UNIVERSITÀ DEGLI STUDI DI TRENTO

CIMeC - Center for Mind/Brain Sciences

Masters of Science in Cognitive Science
as part of the
Erasmus Mundus European Masters Program in Language and Communication Technologies

Playing with Properties:
Visual Concept Representation with Semantically-derived Text-extracted Properties

Supervisor:
Marco Baroni
Co-supervisor:
Gosse Bouma

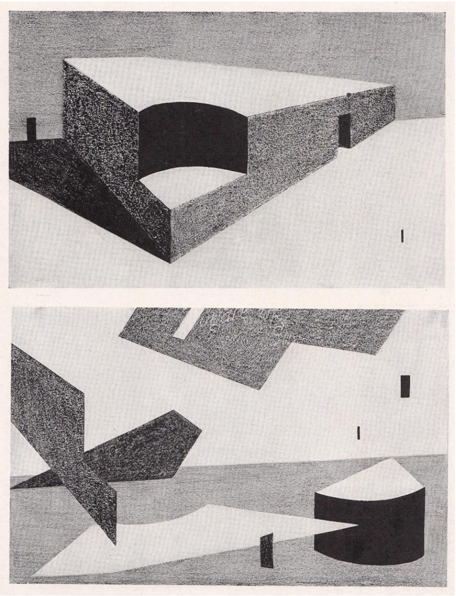
Candidate:
Kim Heiligenstein

Academic Year 2013-2014

KIM HEILIGENSTEIN

PLAYING WITH PROPERTIES

Visual Concept Representation with Semantically-derived Text-extracted Properties



A thesis submitted in partial fulfillment of the degree of

Masters of Science in Cognitive Science
as part of the
Erasmus Mundus European Masters Program in
Language and Communication Technologies

Rijksuniversitat Groningen & Università degli Studi di Trento

July 17th, 2014

E-MAIL:
kim.heiligenstein@gmail.com

The titlepage reproduces a work of Adeline Cross, titled *Study of the advancing and receding space qualities of tone values*.

ABSTRACT

Computer vision models such as that of Farhadi et al. [15] allow us to reach property-based concept representations using unsupervised visual feature-selection methods. The set back is that visual properties are too often not generalizable, and do not properly reflect the way conceptual knowledge is acquired and represented in the mind. In contrast, a more semantically sound approach has been developed using computational linguistic methods, namely those employed by the Strudel [6] model. We suggest using property-based concept descriptions automatically extracted from a corpus of naturally-occurring text to train an image-based concept classification and annotation model to arrive at meaning representations endowed with stronger cognitive qualities. The discussion consists in a qualitative data analysis which encourages the idea that these corpus-harvested properties are in fact plausible candidates to achieve conceptual knowledge grounded in visual perception.

ACKNOWLEDGEMENTS

I wish to express my gratitude and appreciation to my supervisors, Marco Baroni and Gosse Bouma, for their guidance and advice during the whole of the project. I would also like to thank Raffaella Bernardi for her help in keeping me on top of my work, and reminding me of my potential.

A very special acknowledgement is owed to Elia Bruni, without whose direction, support and encouragement throughout the development of the project and beyond, would have made the completion of this work impossible. Jiming Li, thank you for the contribution and assistance you've offered me during the experiments.

I cannot forget my family and friends, near and far, for their continuous encouragement, and for providing me with the time-outs essential to a healthy work pattern.

Finally, my greatest gratitude goes out to my roommate, for supplying me with the motivation to spend day and night working at the library rather than staying at home, subject to insupportable desk drumming and sudden verbal outbursts of anger.

CONTENTS

1 INTRODUCTION

1.1 BACKGROUND

Theories about how concepts are represented in the mind have often adopted a 'distributed, feature-based model of conceptual knowledge' [27]. These grew as an extension of the previous frequency co-occurrence approaches [30, 31, 11] based on the distributional hypothesis that words that appear in similar contexts will share a similar meaning [23]. An attribute-centric approach to concept descriptions has found support in both the computational linguistic community, using both human-generated norms [33] and text-extracted features [25, 1, 6], and more recently in the computer vision field [15, 9, 40].

Computer vision models such as that of Farhadi et al. [15] extend the goal of object recognition to object description, using an unsupervised feature selection method for learning predicted properties from labeled images. The system is successful in that it not only aptly categorizes the target objects, but proves that selecting features and learning classifiers from textual annotations can lead to describing unknown objects and report unusual properties. The properties are interesting because they are not just discriminative, but semantic as well, and define relations such as parts, material and shapes.

Although the focus of their model is to learn object descriptions that generalize well across categories, the classifiers are trained on a set of human-generated, task-oriented annotations. This is a disadvantage in that often, these properties are not representative of the full range of features that a concept represents mainly because the participants only provide a list of features that are generated for the purpose of the task, and omit those that are the most helpful in object discrimination [33]. In spite of the fact that they are reliable and a good reference for feature-based concept representation experiments, the properties are not generated in an automatic fashion, while the rest of the model boasts an unsupervised feature selection method. There are, however, ways in which properties *can* be generated in an unsupervised manner and still retain their semantic characteristics.

An attribute-centric approach to concept descriptions has been adopted by computational linguistic technologies as well, namely by Strudel [6], a corpus-based distributional semantic model that yields structured and comprehensive sets of concept descriptions. The Strudel model automatically extracts concept-property pairs from a corpus of naturally-occurring text from relation pattern templates using possible part-of-speech sequences. Strudel differs from other models in that the collected properties provide strong semantic qualities because they tend to focus on activities and interactions rather than parts and physical attributes.

1.2 PROPOSAL

Although borrowing from both language and vision is not a novel venture [10, 2, 40], the multimodal model presented here takes a different, more cognitive approach to visual concept representation. The goal is to construct a perceptually grounded, linguistically enhanced, distributional model that does not learn exclusively from vision or language, but from both. In order to extract a unified representation of both modalities, semantically-derived text-extracted concept descriptions are used to train an image-based concept classification and annotation model to arrive at representations endowed with stronger semantic qualities.

Since visual models learn classifiers based on the information it is fed, the goal is to shift their focus to learn from a set of automatically extracted text-based properties, instead of ad-hoc human-generated norms. The visual model is inspired by that of Farhadi et al. [15], to reach the goal of property-based concept representation, but instead of learning from the list of task-driven features, swap in the automatically retrieved text ones from the Strudel experiment. Again, the motivation of using text-extracted information is supported by the idea that they are conceptually close to human-generated norms, as seen in the Baroni et al. [6] experiments, but differ from them in that they are acquired in an unsupervised manner and cover a more complete set of features.

1.3 OUTLINE OF PAPER

In the next chapter, I will give a review of the literature pertaining to the scientific fields that have adopted the distributive hypothesis, and demonstrate how they have implemented it backed by findings from past studies. I will then draw support from the related work to present the proposed experiment in chapter 3. In chapter 4, I will give a qualitative analysis of the results, followed by a discussion and conclusion in chapter 5.

2 REVIEW OF LITERATURE

2.1 OVERVIEW

In this chapter, I will address the question of how humans deal with the acquisition and representation of conceptual knowledge in terms of the distributional hypothesis. I will introduce what it proposes, and give an account of how it is applied in varied scientific fields. In linguistics, the distributional hypothesis is often paraphrased as Firth [17]'s famous quote, "you shall know a word by the company it keeps", proposing that the meaning of a word can be characterized by its most typical collocates [14] and by the various circumstances of its common usage [3]. It is implemented in computational linguistics to approximate word meaning in text, using multidimensional feature vectors, whose values describe its context in terms of the distribution of words and phrases it commonly occurs with. Computer vision scientists have adopted it to reach conceptual knowledge visually, extending the distributional approach to models that use visual information extracted from images as values for low-level feature vectors. I will also touch on how the distributional hypothesis plays a role in neuroscience, such that the meaning representation of a word can be characterized by different spatial patterns of neural activation. The foundation for the these findings is reported in the studies below.

Finally, I use the fact that fields concerned with the same goal of reaching meaning representation are tied by the distributional hypothesis, and support the proposed idea of a multiple modality model that borrows from both language and vision.

2.2 THE DISTRIBUTIONAL HYPOTHESIS

The distributional hypothesis in linguistics proposes that words that appear in similar contexts tend to share a similar meaning [23]. This idea can be further extended to explain the behavior and usage of a word based on its distributional context which also helps addresses the issue of word sense disambiguation, where the sense of a polysemous word could be revealed by the context in which it occurs [17]. Furthermore, it has been suggested that distributional semantics play the lead role in the induction and representation of knowledge, as proposed by Landauer and Dumais [30], who consider semantic similarity a key component in explaining how we acquire knowledge.

