Exploiting language models for visual recognition

Dieu-Thu Le
DISI, University of Trento
Povo, 38123, Italy
dle@disi.unitn.it

Jasper Uijlings
DISI, University of Trento
Povo, 38123, Italy
jrr@disi.unitn.it

Raffaella Bernardi
DISI, University of Trento
Povo, 38123, Italy
bernardi@disi.unitn.it

Abstract

The problem of learning language models from large text corpora has been widely studied within the computational linguistic community. However, little is known about the performance of these language models when applied to the computer vision domain. In this work, we compare representative models: a window-based model, a topic model, a distributional memory and a commonsense knowledge database, ConceptNet, in two visual recognition scenarios: human action recognition and object prediction. We examine whether the knowledge extracted from texts through these models are compatible to the knowledge represented in images. We determine the usefulness of different language models in aiding the two visual recognition tasks. The study shows that the language models built from general text corpora can be used instead of expensive annotated images and even outperform the image model when testing on a big general dataset.

1 Introduction

Computational linguistics have created many tools for automatic knowledge acquisition which have been successfully applied in many tasks inside the language domain, such as question answering, machine translation, semantic web, etc. In this paper we ask whether such knowledge generalizes to the observed reality outside the language domain, where we use well-known image datasets as a proxy for observed reality.

In particular, we aim to determine which language model yields knowledge that is most suitable for use in Computer Vision. Therefore we test a variety of language models and a linguistically mined knowledge base within two computer vision scenarios:

Human action recognition : Recognizing <subject, verb, object> triples based on objects (e.g., car, horse) and scenes (the place that the actions occur, e.g., countryside, forest, office) recognized in images. In this scenario, we only consider images with human actions so the "human" subject is always present.

Objects in context : Predicting the most likely identity of an object given its context as expressed in terms of co-occurring objects.

Computer vision can greatly benefit from natural language processing as learning from images requires a prohibitively expensive annotation effort. A major goal of natural language processing is to obtain general knowledge from text and in this paper we test which model provides the best knowledge for use in the visual domain.

Within the two visual scenarios, we compare three state-of-the-art language models and a knowledge base: (1) A window-based model, which counts co-occurrence frequencies within a fixed window; (2) R-LDA (Séaghdha, 2010), an extension of LDA that enables generation of joint probabilities; (3) TypeDM (Baroni and Lenci, 2010), a strong Distributional Memory model; (4) ConceptNet (Speer and Havasi, 2013), an automatically generated semantic graph containing concepts with their relations.

We test the language models in two ways: (1) We directly compare the statistics of the linguistic models with statistics extracted from the visual domain.

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 769–779,
Seattle, Washington, USA, 18-21 October 2013.

(2) We compare the linguistic models inside the two computer vision applications, leading to a direct estimation of their usefulness.

To summarize, our main research questions are: (1) Is the knowledge from language compatible with the knowledge from vision? (2) Can the knowledge extracted from language help in computer vision scenarios?

2 Related Work

Using high level knowledge to aid image understanding has become a recent interest in the computer vision community. Objects, actions and scenes are detected and localized in images using low-level features. This detection and localization process is guided by reasoning and knowledge. Such knowledge is employed to disambiguate locations between objects in (Gupta and Davis, 2008). From the defined relationships between nouns (e.g., above, below, brighter, smaller), the system constrains which region in an image corresponds to which object/noun. Similarly, (Srikanth et al., 2005) exploit ontologies extracted from WordNet to associate words and images and image regions. (Yu et al., 2011) employ relations between scenes and objects introducing an active model to recognize scenes through objects. The reasoning knowledge limits the detector to search for an object within a particular region rather than on the whole image.

Language models have also been employed to generate descriptive sentences for images. (Ushiku et al., 2012) introduce an online learning method for multi-keyphrase estimation to generate a sentence using a grammar model to describe an image. Similarly, from objects and scenes detected in an image, (Yang et al., 2011) estimated a sentence structure to generate a sentence description composed of a noun, verb, scene and preposition.

