Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics[*]

Peratham Wiriyathammabhum
email: peratham@cs.umd.edu

This scholarly paper is submitted
in partial fulfillment of the requirements for
the degree of Master of Science in Computer Science.

Keywords. language and vision, survey, multimedia, robotics, symbol grounding, distributional semantics, computer vision, natural language processing, visual attribute, image captioning, imitation learning, word2vec, word embedding, image embedding, semantic parsing, lexical semantics.

1 Introduction

Languages are common tools to describe the world for human-human communication. There are many forms of languages which may be verbal or nonverbal but all are assistants for understanding. Some examples are texts, gestures, sign languages, face expressions, etc. Languages provide meaning and meaning is grounded in human perception of the world. This is usually referred to as the **symbol grounding problem** [86]. If it is a language without perception, it is a fantasy which is not based on the real world. If it is a pure perception without language, there is no movement and retention of any object or knowledge by the mind.

In human perception, visual information is the dominant modality for acquiring knowledge since a big part of the human brain is dedicated to visual processing. Whether or not there are languages involved in the visual process is still an ongoing argument. However, for an intelligent system that tries to achieve AI, having languages provides interpretability and creates a way for human-machine interaction which gives rises to a lot of interesting applications. To bridge language and vision, we first revisit the major tasks in both language and vision.

1.1 Computer Vision tasks and their relationships to Natural Language Processing

Computer Vision (CV) tasks can be summarized into **the concept of 3Rs** [119] which are **Reconstruction**, **Recognition** and **Reorganization**. Reconstruction involves a

[*] This paper is done under the supervision of Prof. Dr. Yiannis Aloimonos and Dr. Cornelia Fermüller.

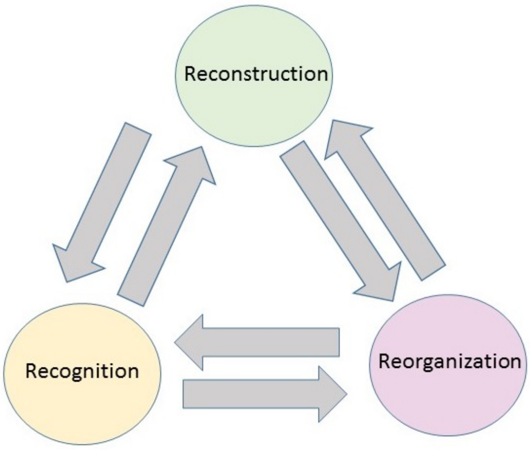


Figure 1: The 3Rs in Computer Vision [119].

variety of processes ranging from feature matching in multiple view geometry to other visual cues like shading, texture or depth from stereo or RGB-D cameras. All processes result in point clouds or 3D models as outputs. Some examples for reconstruction tasks are Structure from Motion (SfM), scene reconstruction, shape from shading. Recognition involves both 2D problems like handwritten recognition, face recognition, scene recognition or object recognition, and 3D problems like 3D object recognition from point clouds which assists robotics manipulations. Reorganization involves bottom-up vision which is pixel segmentation into groups of pixels that can represent facts. Reorganization tasks range from low-level vision like scene segmentation to high-level tasks like semantic segmentation [183, 37, 169] which has an overlapping contribution to recognition tasks. These tasks can be viewed as fact finding from the visual data like images or videos which answers the conventional question "to know what is where by looking." In other words, "vision is the process of discovering from images what is present in the world, and where it is [122]."

Between each of the 3Rs tasks, the output from one task can provide information that helps another task. To give a specific example, 3D faces which are the outputs of a reconstruction task can give more information and assist face recognition [33]. On the other hand, recognition can give a prior knowledge to create an object specific 3D model for a reconstruction task [16]. For reorganization and recognition, reorganization can provide contours, regions and object candidates for object recognition [151]. On the contrary, recognition can generate object proposal regions for segmentation [85]. This can also be viewed as recognition providing context for reorganization [88]. For reorganization and reconstruction, this link is still to be investigated of how low-level features such as edges or contours will provide any information to reconstruction and vice versa.

Recognition tasks are closest to languages since the output is likely to be interpretable as a word. **Lexical semantics** [72, 83] will come into play as an interface in this scenario.

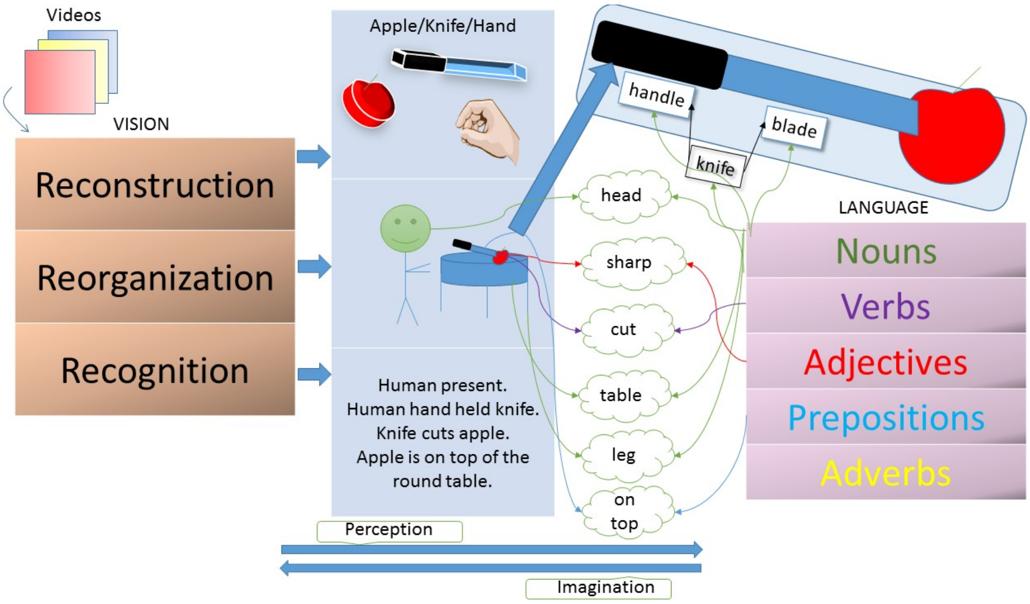


Figure 2: From 3Rs in Computer Vision to Language by Lexical Semantics.

For example, object or scene classes are nouns. Activities are verbs. Object attributes are adjectives. Relations between objects or object and scene are prepositions. Temporal relations of an object and an activity are adverbs. Reorganization tasks deal with a lower-level feature set which can be interpreted as primitive parts of shapes, textures, colors, regions and motions. These primitive parts define a higher-level vision so they do not refer to any specific object or scene that can be described in words for communication but they are essential for learning new words as they implicitly describe object or scene properties. Reconstruction involves geometry of real world physics which provides richer object or scene properties than reorganization tasks. Reconstruction mainly helps in real time high-precision robotics manipulation actions which is its interpretation in the real world.

1.2 Natural Language Processing tasks and their relationships to Computer Vision

Based on the **Vauquois triangle** for Machine Translation [188], Natural Language Processing (NLP) tasks can be summarized into the concept ranged from **syntax** to **semantics** and to **pragmatics** at the top level to achieve communication. Syntax can be in the study of morphology that studies word forms or in the study of compositionality that studies the composition of smaller language units like words to larger units like phrases or sentences. Semantics tries to provide meaning by finding relations between words, phrases, sentences or discourse. Pragmatics tries to interpret the meaning in the presence of a specific context where the standard meaning may change. For instance, an ironic sentence cannot be correctly interpreted without any side information that indicates the nonlinearity in the speaker's intention. Ambiguity in language interpretation

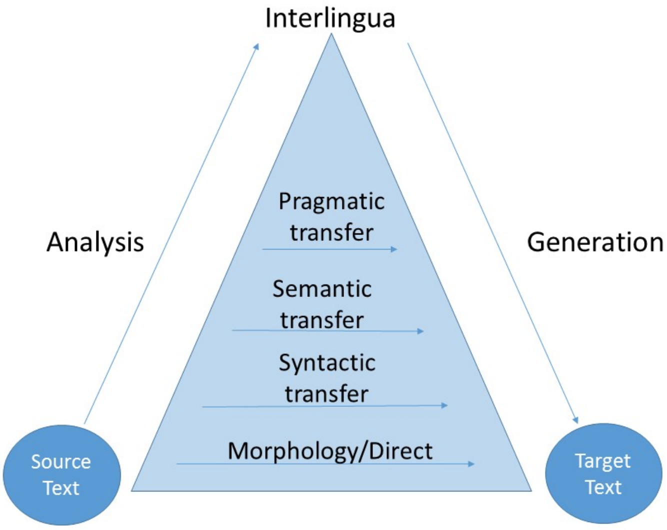


Figure 3: The Vauquois triangle for Machine Translation [188] with added Pragmatic transfer.

is the main obstacle for an intelligence system to overcome and achieve language understanding. Some complex tasks in NLP are machine translation, information extraction, dialog interface, question answering, parsing, summarization etc.

Words can also be viewed as labels, so manipulating words is equivalent to **label manipulation**. Manipulating words which contain their own formal meanings relies on the ambiguous and natural visual context that they refer to. This produces a gap in meaning between the high-level formal data and the low-level natural data. **Bridging the Semantic Gap** [204] is to find a pipeline that will go from visual data like pixels or contours to language data like words or phrases. This is to distinguish the meaning of the constructions of visual data. To give some specific examples, labeling an image patch that contains an object with a word is object recognition. Labeling a background in an image is scene recognition. Assigning words for pixel grouping is semantic segmentation [183, 37, 169]. If we know how the words are related to each other, it will give a clue for visual processing to better disambiguate different visual constructs. For instance, a 'knife' is more likely to 'cut' a 'cucumber' than a 'chair' in a 'kitchen' because the meaning of their interaction is presented as a concept in the real world.