This idea of semantic similarity is translated in computational linguistics as vectors in a high-dimensional semantic space to approximate word meaning. Real language text corpora serve as the source for all extractable information, and therefore allow for an investigation of how the statistics of language influence semantic representation [16], without being limited by the collection of human data. Given a corpus, feature vectors are constructed for target words where their values are quantified information from the context in which the word occurs. Vector space models (VSMs) are used to compare words as points in space, computing their similarity using standard distance measures such as the cosine of the angle between two vectors [10, 44, 12, 44]. To use the example provided by Bruni et al. [10], since *car* and *automobile* occur in similar contexts, meaning they are most often surrounded by the same words, such as *street*, *gas*, and *driver*, they will have comparably populated vectors, suggesting that these two words have similar meanings. This method of using distributional models to quantify word similarity by means of vectors is very useful for applications such as document retrieval and classification, automated thesaurus or bilingual dictionary construction and sentiment analysis, amongst many.

Under the assumption of the distributional hypothesis, words can be represented by a vector denoting its 'context signature' [6]. These signatures can be populated using various approaches. Bullinaria [11] uses basic word co-occurrences, counting the number of times each collocate appears in a window of a particular size around each target word. Evert [14] contrasts this surface co-occurrence with syntactic co-occurrence, an approach he describes as more restrictive in that context words are counted only if they occur in a direct syntactic relation with the target word, such as verb-object, or prenominal adjectives. A different approach to representing word meaning has been explored by topic models, that still use contextual information from corpora, but word meaning is a probability distribution over a set of topics, and each topic over words [16]. Further methods of exploring the context will be discussed below.

2.2.1 Distributional Semantics with Corpus-based Semantic Models

A class of computational methods known as corpus-based semantic models (CSMs) are increasingly employed as a tool to gain insight into the semantic knowledge representation of humans [6]. CSMs take corpora of prepared or naturally occurring linguistic data to derive two types of information from which semantic representations can be learned [3]. *Distributional* data describe the statistical distribution of words across texts in a corpus. *Experiential* data is knowledge about concepts based on their interaction with the world.

CSMs are of interest to us because they are recognized for their ability to aptly model important processes in human cognition and language acquisition, as they have been shown to emphasize the role of learning from simple statistical and distributional cues [28]. They are a good resource in that, like humans, they are also faced with noise and scarcity of explicit and coherent information when dealing with data that consists of large and mixed collections of texts. They have also been found to encounter the same problems as we do in acquiring conceptual knowledge and conceptual categorization [6]. Fortunately, they are in some ways *un*like humans, for whom collecting relevant information on very large scales is extremely time and energy consuming when done manually [44, 13]. Their efficient and robust approach to natural language processing and concept representation from a strong semantic perspective has allowed them to play an important role in practical tasks such as information retrieval and intelligent tutoring systems [28].

There is an infinite number of ways to analyze a corpus. The large variety of CSMs reflect the wide range of information there is to extract, the methods to extract it, and the ways in which it can be interpreted. Here are a few that are worth mentioning.

Notable CSMs

HYPERSPACE ANALOGUE TO LANGUAGE (HAL) HAL is Lund and Burgess [31]'s model for simulating human semantic memory and representing the meanings of words. They use lexical co-occurrence to build high-dimensional semantic spaces and demonstrate how these spaces can model human concept similarity. With a corpus of text, they employ an *n*-window method to automatically extract contextual information of a given word to build feature vectors. These vectors are then analyzed in terms of similarity and multidimensional scaling to examine relationships between words. In their paper, they boast the advantages of employing lexical co-occurrence and position similarity to capture information about word meanings not only in terms of similarity, but also association.

LATENT SEMANTIC ANALYSIS (LSA) Landauer and Dumais [30]'s LSA is another unsupervised high-dimensional linear associative model that captures the similarity of words and documents. Their methods differ from those of HAL in that they rely on an inductive mechanism of dimension optimization. Their model is heavily dependent on the distributional hypothesis as well, using distance in the semantic space and relative frequency of co-occurrence to calculate word similarity. They are aware of the noise and address other flaws of the frequency of co-occurrence method by simultaneously taking all the local estimates of distance into account. It is a great exemplar of CSMs, namely for its implementation of the singular value decomposition (SVD) method that relies on dimensionality reduction, to simulate human word learning and disambiguation.

BOUND ENCODING OF THE AGGREGATE LANGUAGE ENVIRONMENT (BEAGLE) Another CSM worth mentioning is Jones, Kintsch and Mewhort [26]'s BEAGLE. They study the structure of semantic memory using a semantic priming task to learn associative information between words. This approach is different from the previous localist and distributional ones, which Jones, Kintsch and Mewhort [26] claim 'do not actually learn anything'. As a solution to the problems of strict contextual dependence, BEAGLE works by *convoluting* the semantic and associative properties of words in a holographic model that learns word meaning and other stored environmental information, such as word order.

The Semantic Problem

Although these CSMs are proficient in tasks such as synonym and association tests, and conversation analysis, these methods consider words in isolation. Words are found in combination as part of larger structures, such as in sentences, paragraphs, documents, and simple word-level representation is not sufficient in reflecting the actual use of language [28]. Another problem is that there are two types of extractable data from which semantic representations can be learned [3], distributional and experiential data, and these models tend to focus only on the former. Taking a closer look at LSA or HAL, we notice that they address the semantic representation of words in terms of their contextual properties, so the *external* properties of the concepts, but do not explain their similarity in terms of their *internal* properties, leaving the question of how or why they are similar unanswered [6, 26].

Frequency Not Sufficient

Human lexical semantic competence has many facets, such as word association and relation, and taxonomic judgements, which makes it difficult to gather the full content of mental representation of conceptual knowledge. Really grasping the meaning of concept means discovering deeper information about it. What is its function? What is it made of? Where does it come from? These questions cannot be answered by simply counting how often other words appear in its context, for that would assume that a word's meaning is entirely characterized by other words. Supplying a synonym as definition would not be sufficient either, because a similarity relation cannot explain what kind of links ties the two concepts [6]. Let us further exploit the environment: what else can we get out of the context?

Frequency of co-occurrence is the tool used to uncover similarity relations but minimizes representation because intuitively, using frequencies alone do not add to the semantics of concepts as it does not explain how the word *interacts* with its environment [1, 26, 18]. Baroni et al. [6] provide a good example to illustrate how frequently occurring collocates can be insufficient in meaning representation. *year of the tiger* appears much more often than any other pattern that connects *tail* and *tiger*, a pair that would be considered to have a stronger semantic link since more is learnt about the meaning attributed to *tiger*. Moreover, much of the syntactic information is lost as well [12]. Instead of basing meaning representation on similarity relations, we want to discover conceptual knowledge by exploring other types of relations, ones that would address the questions of function, composition, hierarchy or origin. For new goals, we need new tools.

RELATION PATTERNS Let's examine other possible semantic links between concepts and elements in their context. In 1992, Hearst [25] was amongst the first to propose the idea of lexico-syntactic patterns to express high precision semantic relations such as hyponymy and causality. He explored methods involving manual pattern induction, extraction, and ranking [4]. The pattern-based relation extraction has evolved since, and human labor has been equalled by automated techniques, as seen in the Attribute-Value (AV) model by Almuhareb and Poesio [1] and in that of Pantel and Pennacchiotti [36]. Both adopt this idea of surface relations as cues to semantic relations, but instead of finding them using predetermined patterns, discover them in a completely unsupervised manner using concept pairs. Manually created lexico-semantic patterns like those of Hearst showed that links other than those suggested by similarity relations were, if not more, important. However, the findings were tailored to fit the templates, and resulted in relations that could be quantified, instead of qualified. The automatic discovery of relation patterns led to the importance of variety, suggesting there being aspects of interaction between concepts [6].

VARIETY OF RELATION PATTERNS Interaction between the concept and the attribute beyond that of collocation is good evidence of the presence of an inherent semantic link, as Baroni et al. [6] states, proposing the importance of *distinct* patterns rather than their frequency. Moreover, a variety in relation patterns suggests they can be categorized into types. Typed relations are of interest because they can define the kind of relation shared between concepts, as Hearst encouraged, but when carried out in an unsupervised way, reveals much more about the context in terms of semantics.

What's more than defining the relation is being able to make descriptive comments about either of these connected concepts. Almuhareb and Poesio [1] provides a good example to illustrate this, using the highly probably presence of a link between the concept *dog* and the adjective *brown*. It is helpful to know that *dog* can have the value of being *brown*, but it is even more important to the understanding of *dog* to say that it includes such attributes as color, or size, or texture. This leads to the relevance of examining concepts in terms of their most salient features as an extension of their semantic representation.