The studies most similar to ours are (Teo et al., 2012) and (Lampert et al., 2009). In (Teo et al., 2012), the Gigaword corpus is used to extract relationships between tools and actions (e.g., knife - cut, cup - drink) by counting their co-occurences. These relationships are used to constrain and select the most plausible actions within a predefined set of actions in cooking videos. Instead of using this knowledge as a guidance during recognition, we compare different language models and build a general framework that is able to detect unseen actions through their components (verb - object - scene), hence our method does not limit the number of actions in images. (Lampert et al., 2009) use attributes of nouns (e.g., an animal: white, eat fish, water, etc.). They can detect animals without having seen training examples by manually defining the attributes of the target animal. In this work, rather than relying on manual definitions, our aim is to find the best language models built automatically from available corpora to extract relations from natural language.

Currently, human action recognition is popular and mostly studied in video using the Bag-of-Visual-Words method (Delaitre et al., 2010; Everts et al., 2013; Kuehne et al., 2012; Reddy and Shah, 2012; Wang et al., 2013). In this method one extracts small local visual patches of, say, 24 by 24 pixels by 10 frames at every 12^{th} pixel at every 5^{th} frame. For each patch local gradients or local movement (optical flow) histograms are calculated. Then these local visual features are mapped to abstract, predefined "visual words", previously obtained using k-means clustering on a set of random features. While results are good, there are two main drawbacks with this approach. First of all, human actions are semantic and more naturally recognized through their components (human, objects, scene) rather than through a bag of local gradient/motion patterns. Hence we use a component-based method for human action recognition. Second, the number of possible human actions is huge (the number of objects times the number of verbs). Obtaining annotated visual examples for each action is therefore prohibitively expensive. So we learn from language models how components combine into human actions.

3 Two Visual Recognition Scenarios

We now describe the two computer vision scenarios: human action recognition and objects in context.

3.1 Human Action Recognition

We want to identify a human action, defined as a <subject, verb, object> triple. We do this by recognizing the human, the object, and the scene and then determine the most likely verb based on these components. Scenes are only used here as features for

predicting/disambiguating the human action and the final task is to define the human action triple. As in most work in human action recognition, we simplify the problem by considering only images in which human actions occur. This means that a human is always present, leaving the problem of predicting the verb given the object and the scene. While this may seem like a strong assumption, the possibility of having no action in the image at all is largely unexplored in computer vision due to its difficulty.

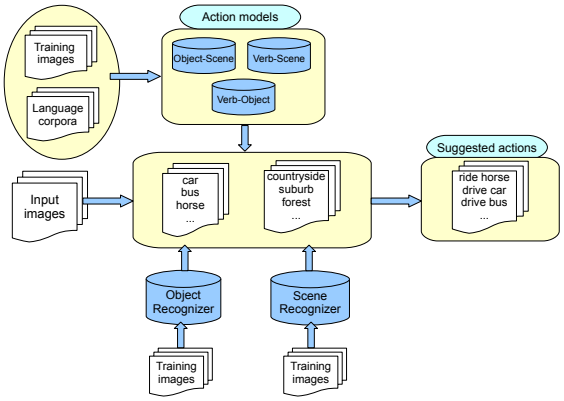

Figure 1: Human action suggestion: based on the objects and scenes recognized in an image, the system suggests the most plausible actions. The action models provide the relationships between objects - scenes - verbs

3.1.1 Human action recognition framework

Our general action recognition framework is presented in Figure 1. Given an image, an object recognizer will predict the probability of each object (e.g, bike, horse) presented in that image. Furthermore, a scene recognizer will provide the probabilities of each scene (e.g., countryside, suburb, forest) given the image. The action model is composed of the conditional probabilities that relate verbs, objects and scenes, which have been learned from training images or language corpora. Given the object and scene probabilities recognized in the image, the action model will guide the action prediction process and finally, the system will suggest the most proper actions (e.g., ride horse, drive car). We will now describe each component in detail:

- **Object and scene recognizers:** To train the object recognizer, we use a set of images where objects have been annotated with bounding boxes (to specify objects' locations). We follow the state-of-the-art method of (Uijlings et al., 2013). The method is based on multiple hierarchical segmentations to sample a limited set of high quality object locations in terms of bounding boxes. A Bag-of-Visual-Words method (Uijlings et al., 2010) is applied to these boxes to localize and recognize objects. For the scene recognizer, we trained the same Bag-of-Visual-Words method on complete images on a dataset annotated with 15 scenes. In both cases we use Support Vector Machines to learn the object/scene models. We use Platt's sigmoid function to obtain the final conditional probabilities $P(o_j|I)$ and $P(s_k|I)$.

- **Action models:** The model captures the relationship between Object - Scene, Verb - Scene and Verb - Object: the probability of an object given a scene $P(o_j|s_k)$, a verb given an object $P(v_i|o_j)$, and a verb given a scene $P(v_i|s_k)$. In one experiment we learn the probabilities from the training images, where each image has been annotated with an object, a verb (of an action) and a scene. All three probabilities are computed using frequency counts in the training set, for example:

$$P(o_j|s_k) = \frac{\#\text{images having } o_j, s_k}{\#\text{images having } s_k} \quad (1)$$

  We aim to replace this learning from annotated training images, which are expensive to obtain, with learning from language corpora. The details of how to extract the probability distributions from language models are explained in section 4.

3.1.2 Component integration

To combine these components in the framework, we use an energy-based model (Lecun et al., 2006) visualized in Figure 2, which includes the image I (an observed variable) and object O, scene S, and verb V. This energy-based formulation allows us to set different weights for energies which come from disparate sources (i.e. language and vision) using Gibbs measure.

Now given an image I, we can compute the score function $\mathbb{S}(a_{ij}; I)$ of an action a_{ij} as:

$$\mathbb{S}(a_{ij}; I) = \mathbb{S}(v_i, o_j; I) = \frac{1}{Z} \exp\left(-\sum_{F \in \mathcal{F}} E_F^{ij}\right),$$

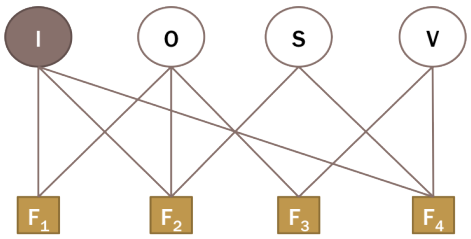

Figure 2: An energy-based model for action recognition

where we define each energy function E_F^{ij} to give lower energies to correct answers and higher energies to incorrect ones, with $\mathcal{S}$ is the set of all scenes:

$$E_{F_1}^{ij}(O, I) = -w_{F_1} \log P(o_j|I) \quad (2)$$

$$E_{F_2}^{ij}(S, I) = -w_{F_2} \log \sum_{S \in \mathcal{S}} P(o_j|S) \times P(S|I) \quad (3)$$

$$E_{F_3}^{ij}(V, O) = -w_{F_3} \log P(v_i|o_j) \quad (4)$$

$$E_{F_4}^{ij}(V, S) = -w_{F_4} \log \sum_{S \in \mathcal{S}} P(v_i|S) \times P(S|I) \quad (5)$$

Let P_i be the position of the correct action in the ranked list of predicted actions for a certain image I_i. The ranked list is sorted in the order of the score $\mathbb{S}$. We evaluate human action recognition in terms of this position average over all images, which we call Average Ranking (AR). Therefore we use Average Ranking as our loss-function:

$$\mathcal{L}(w_F) = AR_N = \frac{1}{N} \sum_{i=0}^{N} P_i. \quad (6)$$

Training the energy model involves finding the factors w_F^* that minimizes the loss:

$$w_F^* = \underset{w_F}{\operatorname{argmin}}\, \mathcal{L}(w_F) \quad (7)$$

As we have only four parameters to learn in our energy model, we do this by performing an exhaustive search and cross validation. We require $w_F \in \{0.0, 0.1, 0.2, ..., 0.9\}$ and set the constraint $\sum_{F \in \mathcal{F}} w_F = 1$. We note that the factor graph formulation of our framework would allow us to use more advanced learning algorithms. We plan to look into this once the model becomes more complex by adding, for example, information about the position of the objects and the human.