NLP concepts were borrowed to solve CV tasks several times and they help provide interpretability to CV tasks [83]. Human actions and object manipulations can be described using language [82]. Action grammar [145, 82] exhibits the use of syntactic information in the compositionality of motions into activities. Semantic Event Chain (SEC) [1] provides an interpretable and grounded framework for manipulation in imitation learning inspired by the mirror-neuron system [155]. Inferring goals and intentions of an agent from cognitive effects [47] demonstrates the pragmatics aspect in CV which

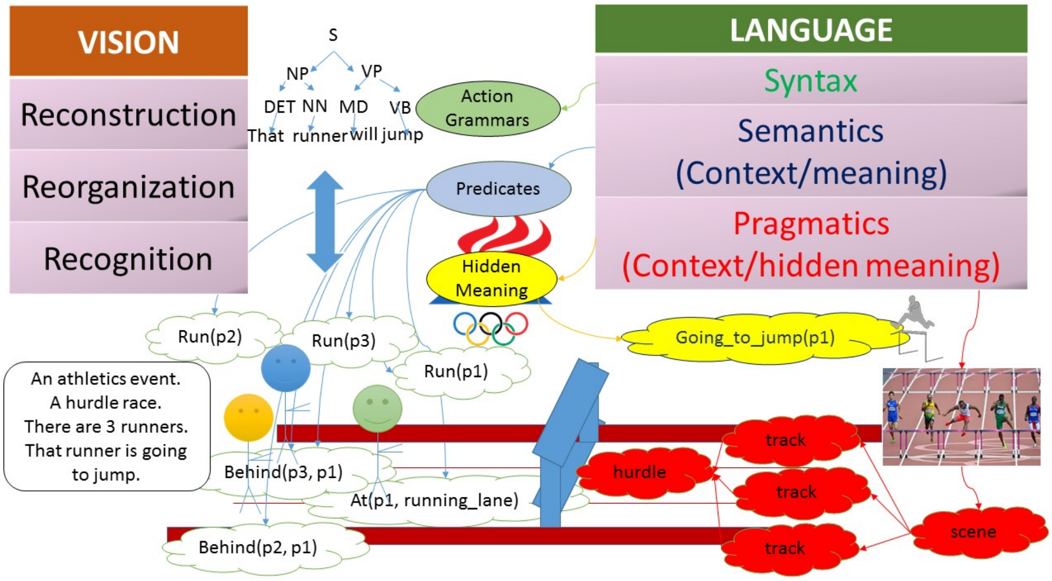


Figure 4: From the Vanquois triangle in NLP to Computer Vision by event semantics. The descriptor may infer that the shower of this image intends to ask the possible action that is going to happen next.

is an essential part in Human Robot Interaction (HRI) for Social Robots.

By having these tasks in mind, one can creatively define new joint tasks from these two domains of language and vision understanding which will better model an arbitrary cognitive phenomenon. This survey will categorize the existing joint tasks of language and vision by the application domain, namely multimedia and robotics. These two domains will further be described in section 2 and section 3 respectively.

This article is structured as follows. Section 2 will provide a survey of language and vision literature in the multimedia domain. Section 3 will provide another survey of language and vision works in robotics applications. Section 4 will focus on distributional semantics in language and vision.

2 Language and Vision for Multimedia

For multimedia, the data is in files from the internet containing images, videos and natural language texts. For example, a news article will contain news that was written by a journalist and a photo related to the news content. There can be a clip video which contains a reporter and a video that depicts the snapshot of the scene where the event in the news occurred. The cooccurence between an image and texts depicts their relations. To give a specific example, an image in the news [27] is likely to have a face and using the accompany news text will help identify the identity of that face including a lot of side information of that person such as occupation, age or gender. This is under the condition that the coreference between the news text and the face can be accurately discovered. Another example of finding relations between images and texts is entry-level categorization[141] by finding the 'naturalness' which is the way people calling an object in the world. A word will make sense when used to describe a specific object depending on contexts. That word will not be weird and can be easily used for an appropriate and unambiguous communication if it is the word normal people are using naturally.

Both language and visual data provide two sets of information that are combined into the whole story. This conforms to the theory of **semiotics** [81] which is the study of the relations of signs and their meanings. Using NLP concepts from the previous section, semiotics also has an equivalent viewpoint to those concepts [140]. First, semiotics studies the relationship between signs and meaning which is equivalent to semantics. Second, the formal relation between signs is equivalent to syntax. Third, the relation of signs to the human interpretors is equivalent to pragmatics. By restricting signs to visual data, this concludes that semiotics can also processed as CV extracting interesting signs for NLP to realize the corresponding meanings.

The tasks for language and vision in multimedia mainly fall into two categories, visual description and visual retrieval.

2.1 Visual Description

2.1.1 Attribute-based Vision

Associating words and pictures [13] is equivalent to the Recognition task in CV. We have words to describe objects and their relations in an image. Object recognition traditionally tries to categorize an image to a fixed set of name tags. [59] argues that an image has more information than just a set of name tags and categorization should change to description. Attribute-based recognition [62, 106, 102] describes and summarizes object **properties** in words in which an unusual property of that object can be detected and recognizing novel objects can be done with a few or zero training examples from category textual descriptions. The attributes may be binary values for some easily recognizable properties like 4-legged animal or walking left. However, some properties may not be easily recognizable like smiling. For these types of properties, the relative attributes [144] help describe the strength of the property in each image by using a Learning to Rank framework (LtR) [115].

The key is that the attributes will provide a set of key contexts as a knowledge vault for recognizing a specific object by its properties. The attributes can be discovered using a classifier which learns a mapping from an image to each property [62, 106]. The attribute words become an immediate representation that will help bridging the semantic gap between the visual space and the label space. In other words, the attributes are textual abstractions of an object. This introduces another dimension for feature engineering in which there are common features that can be shared across tasks [60], such as object parts, and some features that will be task-specific and unique for each task, such as the hardness of a diamond or the softness of a gold when exposed to heat. Mostly, attributes will be shares commonly for objects in the same category.

From this viewpoint, we can incorporate the feature learning framework into attribute-based recognition. That is, we need to determine the possible set of attributes whether they should be words, phrases or sentences. Also, which attributes should be used to recognize what kind of object becomes a feature selection problem that will impact the recognition performance explicitly. In addition, this can be further inferred that a specific set of words will correspond to a specific set of corresponding objects. [70] proposed learned visual attributes with data in an unsupervised setting with feature selection but with visual words not textual words.

Attribute-based vision was found useful in many specific applications where property contexts are crucial and informative including animal recognition [29, 106], face recognition [27, 102], finding iconic images [26], unsupervised attribute recovery on web data [28] and object identification for robotics [173]. Recognizing an image should result in a rich meaning that informatively describe what is going on in the image. Beyond words in isolation, phrases [161] and sentences [61, 196] can expand more dimensions from an image.

2.1.2 Visual Captioning

Beyond unordered words is a sentence. Attractive images usually come with a long corresponding text that tells a story. For example, an image from a sport headline will depict the decisive moment of the game and the corresponding text will describe the details. Generating a caption from an image needs to answer a set of specific questions about that image to ensure understanding. First, the classical "to know what is where by looking" [122] is still applied. Second, the contextual information need to be able to answer "When, For What, and How?" [143] questions to make sure that this information is relevant. The meaning representation needs to capture the answers to these questions.

For the words to be meaningful, they should have interpretations as visual meanings [51]. Sentence generation systems may discard nouns or adjectives that are non-visual from their visual recognition results to reduce bias errors. Scene information can also reduce bias errors in object recognition since only a specific set of objects will naturally occurs in a given scene [201].

Collecting captions from visually similar images can generate good descriptions. [142] finds the best caption from the most visually similar image based on content matching which is the distance measurement consists of the object, people, stuff and scene detectors. [103] goes further by summarizing the captions from the candidate similar images. The motivation for borrowing captions from similar images is that measuring similarity between visual features is easier than measuring in both visual and text features. This also concludes that Nearest Neighbor methods works well for image captioning from the remarkable automatic evaluation scores given a good embedding space by Kernel Canonical Correlation Analysis (KCCA) [90] or Neural Networks [168].