Playing with Properties

After focusing on the relation extraction and relation typing, we must look to what these relations can say about the concepts, and what kind of knowledge we can extract from them. In cognitive science, it is suggested that concept representation consists of some form of decomposition into properties, and that these properties are organized depending on how they relate to the concept [6, 32]. While Barbu [4] refers to them as 'pieces of common sense knowledge', McRae et al. [33] dub properties 'semantic feature production norms' and are often referenced for their collection of feature norms. Their collection is the product of an experiment where the participants are presented with a set of concepts names and asked to generate features about these concepts that they deem the most important. McRae et al. assert their importance in constructing empirically derived conceptual representations in the scope of semantic representation and computation.

Feature-based methods are amongst the most prominent in cognitive science studies. They have their place in concept categorization [13] and hierarchy studies [38], as well as in linguistically-driven experiments concerning noun representation [45], and verbal thematic roles [32], to name a few. Their contribution in uncovering which aspects of meaning are psychologically salient is further supported by Kelly, Devereux and Korhonen [27], who extend these methods to concept-relation-feature triples, and Silberer, Ferrari and Lapata [40], applying them to a computer visual model.

The good news is, they are human-generated. From a cognitive point of view, feature norms are 'the most important properties of basic level concepts' [4] because they are used systematically by participants when generating features [33]. The bad news is, they are human-generated. McRae et al. say it best:

> "[W]hen participants produce features in a norming task, they directly exploit representations that have developed through repeated multisensory exposure to, and interactions with, exemplars of the target category. ... [Barsalou] stated that when participants generate features, they construct a holistic simulation of the target category, then interpret this simulation by using featural and relation simulators. [...] a participant's list of features represents a temporary abstraction that is constructed online for the

> purpose of producing feature names. Therefore, the dynamic nature of feature listing results in substantial variability both across and within participants.”

Faced with these issues, computational linguists found alternatives to borrowing from humans to generate feature-norm-like concept descriptions.

Strudel

Strudel, for Structured Dimension Extraction and Labeling, is a fully unsupervised CSM that extracts sets of property-based concept descriptions from corpora of naturally occurring text. The dimensions of the Strudel semantic space are interpretable as weighted properties comparable to human-generated norms. Strudel also provides concept-property typed relation patterns, which further characterize how the property relates to the concept, addressing the question of *how* they are similar, not just how much they are.

The authors of Strudel define their model based on three fundamental intuitions:

1. The relation between the concepts and the properties is categorized by the pattern that connects them.
2. The number of distinct patterns connecting concepts and properties play a very important part, as variety suggests a stronger semantic link than one instantiated by simple collocational association.
3. The distribution of patterns aids in word sense disambiguation.

Given a lemmatized and part-of-speech (POS) tagged corpus[1], a list of selected target nominal concepts[2], and a set of relation pattern templates, Strudel automatically extracts and annotates neighboring content words (verbs, adjectives or nouns) identified by the templates. The templates are created from possible POS sequences, which allows for the automatic discovery of relation patterns, unlike like most relation extraction algorithms that start with a predefined set of relations.

For example, if Strudel is looking for and nominal property, the template would be a plausible structure such that target concept C and candidate property P are connected by a preposition (or a verb,

---

[1] a version ukWaC [5]

[2] a set of concrete concepts from a corpus of child-directed speech [39]

a possessive *'s* or a relative pronoun such as *whose*). Given the C *onion*, Strudel would find patterns such as *onion with layers* or *layers of an onion*, which would result in templates C_with_P, and P_of_a_C (note normalized determiner *an*). These resulting templates are then used to find other related pairs, like *tigers with tails* or *tail of a tiger*. It is important to note if the relationship between C and P is expressed in a number of different ways, pattern variety denoting a semantic cue [6].

In the first half, the collected candidate properties are filtered then scored based on two factors: the number of distinct patterns connecting them to the concept, and the strength of their statistical association with the concept. The purpose of weighting the properties is to indicate which properties describe the concepts best, but also to demonstrate in what aspects concepts are similar to others. The second half consists in generalizing and classifying the retrieved patterns in order to assign them a type sketch to further indicate the way in which the properties are related to the concept. The type sketches include part-of, hypernymy, location, and function.

The Strudel model is inspired by the Rapp [37] SVD model, the Almuhareb and Poesio [1] AV model, and Padó and Lapata [35] dependency vectors (DV) model, the 'three broad lines of CSM research' [6]. It is similar to LSA in that the original matrix is reduced to a word-by-weight-left-singular-vector matrix, which allows for the dimensions to more accurately capture patterns of correlations, but instead of a word-by-document matrix, uses a word-by-word matrix, like HAL.

In a direct evaluation against the gold standard, the McRae human-generated norms, the property-based concept representations produced by Strudel proved to be reasonable both qualitatively and quantitatively [6]. At an average precision of 23.9%, it outperforms its three competing CSMs in categorization of concepts into superordinates, and has the advantage of being able to generalize over various feature types with no reliance on strict relation patterns [13], all good grounds to validate the use of structured property representations as dimensions of CSMs. Devereux et al. confirm its success by electing Strudel as the best method for their feature-based concept representation experiments. They exploit the fact that Strudel properties tend to focus on activities and interactions, while the McRae norms include physical and parts properties, and suggest a combination of the two types of properties to enrich existing models.

2.2.2 Distributional Semantics with Computer Vision Models

The way in which the distributive hypothesis takes form in computer vision is analogous to that of the computational linguistics approach. Although language and vision are two distinct modalities, they can be investigated using similar methods. Feng and Lapata [16] demonstrate how images, typically described by a continuous feature space, can be treated as words, commonly dealt with as distinct units, by converting visual features into units onto a discrete space.

A common practice in image processing is to segment images into regions, and represent each region by a standard set of features. Local regions in images are like words in text, and can be segmented using different techniques, such as the normalized cut algorithm, fixed grid-layout segmentation, and the Scale Invariant Feature Transform (SIFT) point detector method. Unlike words in a lexicon, these units do not have a 'definition' and must be assigned one to then build a vocabulary of visual words. This is commonly referred to as the bag-of-visual-words (BoVW) method [41], which serves to create discrete representations for images. Each region is characterized by a vector of base features, the information that is extractable from images, such as color, texture, and edges. The feature vectors are then compared to each other and grouped based on their similarity, where the groups are the visual words in the vocabulary, and are assumed to originate from similar objects. This allows for images to be expressed in terms of their BoVW feature vectors.

Again, language and vision are two distinct sources of information, but that does not mean they are competitors in the race to explain how we humans arrive at concept meaning representation. Cognitive analyses of mental representations of meaning propose that conceptualization and comprehension can also be found in non-linguistic representations, such that cognitive structures 'arise from bodily interactions with the world' [22]. Research investigating young children's object naming [29] report a strong relation between perception and conceptual knowledge, showing that when given a new named object, then another object with the same shape, children tend to generalize the name of the first object over the second one. A similar study [46] finds that visual representations of object shape are activated in the brain during sentence comprehension, ultimately suggesting that language helps the receiver construct an experiential simulation of the described event. Such experiments have led to extending views of meaning representation to associate visual perceptual information with linguistic units.

Multimodality

The idea of combining language and vision is not a novel one, though. Cues from the brain, namely functional magnetic resonance imaging (fMRI) recordings of neural signals, indicate that there is a significant correlation between image and brain-based semantic similarities as seen in a study by Anderson et al. [2]. A comparison of a text-based and an image-based distributional model in the experiments carried out by Feng and Lapata [16] also promote unification. Both result reports demonstrate that models based on images and on text are not only complimentary, but possibly mutually beneficial, and that best results can be achieved when combining both modalities.

TEXT TO VISION A good example of a study that encourages the potential of perceptually grounded distributional models that do not gain insight exclusively from text is one conducted by Bruni and Baroni [8]. Their goal of reaching a more human-like notion of meaning by combining techniques from natural language processing and computer vision is presented as a multimodal distributional semantic model. First, they construct text-based and image-based co-occurrence models separately, then combine them. Specifically, they concatenate textual word vectors with their equivalent visual word vector (the BoVW values for images that have been labeled with the same word). They show that combining image-based vectors and text-based distributional vectors leads to qualitatively different results when tested on semantic relatedness tasks and concept clustering, and that the two sources are complementary.

VISION TO TEXT Another example of a model that succeeds in such tasks is presented in Bruni et al. [9]'s work, which also combines both modalities, but instead exploits computer vision techniques to improve text-based models. They show that distributional semantic models based solely on text can be outperformed by models that use visual information to represent words where vision is relevant, more specifically when the focus is on colors. With two different types of visual information, SIFT and LAB features, they demonstrate that words for which their color is typical (e.g. *parsley-green*) are better represented when the distributional textual model is aided by visual information.

Properties: You Again?

One study that also borrows from both text and images, but incorporates an extra ingredient, is that of Farhadi et al. [15]. They too present a distributional visual model, but stress the integration of the attribute-centric framework of meaning representation. While traditional computer vision algorithms describe each concept by assigning it a categorical label (e.g. *goat*), Farhadi et al. incorporate visual attributes (e.g. *has-horns*) to object recognition tasks. Jumping from recognition to describing, they develop a feature-selection method which more closely models the human capacities of representation as it not only names known and unknown objects, but can also comment on the absence of typical attributes, or the presence of unusual ones. Because objects share features, they identify attribute learning as the key problem in recognition, and make it the main component of their framework.

