians wearing face masks ride on a bike path in a park. Topic K: bike, cycling, pedestrian, rider...

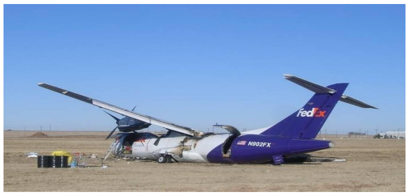

Figure 5: a plane that is on the ground in a field. After: A flight 757 crashed into the ground in the aftermath of the crash.Topic I: aircraft, crash, wreckage, investigator...

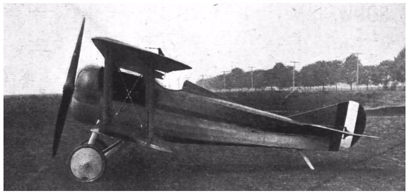

Figure 6: a small plane sitting on the ground in a field. After: A biplane at the airport in 1908. Topic D: aircraft, biplane, raf, wing...

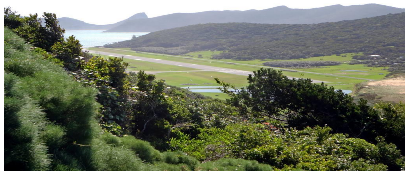

Figure 7: a view of the ocean from the top of a hill. After: A view of the airport from the top of the hill. Topic H: aircraft, airport, airway, flight...

Figure 8: Examples of the model-generated captions before and after re-training with missing topics.

networks [17], it is important to evaluate the models against new and unseen observations. Our approach evaluates the output of perception models against a set of domain knowledge. From point of the XAI our framework keep trained models unchanged and try to evaluate the model behaviour in certain cases at test time [54].

7 THREATS TO VALIDITY

The topics, identified as missing in the dataset could be identified due to inability of the applied neural network to precisely map embeddings, identify the similarity between the embeddings, and establish a relation between the specifications and image captions. We aimed to minimize this threat through adopting a well-trained and state-of-the-art transformer for the identification of the missing topics. The small error in the evaluation results, verified that the error could be negligible. For future work, we will use more advanced image processing techniques, such as scene graph generation and region captioning [45], to extract more information from the visual datasets to be compared against our automotive domain benchmarks. A threat to the construct validity may arise from the evaluations with a single dataset and model. However, since the size of the widely-used datasets is often significantly large, the computational cost of running image processing is extensive. For the future work, we will extend the experiments to several dataset and model commonly used by the community. A threat to the external validity is carrying out the experiments only in two domain (automotive and aviation). We designed a generalizable process, and implemented a general framework, and referred to general knowledge sources. Therefore, no limitation is foreseen for the extension of the application domain. However, due to the expensive computations (in terms of computational resources), we limited the application to the selected concepts for which the accuracy of visual perception tasks is particularly important to achieve functional reliability. In addition, due to the environmental uncertainties, such as noise, in the operational domain, the robustness of the approach requires further investigations.

8 CONCLUSION

We presented an automated process, implemented in a framework called B-AIS, for the evaluation of visual perception of AI-enable software systems (AIS). This generic process first automatically builds a set of semantic variations of primary visual concepts of a domain. Later, refers to the semantically delineated concept variants as a benchmark to measure the coverage of each concept variant, once in the AIS training visual dataset, and once again in the trained AIS perception while treating the model as a black-box.

We evaluated B-AIS in the domains of automotive and aviation for visual detection of pedestrian and aircraft variants. The results showed that B-AIS identified the missing variants of the concept pedestrian and aircraft in the dataset with a F_2 measure of 95% and 45% respectively. As future work, we tend to extend the evaluations to other domains, remove cascading error of the pipeline process, improve the process to be adopted for dataset augmentation purposes, and generate potential failure reports for the system.

ACKNOWLEDGMENTS

This research is partially funded by NSF: CCF: 124606 and partially used resources of ddiLab Laboratory.

REFERENCES

[1] Hala Alrumaih, Abdulrahman Mirza, and Hessah Alsalamah. 2020. Domain ontology for requirements classification in requirements engineering context. *IEEE Access* 8 (2020), 89899–89908.

[2] Grigoris Antoniou and Frank Van Harmelen. 2004. *A semantic web primer.* MIT press.

[3] Oron Ashual, Shelly Sheynin, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. KNN-Diffusion: Image Generation via Large-Scale Retrieval. *arXiv preprint arXiv:2204.02849* (2022).

[4] Michele Banko and Robert C Moore. 2004. Part-of-speech tagging in context. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics.* 556–561.

[5] Hamed Barzamini, Mona Rahimi, Murteza Shahzad, and Hamed Alhoori. 2022. Improving Generalizability of ML-enabled Software through Domain Specification. In *2022 IEEE/ACM 1st International Conference on AI Engineering–Software Engineering for AI (CAIN).* IEEE, 181–192.