To **generate a sentence** for an image, a certain amount of low-level visual information is needed to be extracted. The primitive set of information is the ⟨Objects, Actions, Scenes⟩ triplets to represent meaning as a Markov Random Field (MRF) potential edges [61]. Then, the parameters are learned using human annotated examples. Consider part-of-speech, the quadruplets of ⟨Nouns, Verbs, Scenes, Prepositions⟩ can represent meaning extracted from visual detectors [196]. Visual modules will extract objects that are either a subject or an object in the sentence. Then, a Hidden Markov Model (HMM) is used to decode the most probable sentence from a finite set of quadruplets along with some corpus-guided priors for verb and scene (preposition) predictions. [114] represents meaning using objects (nouns), visual attributes (adjectives) and spatial relationships (prepositions) as ⟨⟨adj1, obj1⟩, prep, ⟨ adj2, obj2⟩⟩. Then, the sentence is generated by phrase fusion using web-scale n-grams for determining probabilities. Babytalk [101] goes further by using Conditional Random Field (CRF) for predicting the best ⟨⟨adj1, obj1⟩, prep, ⟨ adj2, obj2⟩⟩ triplet. Then, the output is decoded using a language model and generated as a template-based sentence. Midge [134] makes an additional improvement by tying the syntactic models to visual detections so that the template is more relaxed and the sentence looks more natural. Visual Dependency Grammar (VDG) [57] proposes dependency constraints, such as spatial relations of pixel, so that the visual detection stage will have a structured output to be fed into a template-based generation system. This step leverages noises from object detectors and provides more stability given gold standard region annotations. Recent methods use a Convolutional Neural Networks (CNN) to detect visual features and using Recurrent Neural Networks (RNN) [97] or Long-Short Term Memory (LSTM) [189] to generate the sentence description. Both methods are implemented in the NeuralTalk2 system[1].

For the recent trend of image captioning datasets, [66] provides a detailed explanation along with an empirical evaluation across standard datasets. Other interesting datasets include the Amazon product data [127, 128] and the Comprehensive Cars (CompCars) dataset [192]. Visual Question Answering (VQA) [6] is also a new interesting task for image captioning in which senses and knowledges from the question should be considered in the visual extraction process.

2.2 Visual Retrieval

Content-based Image Retrieval (CBIR) annotates an image with keyword tags so that the query words will be matched to the precomputed keyword tags. The systems try to annotate an image region with a word similar to semantic segmentation. Some approaches are Co-occurence model on words in image grids[138], Machine Translation on image blobs to words [55], probabilistic models on blobs to words [12], topic models on blobs and words [31], Cross-media Relevance Model (CMRM) on the joint distribution of image blobs and words[94] and Continuous-space Relevance Model (CRM) which further models the semantics rather than color, texture or shape features [107]. Since image class label and its annotated tags are likely to have some relations to each other, [191] proposes an extension of supervised Latent Dirichlet Allocation (sLDA) [129] to jointly model the latent spaces of image classification and image annotation.

[1] https://github.com/karpathy/neuraltalk2

3 Language and Vision for Robotics

3.1 Symbol Grounding

Robotics is a research field involving both **perception** and **manipulation** of the world [181]. There are many modalities of perception, for example, vision, sound, smell, taste, taction or balance, etc. By hardware sensors, each of these modalities provides an sensory information to be processed. To manipulate the world environment, a robot controls its body parts and applies physical forces, embodied from its mechanics, to perform actions which may result in its own movements or a change to its environment. In this section, we will consider research works in language and vision where the resulting applications are in robotics.

For robotics, language becomes symbols and vision dominates perception. Bridging language and vision is equivalent to the **symbol grounding problem** [86] in the lens of robotics which tries to foster an autonomous agent to reason and react in the real world. The symbol grounding problem is about grounding the meaning of the symbols to the perception of the real world. If the grounding process tries to ground the meaning only in symbols, it can be done for some cases in Compositional semantic or Machine Reading [58] but may result in an infinite loop of referencing which will lead to nowhere near the concept understanding if the symbols really need a perceptual information as a context to be grounded in. For example, words like yellow, board, fast, prawn or walk need a real world perception in order to understand their meaning. Another example [137] in Wordnet [133] is the word pair 'sleep' and 'asleep' which have pointers to each other as a loop.

The symbol grounding problem can be categorized into five subcategories [45]. First, the **physical symbol grounding** deals with grounding symbols into perception. This conforms to the original definition in [86]. Second, **perceptual anchoring** [46] is to connect the sensor data from an object to a higher order symbol that refers to that object. Also, the connection must be maintained in time. Third, **grounding words in action** [159] maintains a hierarchical representation of concepts for abstract words: higher-order concepts are grounded into basic concepts and sensorimotor grounded actions. Fourth, **social symbol grounding** [36] tries to share the connection after anchoring for one agent to many agents. This involves the pragmatics level of meaning understanding. Fifth, **grounding symbols in the semantic web** [96] tries to ground the symbols into a large ontology knowledge base from the internet. This is like grounding in text which is the least restricted setting of the symbol grounding problem. The symbol grounding problem represents a gap between symbols and perception.

3.2 Robotics Vision

Visual data can enable perception in a cognitive system that fulfills the grounding process. Current computer vision techniques, however, are limited when only low-level information, like pixels, is being used [2]. Humans process perceptual inputs by using their knowledge about things they perceive in all modalities in the form of words, phrases and sentences [136]. The language may be the knowledge about objects, scenes, actions or events in the real world in which these perceptions can be given by Computer Vision systems. The knowledge needs relations to make sense and understand the meaning.

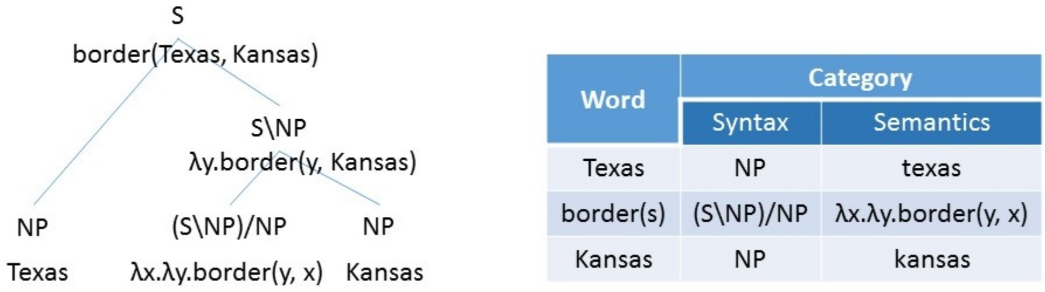
This gives rises to a **top-down active visual process** [2] where language will request some actions from a sensorimotor system which will instantiate a new data collection from perceptual sensors. For example, if an object is far away, a language executive may request for an action that ends in a clearer view point which has a better pose and scale of that object. The action may be decomposable into subtasks and may need a planning system for a robot to perform the action. The visual information from the new viewpoint may be collectible only from direct observation not by inference. This schema is interactive between the conventional bottom-up and this novel top-down vision system. This is like an active learning in education [131] that a student can actively ask a teacher for a clearer understanding of the subject but a student also needs to be smart so that the answer from his question will provide the desired information or lead to an interesting discourse between his teacher and him.

Robotics Vision (RV) [48] is different from the traditional Computer Vision (CV). Nowadays, a large part of CV relies on Machine Learning where the performance relies on the volume of the data. Recognition in CV focuses on category-level recognition which aims to be general and can be acquired from the internet to create a big data paradigm. In contrast, RV uses reliable hardware sensors like depth camera [104] or motion camera [15] so the performance relies on the number of sensors instead. Also, recognition in RV focuses on situated-level recognition where the context environment is limited from the embodiment of the sensor hardware. Robotics Vision tasks relate to how a robot can perform sequences of actions on affordable objects to manipulate the real-world environment. Such tasks need some information involving detecting and recognizing objects, object motion tracking, human activity recognition, etc. This is to give a robot both static and dynamic information about its surrounding contexts.

Interfacing CV and RV needs domain adaptation techniques [49] since online images and real world objects are different. [105] tries to incorporate data available from the internet, namely, Google's 3D Warehouse to solve 3D point cloud object detection in the real world in which the problem describes the need for domain adaptation and the methods involve domain adaptation formulations.

3.3 Situated Language in Robotics

For robotics, languages are used to describe the physical world for a robot to understand its environment. This problem is another form of the symbol grounding problem known as **grounded language acquisition** or embodied language problem. A robot should be able to perceive and transform the information from its contextual perception into language using semantic structures. By this, a language can be grounded into another predefined grammar that can represent meaning. The most well-known approach is **Semantic Parsing (SP)** [202] which transforms words into logic predicates. At first, SP was first introduced as a natural language interface for question answering database system [202]. SP tries to map a natural language sentence to a corresponding meaning



Word	Category	
	Syntax	Semantics
Texas	NP	texas
border(s)	(S\NP)/NP	λx.λy.border(y, x)
Kansas	NP	kansas

Figure 5: An example parse tree for Combinatorial Categorical Grammar (CCG) from [203].

representation which can be a logical form like λ-calculus. The parsing model is induced from a parallel data of sentences and meaning representations. For λ-calculus, the parser will parse a sentence using Combinatorial Categorical Grammar (CCG) [170] as rules to construct a parse tree. The motivation is from the principle of compositionality [69], "The meaning of the whole is determined by the meanings of the parts and the manner in which they are combined."

Combinatorial Categorical Grammar (CCG) [170] defines a syntax in addition to λ-calculus that defines semantics in terms of logic. Similar to Context-Free Grammar, CCG consists of lexica and their parsing rules. However, CCG defines categories or types for each lexicon and their combinatory rules to combine those categories together. For example taken from [203], a word 'Texas' is denoted as an atomic symbol 'NP' with its meaning of the state 'texas'. Another word 'border(s)' which takes arguments representing US states like Texas or Kansas of type 'NP' will have a syntax as $(S\backslash NP)/NP$ and its meaning of a lambda function $\lambda x.\lambda y.$borders(y,x). That is, the word 'border(s)' has a syntax that will take a word of type 'NP' to the right and another word of type 'NP' to the left of the word 'border(s)' and yield the parse tree as in Figure 5. Like many other parsing problems, there can be many possible parse trees in which some of them may be incorrect that they may still contain a lambda form in the tree or the meaning is incorrect. Therefore, a prediction model can be employed to solve this problem as another structured prediction problem[50, 176, 163].