They focus on learning semantic attributes (such as parts, shapes, and material) and non-semantic attributes from localized objects in images. Localizing the object allows for a better focus on the description, because like the text-based models, the context in which an object is located can significantly contribute to the semantic representation [40]. First, from a corpus of images annotated with attribute labels obtained through Amazon's Mechanical Turk[3], they extract and filter base features based on how helpful they are in learning attribute-classifiers. The system is then trained to learn attribute-classifiers from the selected base features, and object categories from the predicted attributes.

The system's ability to generalize across object categories originates from a feature selection method that uses an ℓ_1-regularized logistic regression to decorrelate attribute predictions, in order to focus on within category prediction. This allows for the attributes to be the primary actor in object recognition, meaning that knowledge about concepts can be learned from their visual, and textual properties.

The approach of using semantic attributes for concept discrimination was also adopted by Silberer, Ferrari and Lapata [40]. In their

---

[3] Amazon's Mechanical Turk, a crowdsourcing Internet marketplace, allows individuals to post tasks that cannot currently be carried out by computers, and require the input of human intelligence (https://www.mturk.com/).

paper, they focus on 'physically grounding the meaning of words' by means of high-level visual attributes instead of low-level image features. The model proposed by Silberer, Ferrari and Lapata is unlike that of Farhadi et al. in that instead of using human-generated norms to learn the attribute-classifiers, they learn classifiers from a set of automatically retrieved topically-related attributes from text. They show that the attribute-based bimodal models perform better than the those rooted in a single modality, and outperform those whose word representations are based on human-generated norms.

2.2.3 Evidence from the Brain

Multimodal models for addressing the question of how conceptual knowledge is represented in the human brain have also been implemented in neuroscience. Similar to how the meaning of words is represented in terms of patterns of word co-occurrence, concepts are represented in terms of patterns of neural activation [2, 24]. As previously mentioned, brain imaging studies have shown there to be an association between distinct spatial patterns of neural activity and visual-textual concept representation. Mitchell et al. [34] touch on this in their "Predicting human brain activity associated with the meanings of nouns." paper, which presents a computational model able to make predictions of the fMRI signals associated with thinking about concrete nouns. Going along with the distributional hypothesis in linguistics, they build their model under the assumption that the neural basis of the semantic representation of concrete nouns is related to the distributional properties of those words.

First, each stimulus word is encoded a meaning as a feature vector whose values are extracted from a corpus of text, using a frequency of co-occurrence method similar to the ones previously mentioned. Participants are then shown picture-word pairs of objects organized by category, asked to actively think of properties related to these objects, and a representative fMRI image is created for each word. The next phase predicts the fMRI images as a weighted linear sum of the observed fMRI activation related to each intermediate semantic property. The results reveal a direct correlation between the statistics of word co-occurrence in text and the neural activation from internal word meaning interpretation, which argues that neural representations of concrete nouns do find roots in sensory-motor features.

2.3 SUMMARY

Although the distributional hypothesis finds its origins in linguistics, it has been adopted by many other scientific fields to attempt answering the question of how humans acquire and organize representation of meaning via conceptual knowledge. From a cognitive and technical point of view, it is evident from the literature that our interaction with the physical world plays an important role in how we process this information. More specifically, language and vision can combine forces to reach visual concept knowledge through induction using semantic property-based concept descriptions from text.

The following chapter will introduce the proposed experiment, inspired and supported by the findings from past and current studies which suggest the cognitively and technically plausible combination of language and vision.

3 EXPERIMENT

3.1 OVERVIEW

From the literature, we've seen that concepts can be represented using property-based descriptions both from text and images, independently and dependently. The following experiment is designed to test whether semantically enhanced conceptual knowledge from images can be achieved using linguistic information. Visual models learn classifiers based on the observations it extracts from the training set. The goal here is to shift their focus to learn from a set of semantically derived textual features, instead of ad-hoc human-generated features. The visual model to be used in this experiment is inspired by that of Farhadi et al. [15] but is different in that instead of using the task-driven features to learn classifiers, feed it the text-extracted properties from the Strudel experiment. A visual pipeline of the experiment is depicted in figure 1. Again, using text-extracted information is motivated by the fact that they are conceptually close to human-generated norms, but differ from them in that they are acquired in an unsupervised manner and cover a more complete set of features. Ultimately, the success of the model will be measured in terms of how much knowledge can be gained from images when the predetermined list of visual properties is swapped with an automatically text-derived one.

The following experiment is intended as a preliminary study conducted to evaluate the feasibility the proposed model. The first half of the experiment consists in collecting, formatting and filtering the linguistic data. The results from this portion will indicate if the information extracted from text is relevant for the visual classification and annotation half.

First, concept-property pairs are collected from the Strudel output, and filtered to create a subset that contains only the most salient properties which will be used to train the visual model. The filtering process is achieved by way of clustering and regression, processes that will reveal which properties have the greatest effect in the concept classification task, and thus are the most discriminative.

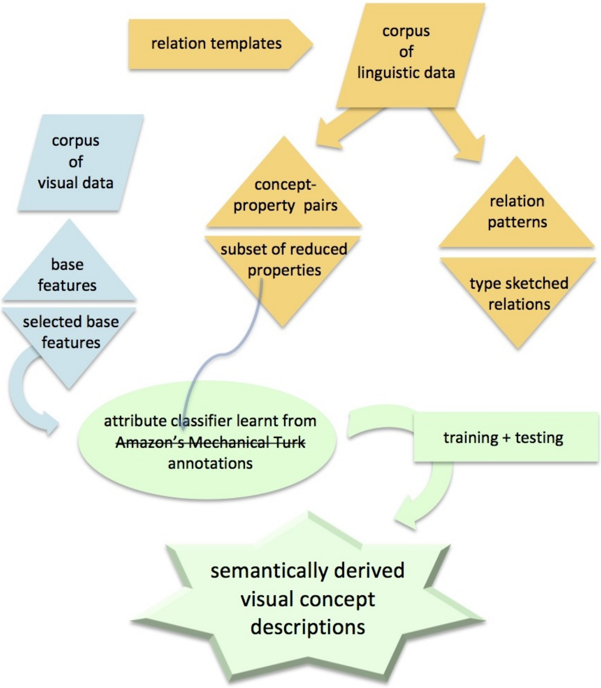

relation templates
corpus of linguistic data
corpus of visual data
concept-property pairs
subset of reduced properties
relation patterns
type sketched relations
base features
selected base features
attribute classifier learnt from ~~Amazon's Mechanical Turk~~ annotations
training + testing
semantically derived visual concept descriptions

Figure 1: Visual Pipeline of Experiment

The next step involves constructing the feature vectors for the visual model training portion. Since this is a pilot experiment, we chose 10 properties from the newly created subset to first make a qualitative analysis. For each of the selected properties, we retrieved the concepts with which they were originally paired, and collected images tagged with that concept from ImageNet[1]. Then, a 10-dimension vector was constructed for each image, where the features are the 10 properties, and the values the log-likelihood associated with the concept-property pair.

The second part of the experiment would consist in training and testing the system with the new properties, where it is evaluated based on the quantity and quality of conceptual knowledge extracted from the visual data.

3.2 DATA COLLECTION

We use the Strudel model output to collect the data because it is relevant and well-structured for the goal at hand. All words having been previously lemmatized, Strudel outputs concepts paired with their

---

[1] ImageNet (http://image-net.org), an ontology of images based on the WordNet (http://wordnet.princeton.edu/) hierarchy.

Concept	Property	LL	Patterns
grass	graze-v	187.4173	_+right+v, on+right+v, _+left+v
grass	blade-n	139.1197	of+right+n
razor	blade-n	220.4947	with+left+n, have+left+n, of+right+n
glow	gold-n	23.7200	above+left+n, with+left+n, like+left+n

Table 1: Example of Strudel output including log-likelihood measures and pattern types

automatically discovered property, and the typed relation patterns with which they occur. Each pair is presented with its log-likelihood, a value based on the probability of the two occurring together. The properties are distinguished from the concepts in that they are labeled with an appended POS tag (*-n* for nouns, *-v* for verbs, or *-j* for adjectives). This convention will be used throughout the current and upcoming chapters. The complete output counts 1216 concepts, 35 262 properties and 147 679 pairs. A small excerpt of the Strudel output is shown in table 1.