3.1.3 Dataset

Recently, researchers have released many image action datasets such as the 7 everyday actions (Delaitre et al., 2010), the Stanford 40 action dataset (Yao et al., 2011), the PASCAL action classification competition (Everingham et al., 2012), and the 89 action dataset (Le et al., 2013). The 89 action dataset was originally created for the recognition of 20 objects. Afterwards also actions were annotated. Therefore, the actions occurring with these objects are mostly unbiased, unlike in other action datasets. Hence we choose to use the 89 action dataset.

In the 89 action dataset, every image has been annotated with human actions, where each action is composed of a verb and an object. We additionally annotated every image with one of the 15 scenes from the 15 scene dataset (Lazebnik et al., 2006).

3.2 Objects in Context

Our other computer vision scenario is about objects in context. Context is useful in visual recognition for two reasons: Firstly, context can significantly reduce the number of possible object categories simplifying the problem. Secondly, when the object appearance is inconclusive for its identity, context can be used for disambiguation. For example, a grey rectangle on a desk may be recognized as a pen, while a grey rectangle on a table may be recognized as a knife. As the recognition systems are not always reliable, the use of context can greatly improve results.

For this scenario we choose a theoretical setting in which we want to predict the identity of one object given that the identities of all other objects in the image are known. We believe that our main conclusions on the linguistic models will transfer to a practical computer vision application where visual recognition systems predict the object identities.

Formally, we can describe this scenario as follows: Given an image I with N objects $\mathcal{O} = \{o_1, o_2, \cdots, o_N\}$, we want to predict the identity of object o_i given all other objects $\mathcal{O} \setminus o_i$. In this paper we use a Naive Bayes assumption, leading to:

$$P(o_i|\mathcal{O} \setminus o_i) = \frac{P(\mathcal{O} \setminus o_i|o_i) \times P(o_i)}{P(\mathcal{O} \setminus o_i)} \approx P(o_i) \times \prod_{o_j \in \mathcal{O} \setminus o_i} P(o_j|o_i). \quad (8)$$

In this scenario, we need conditional relations $P(o_j|o_i)$ and priors. We obtain these from language data or from images directly.

3.2.1 Dataset

For the objects in context scenario, we use the SUN object dataset (Xiao et al., 2010), which contains more than 16 thousand images, more than 79,000 objects whose locations are annotated using polygons. The dataset has been annotated by various people who could choose their own object categories, leading to duplicate categories such as "building" and "buildings", "person" and "person walking". Furthermore, for some images large parts are not annotated leading to an incomplete context. We therefore cleaned the object categories (mapping from around 7,500 objects to over 700 unique object categories) and considered only images whose content was sufficiently annotated.

In our experiments, we used the predefined training and testing parts of the SUN dataset and obtained around 4,500 images for learning the object relations and 10,600 images for testing the object prediction. We obtain conditional probabilities $P(o_j|o_i)$ from frequency counts.

4 Language models & distribution extraction

To extract probability distributions from texts, we use ConceptNet, the Window2 and 20 model, TypeDM and R-LDA. We will now describe how we estimate the four conditional probabilities $P(V|O), P(V|S), P(O|S), P(O|O)$ needed in the two visual scenarios for each language model.

4.1 ConceptNet

ConceptNet (Speer and Havasi, 2013) is a large semantic graph containing concepts and relations between them. It includes everyday basic, cultural and scientific knowledge, which have been automatically extracted from Internet using predefined rules. In this work we use the most current version, ConceptNet 5. As it was mined from free text using rules, the database has uncontrolled vocabulary and contains many false/nonesense statements.