The main challenge of SP is its scalability across domains which is from its supervised training scheme. Some attempts to resolve this problem are by reducing supervision to make unsupervised SP [152], using distributional semantics which encodes rich interactions between lexica [113] or using semi-structured data for a strong typing constraint by the database schema [146]. Some applications of SP are robot control [124, 52], question answering [202, 152, 198, 24] and smart phone voice command [130, 165]. Some software for SP are SEMPRE[2] [24]or Cornell SPF[3] [8] or XYZ parser[4] [11].

Another alternative to SP is Abstract Meaning Representation (AMR) [9]. AMR

[2] http://www-nlp.stanford.edu/software/sempre/

[3] https://bitbucket.org/yoavartzi/spf

[4] http://bioai8core.fulton.asu.edu/kparser/

tries to map a sentence to a graph representation which encodes the meaning of that sentence about “who is doing what to whom”. By this, AMR makes morphology and syntax abstract into predicate-argument structures. A sentence becomes a rooted directed acyclic graph and is labeled on edges for relations and on leaves on concepts. The AMR project aims to output a large scale semantic bank by human annotation. There are some attempts such as [153] that try to use string-to-tree syntactic Machine Translation systems which is a more automatic process for constructing AMR.

Not all words can describe the geometric properties of the world. For example, words like ‘special’ or ‘perhaps’ provide emotional information but not about the real world context. For robotics, most words should involve either **actions** or contexts. Actions can be words like ‘move’, ‘put’, ‘push’ or ‘pick up’ which need an argument that can be the robot itself or other things in the vicinity for execution. The contexts can be words or phrases like ‘a table on the right’, ‘a green button on the wall’ or ‘the red car under the roof’ that show **affordance** [39] of an object so that it can be specified for the robot to perform actions with attention. Prepositions from these phrases, like ‘on’ in ‘on the wall’ or ‘under’ in ‘under the roof’, encode **spatial relations** which are essential contexts to specify where to perform actions for a robot. Affordance is a property of an object related to tools and actions, for example, a cucumber can be cut with a knife or a green button can be pushed with a hand. This defines an affordance of ‘able to cut’ to a knife and ‘can be cut’ to a cucumber. Affordance helps reasoning by giving the relational information about objects, tools, actions along with their pre-conditions and post-conditions after the action is applied. Adverbs depict the details of an action which can be **force** or **speed**. A robot should know how to perform a successful action by adjusting its own control parameters to be precise with tools and object affordances. For example, cutting a ‘cucumber’ and a ‘wood’ need different tools like a ‘knife’ and an ‘axe’ in which a robot should handle them differently and apply a specific amount of forces for each tool. Moreover, to perform a ‘rescue’ task and a ‘finding object’ task need different speed where ‘rescue’ should be executed with full speed while ‘find’ can be done with a normal pace. In addition to vision and language, recently advanced tactile sensors [200] will help in perceiving and adjusting forces by sensing forces and frictions directly.

3.4 Recent Works in Language and Vision for Robotics

Robotics has various interesting applications and we will describe only a salient set of them. We conclude that robotics tasks involving language and vision can be categorized into three main tasks, a robot talking to human, a robot that learns from human actions and a robot that performs navigation.

First, a robot can **interact with human via language** which is situated and grounded in perception. This needs both language understanding and generation as well as some representation that will integrate perception into language. For situated language generation tasks, [40] applies semantic parsing to ground simulated Robocup soccer events into language for commentary. This work goes beyond a manual template system to learning to perform. Automated sport game models (ASPOGAMO) [18] tries to track sportsmen and ball positions via detection and tracking systems on broadcasted football games. ASPOGAMO can handle changing lighting conditions, fast camera motions and distant players. Unifying both system for tracking and sportcasting is another promising direction. For situated language understanding tasks, [124] parses user natural language instructions into a formal representation which commands robots with Probabilistic CCG for semantic parsing [203]. A robot will follow the parsed instructions and execute its control for routing. [123] further incorporates visual attributes for grounding words describing objects based on perception. The model incorporates both semantic parsing for language and visual attribute classification for vision and is trained via EM algorithm which jointly learns language and attribute relations.

To unify generation and understanding, Grounded Situation Model (GSM) [126] is a situated conversation agent that bridges perceptions, language and actions with semantic representation based on parsing. Its belief is updated with a mixture of visual, language and proprioceptive data. GSM can answer questions and perform basic actions via verbal interaction. [177] categorizes robot language tasks into following instructions, asking questions and requesting help. A robot will try to find uncertain parts in the command and ask a targeted question for clarification than it will perform better actions based on the obtained information. To achieve this, the G^3 framework [178] which is a probabilistic model is used to model the inverse semantics from the uncertain part of the world to a word in a sentence. Hence, this involves both language understanding and language generation.

Recently, [190] unifies language and vision for robotics again by bridging visual, language, speech and control data for a forklift robot. A robot can recognize objects based on one example using one-shot visual memory. Its natural language interface is by speech processing or pen gestures. It is equipped with reliable sensors and an anytime motion planner that enables its local actions without global information. It has a nice annunciation and visualization interfaces. A robot also has a safety mechanism for other workers around by pedestrian detection and shout detection. For further information in this topic, [125] provides a survey for verbal and nonverbal human-robot interaction.

Second, a robot can **learn to perform actions** by imitating or observing human actions. This setting is sometimes denoted as robot learning from demonstration (LfD), imitation learning [148] or observational learning [10]. Instead of manual hard coding, a robot can learn from either a human teacher or other robots. LfD helps program robotic controls to perform actions. The motivation is from the **mirror-neuron system** [155] which is a neuron that will fire both when an animal performs and observes a certain action. This phenomenon enables human to see and learn other people's actions including understanding intentions and emotions attached in those actions.

Learning from demonstration (LfD) [7] tries to learn a mapping of state and action pairs from teacher's demonstration (s_t, a_t) as a supervised learning setting so that the learned policy from state S to action A which is $\pi : S \rightarrow A$ will have some performance guarantees. The mapping between actions is defined by $T(s'|s, a) : S \times A \times S \rightarrow [0, 1]$. Moreover, the states may not be fully observable, so the observed state Z is from another mapping $S \rightarrow Z$ and a policy will be $\pi : Z \rightarrow A$ instead. For more information, [7] provides a unified framework for LfD.

Julian Jaynes' bicameral mind [93] theorizes an existence of language intervention in human consciousness and motivates an incorporation of language and vision to LfD. The bicameral mind means a mind with two chambers in which one room is speaking as an executive in language and another room just obeys and perform actions. This defines a model of consciousness in ancient human mind which does not apply to people in nowadays. Whether this is true or not still needs a lot of evidences and may be too hard to prove or disprove but having this model we can view a robot with language as an embodiment of the bicameral mind which will see, talk and perform actions. Language and vision can provide data and models for this scenario. A robot will recognize actions by watching a demonstration from either a real world teacher or a video. Then, it will ground its perceptions to language and learn to plan its own actions. Finally, the planned imitated actions will be carried out by its motor actuators.

Action grammar [82, 145] is the most common interface for language and vision in LfD. Action grammar can model the hand-object-tool relations and incrementally construct an activity tree [172]. An interesting feature is interleaving activities can be recognized using action grammar. This provides a strength from the nature of the sequential data in which there can be many motif sequences to be recognized. [180] introduces a prior knowledge of action-tool relations mined from their cooccurence in Gigawords corpus. This results in a learned prior knowledge that reflects the real world and helps improve activity recognition using textual knowledge. [195] further models the hand-object-tool, object-scene and object-scene-attribute relations as a multi-label classification task with a prior from Gigawords corpus as in [180]. The results suggest that the language prior will rule out some relations that will never occur because they do not make sense such as using a cup of water to fold a t-shirt. Furthermore, the large amount of actions can be easily observed from real world descriptions so having the meaning from texts helps a robot learn starting from words to actions.

Nowadays, there are a large amount of online videos in which demonstration videos can be retrieved easily for human learning. A robot can do the same by watching videos and imitating actions. A robot may needs to be given all language, speech and visual data for a complete understanding [120]. However, using only language and vision can efficiently teach a robot to perform a task successfully as in cooking [194] or t-shirt folding [166]. Semantic Parsing is an enhancement to the action grammar in which a post-condition can be inferred using the semantics of the Combinatorial Categorical Grammar itself [197]. For example, cutting a cucumber will result in divided cucumbers. This is also called as manipulating action consequences [193] which represents object-centric consequences as a knowledge graph. The results can be said that they are the desired goals for the action performers. The corresponding semantics is $\lambda.x\lambda.y\ cut(x,y) \rightarrow divided(y)$ where x is a cutting tool, such as a knife, and y is a cuttable object, such as a cucumber. [166] goes further and introduces the knowledge transfer scheme where a robot will learn from a human demonstration then transfer its knowledge by teaching to another human.

Third, a robot can perform **planning for navigation**. With an understanding of the surrounding world, a robot can perform reasoning and make a plan to achieve its goals. However, real world navigation requires map making with some techniques like Simultaneous Localization And Mapping (SLAM) [54]. For robotics research, a simulated world is created to test the software so a navigation is on a visual world instead. The map can be either egocentric or top view (aerial). The spatial relations are frequently used to refer to a subjective relation between a robot and a map. Therefore, the language used will be more pragmatics since many meanings are hidden as a presupposition or an implicature such as the left-right-straight-backward directions. There are also many scenarios where navigation plans can be distinctively different. The scenario can range from an event or a place where an action should be immediately carried, like an emergency event with a dense human crowd, to a mysterious and dangerous place but with a few people, like a mine, to a household kitchen which is a safe place but the objects are cluttered.