From the relation pattern templates, Strudel discovers pairs of words such that the first member is a nominal concept, and the second is either a noun, verb, or adjective that is a feature of the concept. Concepts can have more than one property, and properties can be features of more than one concept. Let's look at the target concept *razor*. It occurs with the property *blade-n* with a log-likelihood value of 220.4947, and the pair appears in 3 different surface settings:

1. with+left+n: 'razor(s) with blade(s)'
2. have+left+n: 'razors(s) have/has a blade(s)'
3. of+right+n: 'blade(s) of a/the razor(s)'

The first step in the filtering process is to eliminate those that are intuitively the least informative. Although frequency is not a deciding factor in the properties filtering process, we still want the properties to appear in at least two different pairs, because that would mean that they are valuable enough to describe at least two concepts. Every property that appears in only one pair was discarded. The number of concepts dropped to 1207, the properties to 15 936, and the pairs to 128 353. Some concepts were lost in the process because they had only one property, and that property occurred only once, eliminating the pair altogether. The new counts are found in table 2.

There were many suggestions as to what other factors should be considered in the next step of the filtering process. Since the Strudel

Reduction	Concepts	Properties	Pairs
Raw data	1216	35 262	147 679
Frequency > 1	1207	15 963	128 353

Table 2: Concept, property and pair counts after first round of filtering

output also provided the concept-property relation patterns, one proposition was to only keep the pairs whose patterns would suggest a part-of (such as C_have_P, ex: bulls have horns) or location (such as C_in_P, ex: cows in a field) relationship in attempt to target more visually graspable properties. Reconsidered and subsequently rejected, filtering by patterns would not allow for properties that would suggest shapes, colors or textures, important when the second part of the project involves vision. Moreover, with such a strict filtering, we could lose some properties that, although do not seem obviously imageable at first, may play an important role in the visual discrimination in a more covert fashion. The patterns were thus dropped from the data. With the remaining data, we were able to construct a concept-by-property matrix, with the log-likelihoods as values. The resulting matrix is very sparse (0.007%).

Knowing our ultimate goal of selecting the most informative, and thus discriminative properties, some backwards thinking was required to arrive at the next step. In statistics, a regression analysis allows the investigator to make predictions about certain variables in an unsupervised manner, based on the effect of certain factors by analyzing their relationship. In the process, much information is learnt about the independent variables, such as if they are related to the dependent variable, and if so, what is the weight of their effect. If the independent variables are assigned weights, then they can be ranked corresponding to their level of influence: this is where we catch them. Now, since the regression task requires a set of training data that includes both the predictors and the predicted, we needed to get ourselves some dependent variables.

3.3 CLUSTERS SHOULD DO THE TRICK

Clustering is a machine learning technique that involves grouping items into clusters based on a set of item descriptions. Clustering systems are unsupervised in that the items do not need to be preclassified, since the system is able to discover classes by 'learning by observation', instead of from examples [18]. Clusters are formed by grouping items with similar descriptions, or observations. The similarity between the items can be calculated using different metrics, such as the Jaccard index, or the cosine of two angles. With our data, we want to group the concepts into clusters based on their defining properties. The resulting clusters will be assigned a class that will be used to train the regression model. Our data is large, but it is mostly sparse, so for best results, we chose to adopt the affinity propagation clustering technique.

3.3.1 Affinity Propagation Clustering

The affinity propagation (AP) clustering is an algorithm introduced by Frey and Dueck [19], implemented in R as the `apcluster` package [7], that groups data points into clusters using a real-valued message exchange method. This particular clustering method was chosen for two main reasons:

1. it is optimized for sparse data
2. it is completely unsupervised

The goal of the algorithm is to identify the exemplars around which all the other data points have grouped around to form the clusters. An exemplar is a member of the cluster considered to be its representative. The AP technique differs from other clustering methods in that each and every data point is simultaneously considered as a potential exemplar, but still, the aim remains for the squared errors between the center of the cluster and its other members to be as small as possible. The algorithm employs a message passing method where real-valued messages are exchanged through the network of data points until the exemplars are identified and the clusters are formed. This gives it an advantage over other methods of clustering in that there is no particular configuration of the set of exemplars.

There are two kinds of messages: the 'responsibility' message $r(i,k)$, from *i* to *k*, where the suitability of point *k* to serve as an exemplar to point *i* is quantified, and the 'availability' message $a(i,k)$ from *i* to *k*, which refers to how appropriate it is for point *i* to choose candidate *k* as its exemplar. The name 'affinity propagation' arises from the idea of measuring the connection of each message passed from one data point to its currently chosen exemplar, or the affinity one point has for the other as a candidate exemplar [19].

$$a\left(k,k\right) \leftarrow \sum_{i' s.t. i' \neq k} max\left\{0, r(i',k)\right\} \tag{1}$$

Equation 1 defines the algorithm, where the 'self-ability' $a\left(k,k\right)$ of *k* as an exemplar is based on the positive responsibilities sent to candidate exemplar *k* from other points.

As previously mentioned, the AP algorithm runs completely unsupervised. Unlike for the *k*-centers clustering technique, which is dependent on a manually chosen number of exemplars, the number of clusters in AP is a parameter discovered and defined over a number of iterations by the algorithm itself. Although there is an option to manually prespecify this parameter, we choose to continue the theme of unsupervised methods.

The algorithm determines the clusters based on the similarity of the concepts, and thus requires a concept-by-concept similarity matrix as input. The strength of the similarity between pairs of concepts is given by calculating the cosine of their vectors (the rows of the above mentioned concept-by-property sparse matrix). Therefore, the concepts are compared to each other not only according to which properties describe them, but relative to the strength of their relation as well, as indicated by the log-likelihood measures.

3.3.2 Clustering Results

The algorithm concludes when it has reached the point where the exemplars have not changed for 100 iterations, the default. The results include the details of the task, the exemplars and the clusters, and various plots for data visualization.

Table 3 presents the most important results from the AP clustering task. After 270 iterations, the algorithm has refined the number of exemplars to 183, assigning all 1207 samples a group according to their similarity value. Since we did not set a value for the number of clusters, the input and sum of preference is 0. The net similarity is the sum of all similarities between non-exemplar data points and their exemplars plus the preferences. This value indicates how well the similarities have been maximized. These performance measures are illustrated in figure 2.

APC results	
Number of samples	1207
Number of iterations	270
Input preference	0
Sum of similarities	400.0062
Sum of preferences	0
Net similarity	400.0062
Number of clusters	183

Table 3: APC output: specifics

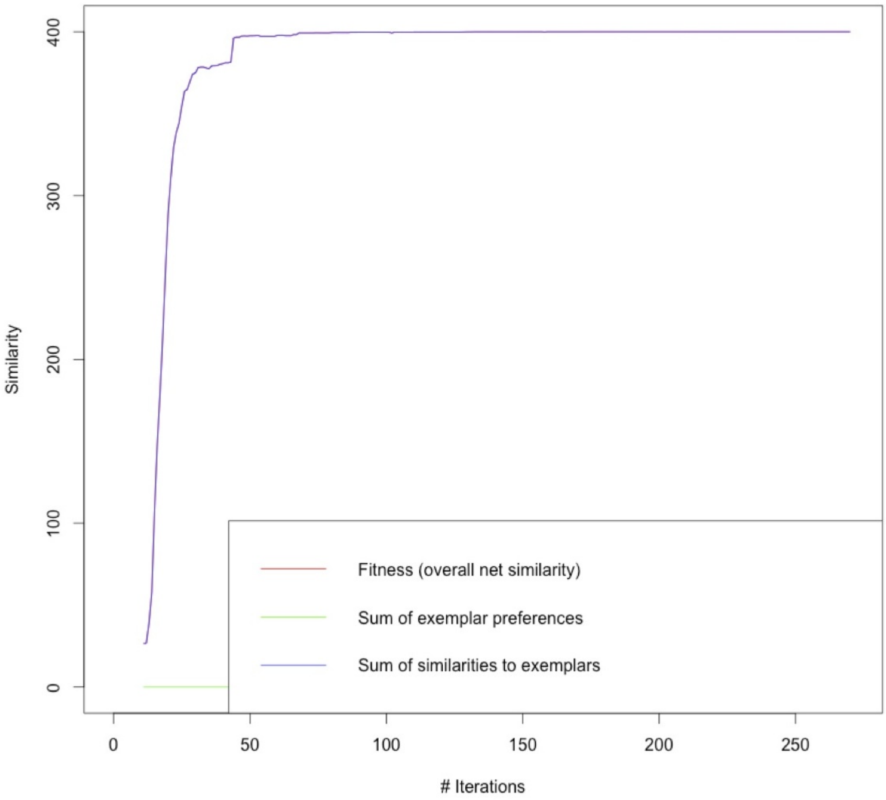


Figure 2: Algorithm Performance

	Exemplar	Cluster members
cluster 5	armchair	armchair bench chair couch cushion desk seat settee sofa stool table waterbed
cluster 16	bottle	barrel cellar corkscrew wineglass jar jug pub
cluster 20	burger	artichoke food grill kebab pizza sandwich sausage stuffing turkey
cluster 152	snowball	dice die grenade paddy wrench
cluster 161	tack	detour tablet

Table 4: APC output excerpt: clusters

As reported, the net similarity is equal to the sum of similarities to exemplars since the sum of exemplar preferences is 0. It can also be noted that the algorithm has not made any changes in the last 200 iterations. This is greater than the default because the size of the data requires more iterations for the algorithm to converge.