[6] Hamed Barzamini, Murtuza Shahzad, Hamed Alhoori, and Mona Rahimi. 2022. A multi-level semantic web for hard-to-specify domain concept, Pedestrian, in ML-based software. *Requirements Engineering* (2022), 1–22.

[7] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[10] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[11] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu Gavrila. 2019. EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* (Feb. 2019).

[12] VerÃ³nica CastaÃ±eda, Luciana Ballejos, Ma Laura Caliusco, and Ma Rosa Galli. 2010. The use of ontologies in requirements engineering. *Global journal of research in engineering* 10, 6 (2010).

[13] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[14] Gobinda G Chowdhury. 2003. Natural language processing. *Annual review of information science and technology* 37, 1 (2003), 51–89.

[15] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).

[16] Ido Dagan, Lillian Lee, and Fernando CN Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine learning* 34, 1 (1999), 43–69.

[17] Hoa Khanh Dam, Truyen Tran, and Aditya Ghose. 2018. Explainable software analytics. In *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results.* 53–56.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[19] P Dollar, C Wojek, B Schiele, and P Perona. 2009. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition.* ieeexplore.ieee.org, 304–311.

[20] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2625–2634.

[21] Dejing Dou, Hao Wang, and Haishan Liu. 2015. Semantic data mining: A survey of ontology-based approaches. In *Proceedings of the 2015 IEEE 9th international conference on semantic computing (IEEE ICSC 2015).* IEEE, 244–251.

[22] Mahak Gambhir and Vishal Gupta. 2017. Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review* 47, 1 (2017), 1–66.

[23] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research* 32, 11 (2013), 1231–1237.

[24] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[25] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[26] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2019), 93.

[27] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. 2020. Generalizable Pedestrian Detection: The Elephant In The Room. (March 2020). arXiv:2003.08799 [cs.CV]

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[29] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European conference on computer vision.* Springer, 3–19.

[30] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV).* 264–279.

[31] Dichao Hu. 2019. An introductory survey on attention mechanisms in NLP problems. In *Proceedings of SAI Intelligent Systems Conference.* Springer, 432–448.

[32] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1115–1124.

[33] Gao Huang, Zhuang Liu, Geoff Pleiss, Laurens Van Der Maaten, and Kilian Weinberger. 2019. Convolutional networks with dense connectivity. *IEEE transactions on pattern analysis and machine intelligence* (2019).

[34] Christian Kaestner. 2020. Machine Learning is Requirements Engineering — On the Role of Bugs, Verification, and Validation in Machine Learning. https://medium.com/analytics-vidhya/machine-learning-is-requirements-engineering-8957aee55ef4.

[35] Haruhiko Kaiya and Motoshi Saeki. 2006. Using domain ontology as domain knowledge for requirements elicitation. In *14th IEEE International Requirements Engineering Conference (RE'06).* IEEE, 189–198.

[36] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence* 24, 7 (2002), 881–892.

[37] ML Kherfi, D Ziou, and A Bernardi. 2003. Atlas WISE: A Web-based image retrieval engine. In *Proceedings of the International Conference on Image and Signal Processing.* 69–77.

[38] Mohammed Lamine Kherfi, Djemel Ziou, and Alan Bernardi. 2004. Image retrieval from the world wide web: Issues, techniques, and systems. *ACM Computing Surveys (Csur)* 36, 1 (2004), 35–67.

[39] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).

[40] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012), 1097–1105.

[41] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence* 35, 12 (2013), 2891–2903.

[42] Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 13 (2000).

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[44] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision.* Springer, 121–137.

[45] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE international conference on computer vision.* 1261–1270.

[46] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[47] Alexander Maedche and Steffen Staab. 2001. Ontology learning for the semantic web. *IEEE Intelligent systems* 16, 2 (2001), 72–79.

[48] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63.

[49] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[50] Sina Mohseni, Mandar Pitale, Vasu Singh, and Zhangyang Wang. 2019. Practical solutions for machine learning safety in autonomous vehicles. *arXiv preprint arXiv:1912.09630* (2019).

[51] Ngrams 2022. *Google Books n-grams*. Retrieved January 8, 2022 from https://www.english-corpora.org/googlebooks

[52] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021).

[53] onelook 2022. *Onelook Dictionary Search*. Retrieved January 8, 2022 from https://www.onelook.com/

[54] Andrés Páez. 2019. The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines* 29, 3 (2019), 441–459.

[55] Maria MP Petrou and Costas Petrou. 2010. *Image processing: the fundamentals*. John Wiley & Sons.

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[57] Mona Rahimi, Jin LC Guo, Sahar Kokaly, and Marsha Chechik. 2019. Toward Requirements Specification for Machine-Learned Components. In *2019 IEEE 27th International Requirements Engineering Conference Workshops (REW)*. IEEE, 241–244.