A planning for navigation problem can be casted as a situated language generation [73]. The task is to generating instructions in virtual environments which is giving directions and instructions [35], such as 'pushing a second button on the wall', to accomplish the final task which is taking the trophy. The contexts in a virtual environment are converted into natural language using Tree-Adjoining Grammar (TAG) which will be further converted into a planning problem [98, 74]. Some visual cues, such as listener gazes [75], can help the system generating a more meaningful discourse because it will have some feedback information that help inferring the mental state of the listener while giving instructions.

4 Distributional Semantics in Language and Vision

4.1 Distributional Semantics

Distributional Semantics [87] relies on the hypothesis that words which occur in the similar contexts are similar. This hypothesis can recover word meaning from cooccurrence statistics between words and contexts in which they appear. Distributional Semantic Models (DSMs) use the vector space and its properties to model the meaning. The semantic vector space will represent a word as a data point and will encode the similarity and relatedness between words in term of measurements between those data points.

Word similarities can be measured by Euclidean distance between the data points or cosine similarity of the angle between a pair of words. Similar words will have similar vector representations and will be near to each other in the vector space. Word relatedness can be observed from the displacements or offsets between the vectors which represent relations. For instance, the word 'King' can be mapped to another vector which is very similar to the word 'Queen' by subtracting with the word vector 'man' and adding with the word vector 'woman'.

Modeling meaning by word cooccurence as the only source of information limits its connection to the real world. One may argue that the meaning can be described in language and provide understanding without any perception as a prerequisite for communication. So, word cooccurence can be derived and provide meaning because an appropriate corpus can be created to serve any specific meaning. However, one can also argue that there are a lot of knowledge that cannot be understood without a grounded perception.

For example, a 'coconut' can occur with other words that represent 'fruits' or 'tropical countries' which describe some aspects of a 'coconut'. Nevertheless, one cannot understand its unique shape, color and texture without perceiving a real coconut. If there are any similar fruits whose shape, color or texture are the same as a coconut, one can effectively describe a coconut by each information aspect but there will be still a question of 'What does it really look like' remains. Therefore, perception is essential in order to answer these questions and provide more information sources for modeling meaning.

For an object, its context can be where this specific object will appear in. This encapsulates the natural information about how scenes and objects can relate to each other [43]. The context also includes the meaning of the real world that similar objects in the same category will be likely to appear in the same context which can be further inferred that those objects cooccur with each other in some specific patterns as well. For example, a 'lion' and a 'deer' are likely to be in a 'forest' or a 'zoo'. If we were to observe them both in a 'forest', a 'lion' is likely to be chasing a 'deer' for its meal. Understanding these meaning will help the reasoning process about the real world.

4.2 Vector Space Models

Vector Space Models (VSMs) [186] is an algebraic model which represent text documents as vectors where each dimension correspond to a term. Putting the vectors together forms a term-document matrix which represent word cooccurence in documents. The best known method for computing the values in this matrix is term frequency-inverse document frequency weighting (tf-idf) [162].

The tf-idf weighting computes the importance of each word to a document in a corpus collection [121]. A word with high term frequency 'tf' will have a high score but it is also proportional by the inverse of the frequency of how it appears across the corpus 'idf'. Given a term 't' and a document 'd', tf-idf weighting computes the score for each term in each document as follows.

$$\begin{aligned} \text{tf-idf}_{t,d} =& tf_{t,d} \times idf_t. \\ idf_t =& \log \frac{N}{df_t}. \end{aligned} \tag{1}$$

where 'N' denotes the total number of the documents in the corpus collection. The tf-idf weighting has many useful applications in which the best well known application is to perform ranking for document retrieval. tf-idf weighting is so simple and computationally efficient that it is usually used as an *important preprocessing step* for various text mining algorithms. There are also some variations in computing the term frequency 'tf' and the inverse document frequency 'idf' which are investigated by [121].

VSMs implement the idea of similarity measurement in the algebraic space. This interpretation of vectors and their similarities are building blocks for the following distributional semantic models.

4.3 Distributed Word Representation

There are two types of connectionist representations, local and distributed [149, 89]. Local representations have a small number of features for each item while distributed representations have a large number of features from nearly all features in the feature pool. Local representations are sparse and capture salient signals from particular features. In contrast, distributed representations are dense, continuous and depict patterns over many signals. Local representations have lower representation power than distributed representations which are dense and compact.

4.3.1 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) [53] or Latent Semantic Indexing (LSI) is the most well known instance of distributed word representation which tries to recover word relations from a corpus. LSA uses Singular Value Decomposition (SVD) on the term-document matrix 'C' which outputs a low-rank approximation that can be used as a weighting scheme.

$$\begin{aligned} C &= U \Sigma V^T. \\ C_k &= U'_k \Sigma'_k V'^T_k. \end{aligned} \tag{2}$$

where 'k' is the approximate rank of the term-document matrix 'C' into the reduced rank matrix'C_k'. An improvement made by LSA is that it can model synonimity and polysemy [53] as well as a grouping of words into concepts while other prior models which rely on term matching cannot. LSA has better performance compared to tf-idf but lacks unit-like interpretability featured in local representation.

Probabilistic Latent Semantic Indexing (pLSI) [91] or aspect model is the probabilistic version of LSA which models each word in a document as a sample from a mixture model of conditionally independent multinomial distributions. Each document consists of topics and each topic consists of words. pLSI has an improvement over LSA in terms of the interpretability of word-topic and topic-document relations.

4.3.2 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) [32] was proposed to overcome the overfitting problem in pLSI by introducing the Dirichlet prior over the multinomial topic distribution. The generative story of LDA [32] is as follows,

1. For each document:

    (a) Choose $N \sim$ Poisson(ξ).

(b) Choose $\theta \sim \mathrm{Dir}(\alpha)$.

(c) For each of the N words w_n:

i. Choose a topic $z_n \sim \mathrm{Multinomial}(\theta)$.

ii. Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

where α is the parameter of the Dirichlet prior of the document-topic distribution, β is the parameter of the Dirichlet prior of the topic-word distribution, θ is the topic distribution for document i, N is the number of words in the document. For a detailed explanation on topic models, some literature surveys on topic models [30, 171, 154, 118] provide an excellent explanation including various new models along with the evolution in the field of topic models created by LDA.

This is to show that distributed representations are essential for extracting interesting patterns from the corpus but it is also beneficial to add interpretablility like local representations for further understanding.

4.4 Count and Prediction-based Word Embedding

4.4.1 Word2Vec

While LSA represents the geometry aspect of meaning and LDA represent the probabilistic generative aspect of meaning [34], the recently proposed Word2Vec [132] represents the neural model of the meaning. In contrast to LSA and LDA, Word2Vec produces a word embedding, in other words, a distributed representation [184] which is equivalent to factorizing the term-term matrix [111]. Word embedding is typically induced by neural language models [23, 139] which predict the context given an input word. The training and testing by prediction was typically slow and scaled by the vocabulary size.

Word2Vec solves this problem by reducing the feed forward neural model to a simple log linear model so that less nonlinearity is involved. The log linear model is just a simple softmax function. However, this made a breakthrough since a high quality word embedding can be obtained by predictive training on a very large corpora. Comparing to the traditional training methods like LSA or LDA which are based on statistics given by counting the terms cooccurence, whether the predictive models will make an improvement in performance [14] or just give more insight information in parameter tuning [112] is still an ongoing open problem.

Word2Vec provides a significant improvement over LSA and LDA from its quality of the output word embedding. The resulting representation is encoded with word meaning so similar words will have similar vector representations, for example, the vector('car') will be similar to the vector('driver'). Moreover, the relationship between words is also preserved in term of displacement between points such that basic vector operations on these points will be meaningful, for example, vector('Paris') - vector('France') + vector('Italy') will be resulting in a vector very similar to the vector('Rome'). Also, the displacement can capture the syntactic relations, for instance, vector('sweet') - vector('sweetest') + vector('fastest') will be resulting in a vector very similar to the vector('fast').

4.4.2 Word2Vec models

Word2Vec has two models, Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram models. Furthermore, there are also two training techniques, namely, Hierarchical Softmax and Negative Sampling. The CBOW model tries to predict the vector representation of the current word given context words. This corresponds to the count-based models [14]. On the other hand, the Skip-gram model tries to predict the vector representations of the context words given the current word. This corresponds to the predict-based models [14]. The CBOW is faster but the Skip-gram model better represents the infrequent words. From the original experiment in [132] and also another experiment in citebaroni2014don, the Skip-gram model is superior in term of accuracy.