Although 183 appears to be a large value for the number of clusters, we must take into consideration that from a sample size of 1207, an average of 7 concepts per cluster seems reasonable, since the number of categories for classifying concepts in this particular setting could be 1207.

The results also include a list of the chosen exemplars and their corresponding cluster, which vary in size and range from 2 to 53 concepts per cluster, with an average of 21. A quick overview of the clusters and their members reveals the task to be successful, where most concepts form sound clusters that can be labeled with tags such as *furniture*, *food* and *animals*. An excerpt of the output is presented in table 4. Where most clusters form an intuitively sound group, such as clusters 5, 16 and 20, some are not as obvious, as seen in cluster 152 or 161. But, if we think in terms of properties, some categories can be inferred, such as ‘things you throw’ for cluster 152.

Seeing that sound clusters can be derived from the concepts and their properties, we are assured that the role the properties played in the classification task is also sound, but to what extent? What we are looking for now are the properties which had the strongest effect and played the greatest role in categorizing the concepts. For this next portion of the experiment, we looked to a multinomial logistic regression model to perform the classification and feature selection task.

3.4 PROGRESSION TOWARDS REGRESSION

Regression analysis is often employed for statistical prediction tasks, as is it a tool for investigating the causal relationships between variables. The analysis is carried out by collecting information about the causal variables of interest and employing regression to estimate and quantify their effect on the dependent variables [42].

Since we want to model the relationship of multiple correlated concepts with multiple correlated properties, we chose to apply our data to a multinomial logistic regression model that uses an ℓ_1 penalty regularizer parameter, the lasso, as it is efficient for automatic feature selection on sparse data [20].

3.4.1 Lasso Regression

Lasso, for 'Least Absolute Shrinkage and Selection Operator', is a regression method that involves penalizing the absolute size of the regression coefficients. This method is ideal to solve our problem, as it will allow us to determine the strength of the predictors based on their weighted coefficient: the smaller the coefficient of x, the less they contribute to a good prediction of y [43].

The main idea behind the lasso is to preserve a minimal residual sum of squares while constraining the sum of the absolute value of the coefficients under a certain threshold. To achieve this balance, a regularizer parameter λ is introduced to control the shrinking of the coefficients towards 0, and ultimately define the number of predictors in the regression model. The larger the value of λ, the more relaxed the penalty, meaning the greater the number of predictors retained. For a predictor to be retained, its coefficient must be greater than 0, which means that it has not been forced to shrink to 0. Conversely, as the penalty becomes more constrained, the shrinkage is allowed to increase and force the weaker coefficients to 0. These are consequently eliminated, giving a more interpretable model for which the subset of predictors includes the most discriminative ones.

$$(\hat{\alpha}, \hat{\beta}) = \arg\min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \text{subject to} \sum_j |\beta_j| \leqslant t. \tag{2}$$

The lasso estimate $(\hat{\alpha}, \hat{\beta})$ is defined by equation 2 above, where x_{ij} are the standardized predictor variables, and y_i are the responses. Here $t \geqslant 0$ is the regularizer that controls the amount of shrinkage that is applied to the coefficients.

We chose this method because, unlike the ridge regression where they continuously shrink, the coefficients in the lasso regression actually fall to 0, improving the accuracy by reducing variance of the predicted values. Moreover, the lasso provides support for sparse data, and improved interpretation when the goal is to determine a smaller subset of predictors that exhibit the strongest effects [43].

3.4.2 Regression Results

Regression is employed to discover the best model to predict the category y of a given dependent variable based on the predictor variables x. Here, the y variables are the cluster labels for a given concept, and the x variables are the retrieved properties. To run the regression task, we used the `glmnet` package [21] in R. The first step of the task consists in fitting the model, a process for which all the relevant details are available and can be visually presented, as in figure 3.

Each curve in 3 corresponds to a single predictor. As the penalty becomes more relaxed, represented by the increasing ℓ_1 Norm values, the number of non-zero coefficients (measured using the topmost axis) and their value (x-axis, which also indicates the number of predictors: 183) increase as well. The package also allows one to retrieve a list of the path of each predictor at each step of the fit, as well as at a specific λ.

To get the best model with the optimal regularizer value, we run a cross-validation and are returned two selected values: `lambda.min`, the value of λ that gives minimum mean cross-validated error, and `lambda.1se`, which gives the most regularized model such that the error is within one standard error of the minimum. The cross-validation curve is plotted in figure 4, represented by the red dotted line, between the upper and lower standard deviation curves. The two vertical lines indicate the two selected values for λ.

The prediction algorithm and its learned ability to predict cluster labels is not what is of interest to us, though. What we want to look at is a particular side product of the building process that led to this final model. From this model, we are able to locate and extract the properties that are active in the prediction task, meaning the pre-

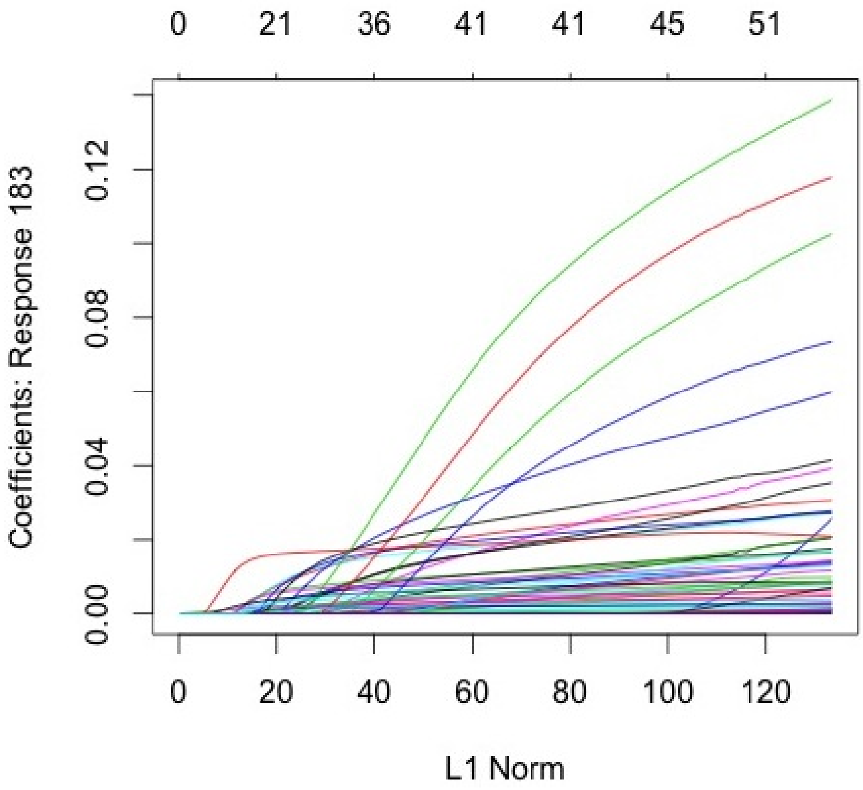


Figure 3: Model fit plot

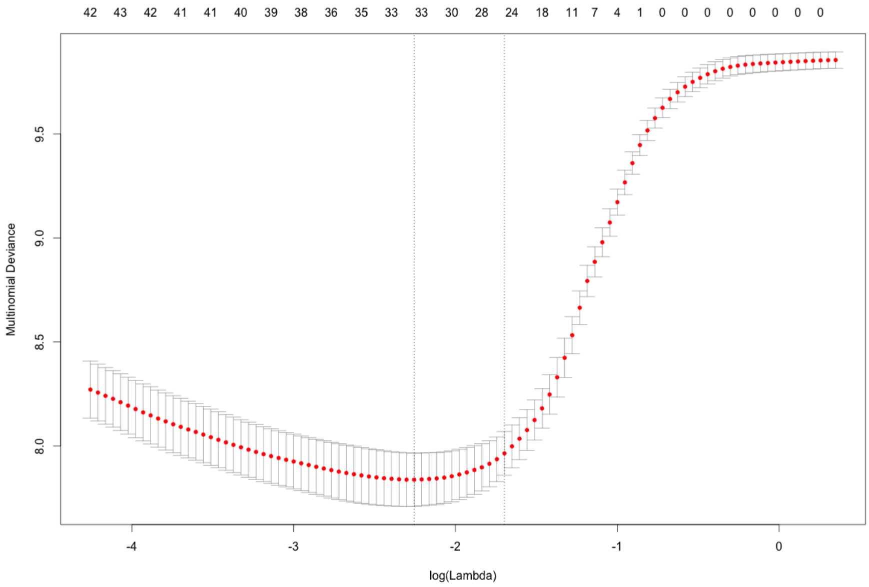


Figure 4: Cross-validation fit plot

Property of cluster 152	Coefficient
husking-v	0.00005
wafer-n	0.01733
bias-v	0.01388
incendiary-j	0.00470
throw-v	0.00427
rolling-n	0.00281
roll-v	0.00148
lob-v	0.00134
mutineer-n	0.00094
cast-v	0.00027

Table 5: Cluster 152 properties and their coefficients

Reduction	Concepts	Properties	Pairs
Raw data	1216	35 262	147 679
Frequency > 1	1207	15 963	128 353
Feature selection	1185	3659	X

Table 6: Concept, property and pair counts
after second round of reduction

dictor variables whose coefficient was not forced to shrink to 0 by the regularizer. Let's look at cluster 152 (*snowball: dice die grenade paddy wrench*) from table 4, for which the regression results are presented in table 5. It is interesting to see which properties were preserved, and in fact, the presumed category 'things you throw' fits!