[58] Mona Rahimi, Jin L.C. Guo, Sahar Kokaly, and Marsha Chechik. 2019. Toward Requirements Specification for Machine-Learned Components. In *Proceedings of the 24th international conference on requirements engineering*. IEEE, porceeding.

[59] Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. 2012. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, Vol. 4. 1.

[60] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[61] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2018. It's Not All About Size: On the Role of Data Properties in Pedestrian Detection. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[62] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[63] relatedwords 2022. *Related Words*. Retrieved January 8, 2022 from https://www.relatedwords.org/

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[65] Mosab Rezaei, Elhamossadat Ravanbakhsh, Ehsan Namjoo, and Mohammad Haghighat. 2019. Assessing the effect of image quality on ssd and faster r-cnn networks for face detection. In *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 1589–1594.

[66] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[67] Rick Salay and Krzysztof Czarnecki. 2018. Using Machine Learning Safely in Automotive Software: An Assessment and Adaption of Software Process Requirements in ISO 26262. *ArXiv* abs/1808.01614 (2018).

[68] Rick Salay and Krzysztof Czarnecki. 2019. Improving ML Safety with Partial Specifications. In *Computer Safety, Reliability, and Security*, Alexander Romanovsky, Elena Troubitsyna, Ilir Gashi, Erwin Schoitsch, and Friedemann Bitsch (Eds.). Springer International Publishing, Cham, 288–300.

[69] Rick Salay and Czarnecki Krzysztof. 2018. Using machine learning safely in automotive software: An assessment and adaption of software process requirements in ISO 26262. *arXiv preprint arXiv:1808.01614* (2018).

[70] Axel Sauer, Katja Schwarz, and Andreas Geiger. 2022. Stylegan-xl: Scaling stylegan to large diverse datasets. *arXiv preprint arXiv:2202.00273* (2022).

[71] Stan Sclaroff, Leonid Taycher, and Marco La Cascia. 1997. Imagerover: A content-based image browser for the world wide web. In *1997 Proceedings IEEE workshop on content-based access of image and video libraries*. IEEE, 2–9.

[72] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. 2006. The semantic web revisited. *IEEE intelligent systems* 21, 3 (2006), 96–101.

[73] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.

[74] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[75] John R Smith and Shih-Fu Chang. 1997. Visually searching the web for content. *IEEE multimedia* 4, 3 (1997), 12–20.

[76] Bernd Spanfelner, Detlev Richter, Susanne Ebel, Ulf Wilhelm, Wolfgang Branz, and Carsten Patz. 2012. Challenges in applying the ISO 26262 for driver assistance systems. *Tagung Fahrerassistenz, München* 15, 16 (2012), 2012.

[77] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. , 4444–4451 pages. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[78] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2443–2449.

[79] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. 2017. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 667–676.

[80] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 3716–3725.

[81] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6619–6628.

[82] Oguzhan Tas and Farzad Kiyani. 2007. A survey automatic text summarization. *PressAcademia Procedia* 5, 1 (2007), 205–213.

[83] Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626* (2020).

[84] Remco C Veltkamp and Mirela Tanase. 2000. Content-based image retrieval systems: A survey. (2000).

[85] Giulia Vilone and Luca Longo. 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093* (2020).

[86] Andreas Vogelsang and Markus Borg. 2019. Requirements Engineering for Machine Learning: Perspectives from Data Scientists. *arXiv preprint arXiv:1908.04674* (2019).

[87] Atro Voutilainen. 2003. Part-of-speech tagging. *The Oxford handbook of computational linguistics* (2003), 219–232.

[88] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al. 2001. Constrained k-means clustering with background knowledge. In *Icml*, Vol. 1. 577–584.

[89] Hanna M Wallach. 2006. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*. 977–984.

[90] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. *arXiv preprint arXiv:2202.03052* (2022).

[91] Yulin Wang, Xuran Pan, Shiji Song, Hong Zhang, Gao Huang, and Cheng Wu. 2019. Implicit semantic data augmentation for deep networks. *Advances in Neural Information Processing Systems* 32 (2019).

[92] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5410–5419.

[93] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. PMLR, 2048–2057.

[94] Pengfei Xu, Xiaojun Chang, Ling Guo, Po-Yao Huang, Xiaojiang Chen, and Alexander G Hauptmann. 2020. A survey of scene graph: Generation and application. *IEEE Trans. Neural Netw. Learn. Syst* (2020).

[95] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[96] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. 2018. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[97] Ke Zhai, Jordan Boyd-Graber, Nima Asadi, and Mohamad L. Alkhouja. 2012. Mr. LDA: A Flexible Large Scale Topic Modeling Package Using Variational Inference in MapReduce. In *Proceedings of the 21st International Conference on World Wide Web* (Lyon, France) *(WWW '12)*. ACM, New York, NY, USA, 879–888. https://doi.org/10.1145/2187836.2187955

[98] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[99] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. openaccess.thecvf.com, 3213–3221.

[100] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.