Given a sequence of words $w_1, w_2, w_3, \ldots, w_T$, the CBOW model tries to maximize the objective function of a log probability of a word w_t,

$$\frac{1}{T}\sum_{t=1}^{T} \log p(w_t | w_{t-c}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+c}), \tag{3}$$

where c is the size of training context. The Skip-gram model tries to maximize the objective function in terms of average log probability,

$$\frac{1}{T}\sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \tag{4}$$

where c is the size of training context. The term $p(w_t | w_{t-c}, \cdots, w_{t-1}, w_{t+1}, \cdots, w_{t+c})$ and $p(w_{t+j} | w_t)$ are defined by the softmax function:

$$p(w_{Output} | w_{Input}) = \frac{\exp(v'^{T}_{w_{Output}} v_{w_{Input}})}{\sum_{w=1}^{W} \exp(v'^{T}_{w_{Output}} v_{w_{Input}})} \tag{5}$$

where v_w and v'_w are the input and output vector representation of the word w and W is the number of words in the vocabulary list. This softmax function tries to predict the output words w_{Output} given the input words w_{Input}. However, this formulation is impractical because of the cost when computing the gradient is proportional to W which is typically very large. So, the hierarchical softmax and negative sampling are proposed to compute the approximation to the softmax function instead.

4.4.3 Word2Vec training

The hierarchical softmax [139] reduces the number of nodes to be evaluated from W to $\log_2(W)$ nodes by defining a binary tree over W as leaves and the relative probabilities of W as the intermediate nodes. This creates a random walk model over W. That is, there exists a path from the root of the tree to each leaf node w. Let $n(w, j)$ be the

j^{th} node on that path and $L(w)$ denotes the length of the path. Also, let $ch(n)$ be an arbitrary fixed child node of n and an operation $[\![x]\!]$ will return 1 if x is true and -1 otherwise. The $p(w_{Output}|w_{Input})$ of hierarchical softmax is as follows,

$$\prod_{j=1}^{L(w)-1} \Sigma\Big([\![n(w, j+1) = ch(n(w,j))]\!] \cdot v'^{T}_{n(w,j)} v_{w_{Input}}\Big), \tag{6}$$

where $\Sigma = 1/(1 + exp(-x))$. This hierarchical softmax has only one representation for each word v_w and one representation for each intermediate node n as v'_n.

The Negative sampling (NEG) simplifies Noise Contrastive Estimation (NCE) [84] which is an alternative to hierarchical softmax that approximates the log probability of the softmax model. NEG concerns only with the quality of the word representation rather than differentiate data from noises. The term $p(w_{Output}|w_{Input})$ of NEG is as follows,

$$\log\sigma(v'^{T}_{w_{Output}} v_{w_{Input}}) + \sum_{i=1}^{K} \mathbb{E}_{w_i \sim P_n(w)} \log\sigma(-v'^{T}_{w_{Output}} v_{w_{Input}}) \tag{7}$$

which is a replacement to $\log p(w_{Output}|w_{Input})$ in the CBOW and Skip-gram objective functions. NEG tries to distinguish w_{Output} from noises $P_n(w)$ using logistic regression where k is the number of negative samples for each data sample. $P_n(w)$ is a free parameter in which [132] suggests to use the unigram distribution $U(w)$ raised to the 3/4rd power as the best choice. From the original experiment in [132], the Skip-gram model with NEG provides the best word embedding. In addition, the Skip-gram model with NEG was shown to be equivalent to factorizing the PMI word cooccurence matrix [111]. [156, 79, 56] provides a more detailed explanation of Word2Vec training methods.

4.4.4 State-of-the-art Word2Vec-based models

There are a lot of interesting follow up works which try to improve Word2Vec in various ways. For example, GloVe [147] tries to incorporate global word cooccurence information. DEPs [110] tries to include syntactic dependency information. MCE [42] uses pairwise ranking loss function in the learning to rank framework to model antonyms on an antonym word list. SPs [164] creates its word embedding based on symmetric patterns extracted from corpora such as “X such as Y” or “X is a Y” or “X and Y” which also enables antonym modeling. Modeling lexical contrast [135] is an important contribution because it solves the fundamental problem of cooccurence between the target word and its synonym or antonym which previously was a blind spot of modeling distributional semantics by word cooccurence. These works provide another breakthrough in distributional semantics by word embedding.

Other interesting works try to make an improvement to word embedding in term of interpretability because word embedding models typically output dense vectors that cannot be easily interpreted by human. [64, 199] tries to add interpretability into a word embedding by incorporating sparsity using sparse coding methods. AutoExtend [157] can extend existing word embedding from synsets and lexemes to make a better embedding. Retrofitting [63] is a graph-based method which also extends an existing word embedding. Word2Vec triggered a lot of interest from its high quality word representation output so that it creates a renaissance of word embedding research.

Other interesting works try to generalize word embedding to logic for better reasoning. Traditionally, this was done by the method of binding roles and fillers by using operations such as tensor product [44]. Roles are logical predicates and fillers are logical atoms which are both represented as vectors. Recently, there are efforts that try to map boolean structures to distributional semantics for recognizing textual entailment (RTE) which decides the entailment between two sentences. The proposed approaches are Markov Logic Networks [21, 76] and learning a mapping function with BDSM [100].

4.5 Bag-of-Visual-Words and Image Embedding

4.5.1 Bag-of-Visual Words (BoVW)

In Computer Vision (CV), the idea of bag-of-words representation (BoW) was long borrowed from Natural Language Processing (NLP) community in solving recognition tasks under the name of bag-of-visual-words representation (BoVW). BoW representation discards spatial and temporal relations between words and creates a representation of a document based on only word frequencies which outputs the term-document matrix. Similarly, BoVW discards location and shape information.

For an image, BoVW representation is a descriptor-image matrix where descriptors are local descriptors in the visual codebook. The descriptors are salient keypoints of an image extracted by using techniques such as SIFT [116] or SURF [17] which can reliably find descriptors across images under difficulties like rotation, translation or illumination changes. Then, the descriptors are clustered together by a clustering algorithm such as k-means [117] to create a visual codebook. The reason is that there are varying numbers of visual descriptors in each image unlike text documents whose words come off-the-shelf so the codebook step is needed for visual data. Thus, the clustering step is needed in order to make the frequency histogram comparable across images by fixing the codewords. This BoVW model still does not go beyond point descriptors to edges or lines. The local descriptors will correspond to image patches with similar appearance but may not be correlated with the object-level parts in an image [80].

Some landmark works incorporate location and shape information into BoVW model and achieve a test-of-the-time result like Spatial Pyramid Matching (SPM) [108] or Constellation model [65].

4.5.2 Image Embedding

Since natural images lie in a low dimensional manifold in the space of all possible images, the efforts to model that manifold result in image embedding techniques [150]. Image embedding is similar to word representation because it is also represented as a dense low dimensional feature vector. Besides, the images that are close to each other are similar and each dimension captures factors of variations in the images such as pose, illumination or translations.

The image embedding is an output or intermediate output from a representation learning methods such as dimensionality reduction methods [187] including deep learning techniques [22]. One of the most dominant examples for image embedding is in face recognition [205]. Eigenfaces [185] uses Principal Component Analysis (PCA) which is analogous to Singular Value Decomposition (SVD) used in LSA to create a projection to a low dimensional image manifold which represent faces. Fisherfaces [19] also uses a dimensionality reduction method, namely Fisher's Linear Discriminant Analysis (FLD) [68], to compute a better projection. There are more creative dimensionality reduction techniques based on SVD, for instance, ISOMAP [179], Locally Linear Embedding (LLE) [158] or Laplacian Eigenmaps [20] can capture nonlinear variations in images such as pose or facial expressions. These techniques fall into the category of Spectral Methods which is still an interesting ongoing research [3, 4, 5] in Machine Learning.

4.5.3 State-of-the-art Image Embedding models

Both BoVW and image embedding are used as a feature set for classification mainly for recognition tasks but are not limited to them. For example, Recursive Neural Networks (RNN) was applied to another recognition task of semantic segmentation in the context of scene parsing [169]. Recently, image embedding from deep Convolutional Neural Networks (CNNs) [109] which exhibits similar characteristics to Word2Vec [71] is applied in various tasks both in Reorganization (like Optical Flow [67]) and Recognition using Siamese CNNs (like visual analogy [160]). Moreover, the CNN models from Caffe [95], including models for recognition like AlexNet [99] or GoogLeNet [174] or VGG net [167] and models for detection like R-CNN [78] or Fast R-CNN [77], tend to be the current state-of-the-art image embedding models.

In short, one can conclude that performing representation learning on image data will result in image embedding similar to Word2Vec. However, image embedding is likely to be more domain specific and has more data set bias [182]. Even though it is trained on a data set of millions or billions images like AlexNet [99] and provides a breakthrough in recognition on the ImageNet LSVRC-2010 challenge [25], the coverage of real world objects is just around $1,000$ categories and is still far from learning from text alone like training Word2Vec on Google's Billion Words corpus which has $793,471$ vocabulary words [38]. For face recognition, the recently proposed DeepFace [175] model was created to recognize around $4,000$ identities on the Labelled Face in the Wild (LFW) data set [92] which is very remarkable but still far from a system that can recognize everybody's faces on the fly.

Solving this problem by learning on a larger scale image collection of size trillions or zillions, such as learning from the internet [41], to provide a general image embedding like Word2Vec which also models infrequent data, is an interesting future direction.

References

[1] Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, Kejun Ning, Babette Dellen, and Florentin Wörgötter. Learning the semantics of object–action relations by observation. *The International Journal of Robotics Research*, page 0278364911410459, 2011.

[2] Yiannis Aloimonos and Cornelia Fermüller. The cognitive dialogue: A new model for vision implementing common sense reasoning. *Image and Vision Computing*, 34:42–44, 2015.

[3] Anima Anandkumar, Yi-kai Liu, Daniel J Hsu, Dean P Foster, and Sham M Kakade. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 917–925, 2012.