The final subset of properties now includes 3659 properties from the last 15 936. The number of concepts dropped as well, from 1207 to 1185. Some concepts were lost because while the properties were filtered for having a shared factor of greater than 1, the concepts were not. For example, take the concept *chipmunk*. In the retrieved pairs from the Strudel data, *chipmunk* only appears once, with property *scurry-v*. *scurry-v* was not removed in the first round (when removing properties with shared factor =1) because it occurs with 11 other concepts. However, after the regression, *scurry-v* was not amongst the active properties and so was taken out of the 15 936, along with *chipmunk*. This happened with 22 other concepts.

3.5 THE VISION PART

This is where the real experiment begins: if the resulting subset of selected properties are the most discriminative for the prediction of concepts in *text*, can they also reveal themselves to be the most discriminative in predicting and describing concepts in *images*? To evaluate this, we put them to the test by setting them as the predictors in a visual distributional model. The visual distributional model takes images and their corresponding feature vector to train the system to recognize the properties that we have selected. Since this is a pilot experiment, we chose 10 properties (see table 7 in chapter. 4) from the newly created subset to first make a qualitative analysis. For each of the selected properties, we retrieved the concepts with which it was originally paired, and collected images tagged with that concept from ImageNet. Then, a 10-dimension vector was constructed for each image, where the features are the 10 properties, and the values the log-likelihood associated with the concept-property pair. While most vectors are usually populated with binary values because the properties are coded manually, we have decided to use the log-likelihood values which gives the advantage of a more precise training process based on weights.

3.6 SUMMARY

So far, we've taken a set of concept-property pairs, reduced them to a subset of the most salient properties, and before subjecting them to the visual model training task, I will assess their potential in a qualitative analysis. In the following chapter, I present 10 of the 3659 preserved properties, accompanied by a detailed explanation of intentions and expected results.

4 EVALUATION

4.1 OVERVIEW

In the following qualitative analysis, ten properties have been selected in an attempt to demonstrate the findings, represent the intentions and achievements, and express the failures of the proposed experiment.

The selected properties were hand chosen to explore all aspects of the expected results. They include four nouns, two adjectives and three verbs. It must be noted that although *gold-n* is tagged as a noun, it seems to play the same role as the color, and therefore can be considered an adjective. Similarly, *evil-n*, although tagged as a noun, resembles more an adjective. In table 7, the properties are listed followed by the all the concepts with which they were originally paired. In other words, each of these sets of concepts share the same property. Pictures are supplied to aid the illustration.

4.2 ANALYSIS

Concrete-Abstractness

Amongst the selected ten are properties that fall on either side of the concrete-abstract scale, such as *horn-n* and *evil-n*, respectively. Learning a classifier for *horn-n* seems easy enough, since the images supplied for training depict all the concepts for which this property is physically present. On the other hand, it appears more difficult to visually capture aspects of the more abstract *evil-n*. It is our intuition however, that not-so-obvious features can be returned, perhaps even beyond our scope of perception. We could expect the classifier to catch features that inspire evil, such as darker, colder hues, maybe the presence red, sharp contrasts, or even objects or shapes tied to the theme.

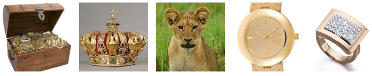
Property	Concepts they are properties of
blade-n	axe dagger grass helicopter knife oar razor scissors
bushy-j	beard mane mustache plant
sharp-j	arrow axe knife needle pencil razor scissors sword
evil-n	demon monster witch
gold-n	chest coin crown glow lion plate ring 
sparkle-v	chandelier ring water
graze-v	bison deer goat grass hill horse meadow rabbit sheep
horn-n	antelope bison buffalo bull deer goat rhino sheep
preach-v	church synagogue vicar
pair-n	earring jean mitten scissors shoe sock

Table 7: 10 examples of preserved properties and the concepts they describe

Imageability

Properties were also selected to explore the imageability scale, where, for example *gold-n* would be considered much easier to build a classifier for than say, *preach-v*. Again, the idea here is to try and reach for the features that capture the meaning of the concept in the manner humans do, even if covertly. In the training, the images of *vicar* could place the need for a person to be included in the classifier, while those of *church* and *synagogue* could provide the setting. A successful classifier for *preach-v* would be able to connect these two, and learn a meaning of action. Ideally, when presented with the concept of *church* or *synagogue*, the model would understand that such scenes are plausible settings for a vicar to carry out his preaching. Conversely, a vicar would be recognized as having his place in a religious setting.

Actions

Similarly, if we look at the property *graze-v*, defined as 'to eat grass in a field', it entails there being an entity performing the action of eating, and the thing being eaten to be grass. Therefore, not only does the property suggest these two sets be present in the context, but requires them to be. Provided with images representing the concepts, so *hill*, *grass* and *meadow* on one side, and *bison*, *rabbit* and *sheep*, on the other, a classifier for *graze-v* could be successfully achieved if the model were able to capture the relation between the two. Therefore, it is expected that when faced with such concepts, the model will have learned that these animals are often found surrounded by grass.

Colors and Textures

To remain on the topic of imageability, if *gold-n* can be visually represented by a color, then it can be assumed that *bushy-j* can be visually represented by a texture. 'bushy' can be used to characterize concepts of various categories, and therefore is a texture that does not have a definitive color or setting. The model would be considered successful if it were able to focus on the arrangement of visual features that would qualify a concept as being bushy and detect a visual pattern that is in fact, a sensory texture.

Cross-category Properties

Additionally, we wanted to include properties that describe some same concept. The concept *scissors* is an example of this, where they have the properties of being *sharp-j*, having a *blade-n* and occurring in a *pair-n*. We chose to include these because it would allow for the discovery of possible relationships between the properties as well. If more than one concept has the property of being sharp and having a blade, could there be a connection between the two? If so, is it detectable by the model?

Counting

Finally, the property *pair-n* would be a good example to test for the ability of the model to not just count, but understand the meaning of a number. For humans, mittens are understood to occur in pairs, and when alone, are known to be only one half of the whole it represents. If presented with an image depicting a pair of mittens, or a pair of socks, the model could easily detect there being two of the same object. If, however, we want the model to achieve a semantic representation closer to that of humans, then we want it to perform beyond its ability to just count. When ascribed the property *pair-n*, the model must recognize the necessity of there being two of the same object for its function to be realized, or better, when presented with only one, understand that it is incomplete.

4.3 SUMMARY

The above analysis is to give an account of the intentions and expected results of the proposed visual model. From the small sample of properties, we can easily define what we would like to see be realized in the next steps of the experiment, and give a qualitatively supportive rendering of what the model is set out achieve. From the visual classification task, we expect the proposed model to develop feature-selection methods that more closely model the human capacities of recognition by learning classifiers for properties beyond just low-level base features. From the annotation task, we hope the concept annotations reveal meaningful observations tied to function, or origin, over just physical descriptions, especially when presented with unknown concepts.

5 CONCLUSION

Past studies have shown us that it is possible to achieve conceptual knowledge and meaning representation close to that of humans using computational models rooted in language and vision, with neuroscientific support. More importantly, we've seen that best results are obtained when both modalities are combined. Knowing that humans employ both visual and linguistic cues when acquiring and organizing knowledge about concepts has provided us with the motivation to reach the middle ground between visual features and linguistic words. With this in mind, we set out to build a visual model enhanced with semantically-induced concept descriptions from text capable of object recognition and annotation.