[4] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[5] Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. *arXiv preprint arXiv:1505.00468*, 2015.

[7] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[8] Yoav Artzi and Luke Zettlemoyer. UW SPF: The University of Washington Semantic Parsing Framework, 2013.

[9] Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. Abstract meaning representation (amr) 1.0 specification. In *Parsing on Freebase from Question-Answer Pairs." In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle: ACL*, pages 1533–1544, 2012.

[10] Albert Bandura. *Psychological modeling: Conflicting theories.* Transaction Publishers, 1974.

[11] Chitta Baral, Juraj Dzifcak, Kanchan Kumbhare, and Nguyen H Vo. The nl2kr system. *Language Processing and Automated Reasoning (NLPAR) 2013*, 2013.

[12] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M Blei, and Michael I Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003.

[13] Kobus Barnard and David Forsyth. Learning the semantics of words and pictures. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 408–415. IEEE, 2001.

[14] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 238–247, 2014.

[15] F. Barranco, C. Fermuller, and Y. Aloimonos. Contour motion estimation for asynchronous event-driven cameras. *Proceedings of the IEEE*, 102(10):1537–1556, Oct 2014.

[16] Jonathan Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(8):1670–1687, 2015.

[17] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006.

[18] Michael Beetz, Suat Gedikli, Jan Bandouch, Bernhard Kirchlechner, Nico von Hoyningen-Huene, and Alexander Perzylo. Visually tracking football games based on tv broadcasts. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2007.

[19] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[20] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[21] Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J Mooney. Representing meaning with a combination of logical form and vectors. *arXiv preprint arXiv:1505.06816*, 2015.

[22] Yoshua Bengio, Aaron Courville, and Pierre Vincent. Representation learning: A review and new perspectives. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1798–1828, 2013.

[23] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[24] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, pages 1533–1544, 2013.

[25] A Berg, J Deng, and L Fei-Fei. Large scale visual recognition challenge (ilsvrc), 2010. *URL http://www. image-net. org/challenges/LSVRC*, 2010.

[26] Tamara Berg and Alexander C Berg. Finding iconic images. In *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2009.

[27] Tamara L Berg, Alexander C Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A Forsyth. Names and faces in the news. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–848. IEEE, 2004.

[28] Tamara L Berg, Alexander C Berg, and Jonathan Shih. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*, pages 663–676. Springer, 2010.

[29] Tamara L Berg, David Forsyth, et al. Animals on the web. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1463–1470. IEEE, 2006.

[30] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[31] David M Blei and Michael I Jordan. Modeling annotated data. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134. ACM, 2003.

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

[33] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. Three-dimensional face recognition. *International Journal of Computer Vision*, 64(1):5–30, 2005.

[34] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *J. Artif. Intell. Res.(JAIR)*, 49:1–47, 2014.

[35] Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. Generating instructions in virtual environments (give): A challenge and an evaluation testbed for nlg. 2007.

[36] Angelo Cangelosi. The grounding and sharing of symbols. *Pragmatics & Cognition*, 14(2):275–285, 2006.

[37] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3241–3248. IEEE, 2010.

[38] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[39] Anthony Chemero. An outline of a theory of affordances. *Ecological psychology*, 15(2):181–195, 2003.

[40] David L Chen and Raymond J Mooney. Learning to sportscast: a test of grounded language acquisition. In *Proceedings of the 25th international conference on Machine learning*, pages 128–135. ACM, 2008.

[41] Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. Neil: Extracting visual knowledge from web data. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1409–1416. IEEE, 2013.

[42] Zhigang Chen, Wei Lin, Qian Chen, Xiaoping Chen, Si Wei, Hui Jiang, and Xiaodan Zhu. Revisiting word embedding for contrasting meaning. In *Proceedings of ACL*, 2015.

[43] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012.

[44] Stephen Clark and Stephen Pulman. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*, pages 52–55, 2007.

[45] Silvia Coradeschi, Amy Loutfi, and Britta Wrede. A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intelligenz*, 27(2):129–136, 2013.

[46] Silvia Coradeschi and Alessandro Saffiotti. Anchoring symbols to sensor data: preliminary report. In *AAAI/IAAI*, pages 129–135, 2000.

[47] Christopher Crick, Marek Doniec, and Brian Scassellati. Who is it? inferring role and intent from agent motion. In *Development and Learning, 2007. ICDL 2007. IEEE 6th International Conference on*, pages 134–139. IEEE, 2007.

[48] Trevor Darrell. Learning representations for real-world recognition, 7 2010. UCB EECS Colloquium [Accessed: 2015 11 1].

[49] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[50] Hal Daumé III, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine learning*, 75(3):297–325, 2009.

[51] Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 762–772. Association for Computational Linguistics, 2012.

[52] Kais Dukes. Semeval-2014 task 6: Supervised semantic parsing of robotic spatial commands. *SemEval 2014*, page 45, 2014.

[53] Susan T Dumais. Lsa and information retrieval: Getting back to basics. *Handbook of latent semantic analysis*, pages 293–321, 2007.

[54] Hugh Durrant-Whyte and Tim Bailey. Simultaneous localization and mapping: part i. *Robotics & Automation Magazine, IEEE*, 13(2):99–110, 2006.

[55] Pinar Duygulu, Kobus Barnard, Joao FG de Freitas, and David A Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision—ECCV 2002*, pages 97–112. Springer, 2002.

[56] Chris Dyer. Notes on noise contrastive estimation and negative sampling. *arXiv preprint arXiv:1410.8251*, 2014.

[57] Desmond Elliott and Frank Keller. Image description using visual dependency representations. In *EMNLP*, pages 1292–1302, 2013.

[58] Oren Etzioni, Michele Banko, and Michael J Cafarella. Machine reading. In *AAAI*, volume 6, pages 1517–1519, 2006.

[59] Ali Farhadi. Designing representational architectures in recognition. 2011.

[60] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2352–2359. IEEE, 2010.

[61] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010.

[62] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[63] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

[64] Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004*, 2015.

[65] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[66] Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. On available corpora for empirical methods in vision & language. *CoRR*, abs/1506.06833, 2015.

[67] Philipp Fischer, Alexey Dosovitskiy, Eddy Ilg, Philip Häusser, Caner Hazırbaş, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. *arXiv preprint arXiv:1504.06852*, 2015.

[68] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[69] Gottlob Frege and John Langshaw Austin. *The Foundations of Arithmetic: A logico-mathematical enquiry into the concept of number.* Northwestern University Press, 1980.

[70] Jianlong Fu, Jinqiao Wang, Xin-Jing Wang, Yong Rui, and Hanqing Lu. What visual attributes characterize an object class? In *Computer Vision–ACCV 2014*, pages 243–259. Springer, 2015.

[71] D Garcia-Gasulla, J Béjar, U Cortés, E Ayguadé, and J Labarta. Extracting visual patterns from deep learning representations. *arXiv preprint arXiv:1507.08818*, 2015.

[72] Peter Gärdenfors. *The geometry of meaning: Semantics based on conceptual spaces.* MIT Press, 2014.

[73] Konstantina Garoufi. Planning-based models of natural language generation. *Language and Linguistics Compass*, 8(1):1–10, 2014.

[74] Konstantina Garoufi and Alexander Koller. Automated planning for situated natural language generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1573–1582. Association for Computational Linguistics, 2010.

[75] Konstantina Garoufi, Maria Staudte, Alexander Koller, and Matthew W Crocker. Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science*, 2015.

[76] Dan Garrette, Katrin Erk, and Raymond Mooney. A formal approach to linking logical form and vector-space lexical semantics. In *Computing Meaning*, pages 27–48. Springer, 2014.

[77] Ross Girshick. Fast r-cnn. *arXiv preprint arXiv:1504.08083*, 2015.

[78] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[79] Yoav Goldberg and Omer Levy. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[80] Kristen Grauman and Bastian Leibe. *Visual object recognition*. Number 11. Morgan & Claypool Publishers, 2010.

[81] Douglas Greenlee. Semiotic and significs. *International Studies in Philosophy*, 10:251–254, 1978.

[82] Gutemberg Guerra-Filho and Yiannis Aloimonos. A language for human action. *Computer*, 40(5):42–51, 2007.

[83] Abhinav Gupta. Beyond nouns and verbs. 2009.

[84] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13(1):307–361, 2012.

[85] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2014.

[86] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990.

[87] Zellig S Harris. Distributional structure. *Word*, 1954.

[88] Geremy Heitz and Daphne Koller. Learning spatial context: Using stuff to find things. In *Computer Vision–ECCV 2008*, pages 30–43. Springer, 2008.

[89] Geoffrey E Hinton. Distributed representations. 1984.

[90] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[91] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.

[92] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[93] Julian Jaynes. *The origin of consciousness in the breakdown of the bicameral mind*. Houghton Mifflin Harcourt, 2000.

[94] Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 119–126. ACM, 2003.

[95] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[96] Benjamin Johnston, Fangkai Yang, Rogan Mendoza, Xiaoping Chen, and Mary-Anne Williams. Ontology based object categorization for robots. In *Practical Aspects of Knowledge Management*, pages 219–231. Springer, 2008.

[97] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*, 2014.

[98] Alexander Koller and Matthew Stone. Sentence generation as a planning problem. *ACL 2007*, page 336, 2007.

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

[100] German Kruszewski, Denis Paperno, and Marco Baroni. Deriving boolean structures from distributional vectors. *Transactions of the Association for Computational Linguistics*, 3:375–388, 2015.

[101] Gaurav Kulkarni, Visruth Premraj, Vicente Ordonez, Sudipta Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara Berg. Babytalk: Understanding and generating simple image descriptions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2891–2903, 2013.

[102] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[103] Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 359–368. Association for Computational Linguistics, 2012.

[104] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[105] Kevin Lai and Dieter Fox. Object recognition in 3d point clouds using web data and domain adaptation. *The International Journal of Robotics Research*, 29(8):1019–1037, 2010.

[106] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[107] Victor Lavrenko, R Manmatha, and Jiwoon Jeon. A model for learning the semantics of pictures. In *Advances in neural information processing systems*, page None, 2003.

[108] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[109] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[110] Omer Levy and Yoav Goldberg. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308, 2014.

[111] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[112] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225, 2015.

[113] Mike Lewis and Mark Steedman. Combining distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192, 2013.

[114] Siming Li, Girish Kulkarni, Tamara L Berg, Alexander C Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 220–228. Association for Computational Linguistics, 2011.

[115] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[116] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[117] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.

[118] ASM Ashique Mahmood. Literature survey on topic modeling.

[119] Jitendra Malik. The three r's of vision, 7 2013. ENS/INRIA Visual Recognition and Machine Learning Summer School [Accessed: 2015 10 25].

[120] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. *arXiv preprint arXiv:1503.01558*, 2015.

[121] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[122] David Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* Henry Holt and Co., Inc., New York, NY, USA, 1982.

[123] Cynthia Matuszek*, Nicholas FitzGerald*, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.

[124] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer, 2013.

[125] Nikolaos Mavridis. A review of verbal and non-verbal human–robot interactive communication. *Robotics and Autonomous Systems*, 63:22–35, 2015.

[126] Nikolaos Mavridis and Deb Roy. Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 4690–4697. IEEE, 2006.

[127] Julian McAuley, Rahul Pandey, and Jure Leskovec. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2015.

[128] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[129] Jon D Mcauliffe and David M Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[130] P.J.M. Mengibar and M.E. Epstein. Speech and semantic parsing for content selection, October 8 2015. US Patent App. 13/844,312.

[131] Chet Meyers and Thomas B Jones. *Promoting Active Learning. Strategies for the College Classroom.* ERIC, 1993.

[132] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[133] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[134] Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics, 2012.

[135] Saif M Mohammad, Bonnie J Dorr, Graeme Hirst, and Peter D Turney. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590, 2013.

[136] Raymond J Mooney. Learning to connect language and perception. In *AAAI*, pages 1598–1601, 2008.

[137] Raymond J. Mooney. Grounded language learning, 7 2013. 27th AAAI Conference on Artificial Intelligence, Washington 2013 [Accessed: 2015 11 1].

[138] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, pages 1–9. Citeseer, 1999.

[139] Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In *Proceedings of the international workshop on artificial intelligence and statistics*, pages 246–252. Citeseer, 2005.

[140] Charles William Morris. Foundations of the theory of signs. 1938.

[141] Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. From large scale image categorization to entry-level categories. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[142] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, pages 1143–1151, 2011.

[143] Devi Parikh. Modeling context for image understanding: When, for what, and how? 2009.

[144] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[145] Katerina Pastra and Yiannis Aloimonos. The minimalist grammar of action. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1585):103–117, 2012.

[146] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

[147] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.

[148] Jean Piaget. *Play, dreams and imitation in childhood*, volume 25. Routledge, 2013.

[149] Tony Plate. A common framework for distributed representation schemes for compositional structure. *Connectionist systems for knowledge representation and deduction*, pages 15–34, 1997.

[150] Robert Pless and Richard Souvenir. A survey of manifold learning for images. *IPSJ Transactions on Computer Vision and Applications*, 1:83–94, 2009.

[151] Jordi Pont-Tuset, Pablo Arbelaez, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping for image segmentation and object proposal generation. *arXiv preprint arXiv:1503.00848*, 2015.

[152] Hoifung Poon and Pedro Domingos. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, pages 1–10. Association for Computational Linguistics, 2009.

[153] Michael Pust, Ulf Hermjakob, Kevin Knight, Daniel Marcu, and Jonathan May. Parsing english into abstract meaning representation using syntax-based machine translation. *EMNLP 2015*, pages 1143–1154, 2015.

[154] Li Ren. A survey on statistical topic modeling.

[155] Giacomo Rizzolatti and Laila Craighero. The mirror-neuron system. *Annu. Rev. Neurosci.*, 27:169–192, 2004.

[156] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[157] Sascha Rothe and Hinrich Schütze. Autoextend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the ACL*, 2015.

[158] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[159] Deb Roy. Grounding words in perception and action: computational insights. *TRENDS in Cognitive Sciences*, 9(8):390, 2005.

[160] Fereshteh Sadeghi, C Lawrence Zitnick, and Ali Farhadi. Visalogy: Answering visual analogy questions. In *Advances in Neural Information Processing Systems (NIPS-15)*, 2015.

[161] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE, 2011.

[162] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[163] Bernhard Schölkopf, Alexander J Smola, Ben Taskar, and SVN Vishwanathan. Predicting structured data, 2006.

[164] Roy Schwartz, Roi Reichart, and Ari Rappoport. Symmetric pattern based word embeddings for improved word similarity prediction. *CoNLL 2015*, page 258, 2015.

[165] J Sherwani, Dong Yu, Tim Paek, Mary Czerwinski, Yun-Cheng Ju, and Alex Acero. Voicepedia: towards speech-based access to unstructured information. In *INTERSPEECH*, volume 7, pages 146–149, 2007.

[166] Nishant Shukla, Caiming Xiong, and Song-Chun Zhu. A unified framework for human-robot knowledge transfer. In *2015 AAAI Fall Symposium Series*, 2015.

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

[168] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[169] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011.

[170] Mark Steedman. Surface structure and interpretation. 1996.

[171] Mark Steyvers and Tom Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440, 2007.

[172] Douglas Summers-Stay, Ching L Teo, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Using a minimal action grammar for activity understanding in the real world. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pages 4104–4111. IEEE, 2012.

[173] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 2096–2103. IEEE, 2013.

[174] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[175] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701–1708. IEEE, 2014.

[176] Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.

[177] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems, Berkeley, USA*, 2014.

[178] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, 2011.

[179] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[180] Ching L Teo, Yezhou Yang, Hal Daumé III, Cornelia Fermüller, and Yiannis Aloimonos. Towards a watson that sees: Language-guided action recognition for robots. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 374–381. IEEE, 2012.

[181] Sebastian Thrun, Wolfram Burgard, and Dieter Fox. *Probabilistic robotics*. MIT press, 2005.

[182] Antonio Torralba, Alexei Efros, et al. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011.

[183] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005.

[184] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[185] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[186] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188, 2010.

[187] Laurens JP van der Maaten, Eric O Postma, and H Jaap van den Herik. Dimensionality reduction: A comparative review. *Journal of Machine Learning Research*, 10(1-41):66–71, 2009.

[188] Bernard Vauquois. Structures profondes et traduction automatique. le système du ceta. *Revue Roumaine de linguistique*, 13(2):105–130, 1968.

[189] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*, 2014.

[190] Matthew R. Walter, Matthew E. Antone, Ekapol Chuangsuwanich, Andrew Correa, Randall Davis, Luke Fletcher, Emilio Frazzoli, Yuli Friedman, James R. Glass, Jonathan P. How, Jeong Hwan Jeon, Sertac Karaman, Brandon Luders, Nicholas Roy, Stefanie Tellex, and Seth J. Teller. A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments. *J. Field Robotics*, 32(4):590–628, 2015.

[191] Chong Wang, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

[192] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. A large-scale car dataset for fine-grained categorization and verification. June 2015.

[193] Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE, 2013.

[194] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*, 2015.

[195] Yezhou Yang, Ching L Teo, Cornelia Fermuller, and Yiannis Aloimonos. Robots with language: Multi-label visual recognition using nlp. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 4256–4262. IEEE, 2013.

[196] Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 444–454. Association for Computational Linguistics, 2011.

[197] Yang Yezhou, Yiannis Aloimonos, Cornelia Fermüller, and Eren Erdal Aksoy. Learning the semantics of manipulation action. In *Association for Computational Linguistics (ACL)*, 2015.

[198] Wen-tau Yih, Xiaodong He, and Christopher Meek. Semantic parsing for single-relation question answering. In *Proceedings of ACL*, 2014.

[199] Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah A Smith. Learning word representations with hierarchical sparse coding. *arXiv preprint arXiv:1406.2035*, 2014.

[200] Nivasan Yogeswaran, Wenting Dang, William Taube Navaraj, Dhayalan Shakthivel, Saleem Khan, Emre Ozan Polat, Shoubhik Gupta, Hadi Heidari, Mohsen Kaboli, Leandro Lorenzelli, et al. New materials and advances in making electronic skin for interactive robots. *Advanced Robotics*, 29(21):1359–1373, 2015.

[201] Xiaodong Yu, Cornelia Fermuller, Ching Lik Teo, Yezhou Yang, and Yiannis Aloimonos. Active scene recognition with vision and language. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 810–817. IEEE, 2011.

[202] John M Zelle and Raymond J Mooney. Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1050–1055, 1996.

[203] Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.

[204] Rong Zhao and William I Grosky. Bridging the semantic gap in image retrieval. *Distributed multimedia databases: Techniques and applications*, pages 14–36, 2002.

[205] Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399–458, 2003.