The reported experiment is a preliminary study conducted to evaluate the feasibility of the proposed model. It consists in filtering the language data to assess if the value of the extracted information is relevant for the visual classification and annotation portion of the experiment. The evaluation presented in the qualitative analysis encourages the idea that the proposed multimodal distributional model is indeed plausible, both from a cognitive and technical perspective. Such a model would have many advantages for all fields concerned with this topic, as it would present a different approach to achieve the human capacity of conceptual knowledge grounded in visual perception in a cognitively plausible and completely unsupervised manner.

5.1 FUTURE WORK

The experiment presented above lays out a strong ground for the realization of the proposed model. The next steps include training and testing a visual system inspired that Farhadi et al. [15]'s model with our 10 attributes, and carry out a small-scale quantitative analysis. Its success will determine the potential of a full-scale experiment, including all 3659 properties.

BIBLIOGRAPHY

[1] Abdulrahman Almuhareb and Massimo Poesio. "Attribute-Based and Value-Based Clustering: An Evaluation". In: *Proceedings of the Conference on Empirical Methods on Natural Language Processing*. 2004.

[2] Andrew J Anderson et al. "Of Words, Eyes and Brains : Correlating image-based distributional semantic models with neural representations of concepts". In: *Proceedings of the Conference on Empirical Methods on Natural Language Processing*. 2013.

[3] Mark Andrews, Gabriella Vigliocco and David Vinson. "Integrating experiential and distributional data to learn semantic representations." In: *Psychological review* 116.3 (July 2009), pp. 463–98.

[4] Eduard Barbu. "Combining methods to learn feature-norm-like concept descriptions". In: *Proceedings of the ESSLLI Workshop on Distributnioal Lexical Semantics*. 2008, pp. 9–16.

[5] Marco Baroni and Alessandro Lenci. "One distributional memory, many semantic spaces". In: *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics* (2009), pp. 1–8.

[6] Marco Baroni et al. "Strudel: a corpus-based semantic model based on properties and types." In: *Cognitive Science* 34.2 (Mar. 2010), pp. 222–54.

[7] Ulrich Bodenhofer, Andreas Kothmeier and Sepp Hochreiter. "APCluster: an R package for affinity propagation clustering." In: *Bioinformatics (Oxford, England)* 27.17 (Sept. 2011), pp. 2463–4.

[8] Elia Bruni and Marco Baroni. "Distributional semantics from text and images". In: *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*. 2011, pp. 22–32.

[9] Elia Bruni et al. "Distributional semantics in technicolor". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Vol. 1. 2012, pp. 136–145.

[10] Elia Bruni et al. "Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning". In: *Proceedings of the 20th ACM International Conference on Multimedia*. 2012, pp. 1219–1228.

[11] John A Bullinaria. "Semantic Categorization Using Simple Word Co-occurrence Statistics". In: *Proceedings of the ESSLLI Workshop on Distributnioal Lexical Semantics*. 2008, pp. 1–8.

[12] Stephen Clark. "Vector space models of lexical meaning". In: *Handbook of Contemporary Semantics*. Ed. by Shalom Lappin and Chris Fox. 2nd ed. March. 2014, pp. 1–43.

[13] Barry Devereux et al. "Towards Unrestricted, Large-Scale Acquisition of Feature-Based Conceptual Representations from Corpus Data". In: *Research on Language and Computation* 7.2-4 (July 2010), pp. 137–170.

[14] Stefan Evert. "Corpora and collocations". In: *Corpus Linguistics. An International Handbook*. Ed. by A. Lüdeling and M. Kytö. Berlin: Mouton de Gruyter, 2008.

[15] Ali Farhadi et al. "Describing objects by their attributes". In: *IEEE Conference on Computer Vision and Pattern Recognition* (June 2009), pp. 1778–1785.

[16] Yansong Feng and Mirella Lapata. "Visual information in semantic representation". In: *Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistcs* June (2010), pp. 91–99.

[17] J R Firth. "A synopsis of linguistic theory 1930-55." In: *The Philological Society* 1952-59 (1957), pp. 1–32.

[18] Douglas H. Fisher. "Knowledge Acquisition Via Incremental Conceptual Clustering". In: *Machine Learning* 2 1980 (1987), pp. 139–172.

[19] Brendan J Frey and Delbert Dueck. "Clustering by passing messages between data points." In: *Science* 315.5814 (Feb. 2007), pp. 972–6.

[20] Jerome Friedman, Trevor Hastie and Rob Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), p. 22.

[21] Jerome Friedman et al. *Package 'glmnet': Lasso and elastic-net regularized generalized linear models*. 2014.

[22] Arthur M. Glenberg. "What memory is for." In: *The Behavioral and Brain Sciences* 20.1 (Mar. 1997), 1–19; discussion 19–55.

[23] Zellig S. Harris. "Distributional structure". In: *Word* 10 (1954), pp. 146–162.

[24] James Haxby et al. "Distributed and overlapping representations of faces and objects in ventral temporal cortex." In: *Science (New York, N.Y.)* 293 (2001), pp. 2425–2430.

[25] Marti A. Hearst. "Automatic Acquisition of Hyponyms from Large Text Corpora". In: *Proceedings of the 14th Conference on Computational Linguistics*. 1992, pp. 539–545.

[26] Michael N. Jones, Walter Kintsch and Douglas J.K. Mewhort. "High-dimensional semantic space accounts of priming". In: *Journal of Memory and Language* 55.4 (Nov. 2006), pp. 534–552.

[27] Colin Kelly, Barry Devereux and Anna Korhonen. "Acquiring Human-like Feature-Based Conceptual Representations from Corpora". In: *Proceedings of the NAACL HLT 2010 First Workshop on Computational Neurolinguistics*. June. 2010, pp. 61–69.

[28] Kirill Kireyev. "Beyond Words : Semantic Representation of Text in Distributional Models of Language". In: *Proceedings of the ESSLLI Workshop on Distributnioal Lexical Semantics*. 2008, pp. 25–33.

[29] Barbara Landau, Linda Smith and Susan Jones. "Object perception and object naming in early development." In: *Trends in cognitive sciences* 2.1 (Jan. 1998), pp. 19–24.

[30] Thomas K. Landauer and Susan T. Dumais. "A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge". In: *Psychological Review* 104.2 (1997), pp. 211–240.

[31] Kevin Lund and Curt Burgess. "Producing high-dimensional semantic spaces from lexical co-occurrence". In: *Behavior Research Methods, Instruments, & Computers* 28.2 (June 1996), pp. 203–208.

[32] Ken McRae, Todd Ferretti and Liane Amyote. "Thematic Roles as Verb-specific Concepts". In: *Language and Cognitive Processes* 12 (1997), pp. 137–176.

[33] Ken McRae et al. "Semantic feature production norms for a large set of living and nonliving things". In: *Behavior Research Methods, Instruments, & Computers* 37.4 (2005), pp. 547–559.

[34] Tom M Mitchell et al. "Predicting human brain activity associated with the meanings of nouns." In: *Science (New York, N.Y.)* 320.5880 (May 2008), pp. 1191–5.

[35] Sebastian Padó and Mirella Lapata. "Dependency-based construction of semantic space models". In: *Computational Linguistics* (2007).

[36] Patrick Pantel and Marco Pennacchiotti. "Espresso: Leveraging generic patterns for automatically harvesting semantic relations". In: *Proceedings of the 21st International Conference on Computational Linguistics*. Hindle 1990. 2006.

[37] Reinhard Rapp. "Word Sense Discovery Based on Sense Descriptor Dissimilarity". In: *Proceedings of the Ninth Machine Translation Summit* (2003), pp. 315–322.

[38] Eleanor Rosch and Carolyn B Mervis. *Family resemblances: Studies in the internal structure of categories*. 1975.

[39] Caroline F Rowland et al. "The incidence of error in young children's Wh-questions." In: *Journal of speech, language, and hearing research : JSLHR* 48 (2005), pp. 384–404.

[40] Carina Silberer, Vittorio Ferrari and Mirella Lapata. "Models of semantic representation with visual attributes". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia, Bulgaria, 2013.

[41] Josef Sivic and Andrew Zisserman. "Video Google: a text retrieval approach to object matching in videos". In: *Proceedings Ninth IEEE International Conference on Computer Vision*. Vol. 2. 2003, pp. 1–8.

[42] Alan O. Sykes. "An introduction to regression analysis". Chicago, 1993.

[43] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B* 58.1 (1996), pp. 267–288.

[44] Peter D Turney and Patrick Pantel. "From Frequency to Meaning : Vector Space Models of Semantics". In: *Journal of Artificial Intelligence Research* 37 (2010), pp. 141–188.

[45] Gabriella Vigliocco et al. "Representing the meanings of object and action words: The featural and unitary semantic space hypothesis". In: *Cognitive Psychology* 48 (2004), pp. 422–488.

[46] Rolf A. Zwaan. "The Immersed Experiencer: Toward An Embodied Theory Of Language Comprehension". In: *The Psychology of Learning and Motivation* 44 (2003), pp. 35–62.