To extract relations from ConceptNet5, we first examine all relations in the database and define those that are relevant to our scenarios (Figure 3). For

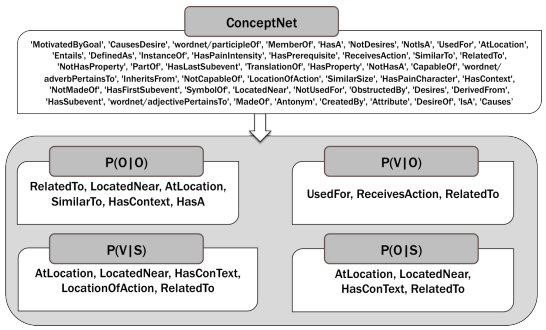

Figure 3: List of relations in ConceptNet

example, for the conditional probability of objects given scenes, relations such as "At Location", "Located Near" are extracted. For the human action recognition scenario, we used a list of 19 objects, 15 scenes and around 5 thousand verbs for computing $P(V|O), P(O|S), P(V|S)$. For the objects in context scenario, we used 700 objects for computing $P(O|O)$. Examples of relations extracted from ConceptNet are illustrated in Table 1, such as: Oil - Located near - Car, Horse - Related to - Zebra. From these relations, we define the four conditional probabilities using their frequency counts. For example, to compute the conditional probability of an object given a scene $P(o_i|s_j)$, we extract all triples having the form <object, rel, scene>, where "rel" can be "AtLocation", "LocatedNear", etc.

$$P(o_i|s_j) = \frac{freq(< o_i, rel, s_j >)}{\sum_{o_m \in \mathcal{O}} freq< o_m, rel, s_j >} \quad (9)$$

4.2 Window model

One of the most famous and basic statistical model is based on counting co-occurrences within a window of fixed width, which follows the tradition of hyperspace analogue to language (Lund and Burgess, 1996). We took the Window2, 20 models which have been built in (Bruni et al., 2012) using the ukWaC (1.9B tokens) and Wackypedia (820M tokens). As the Window2 model only looks at 2 words on the left and right of the current one, it reflects the relationships between words occurring near each other, while the Window20 searches for a broader view of how words are related to each other. The weights of each pairs of words are calculated using the Local Mutual Information (LMI). To compute the conditional probabilities, we use the LMI scor-

LocatedNear		RelatedTo		UsedFor		AtLocation	
oil car	seatbelt car	horse zebra	plant garden	bottle store_liquid	horse race	bus city	car city
chair your_bottom	chair school	horse pony	sheep baa	boat fish	table eat_off_of	bike street	dog city
plant everywhere	muzzle dog	plant green	sheep cloud	dog companionship	chair rest	bird countryside	dog street
trailer car	dog bark_bone	boat ship	cow bull	horse riding	bus travel	car street	chair city
salt table	horse cowboy	chair table	horse riding	chair sitting	table eat_meal	cat store	bus city
stool table	carriage horse	dog wolf	sheep farm	chair sit_on	boat travel	car street	chair store
pasture cow	horse fence	dog cat	cow milk	car transportation	bottle hold_liquid	car street	bicycle store
cat dog	whisker cat	sheep lamb	table desk	sheep wool	boat float_on_water	bird forest	chair store
horse zebra	desk chair	sheep wool	cat feline	table put_thing_on	table eat_at	car city	bottle store
cat household	train railroad	dog a wolf	dog canine	boat travel_on_water	cat catch_mouse	table kitchen	chair office
horsehair horse	sheep wool	cat dog	plant flower	chair sit	cow milk	chair office	chair city

Table 1: Examples of relations extracted from ConceptNet 5

ing function provided by the models, for example:

$$P(v_i|o_j) = \frac{LMI_{v_i,o_j}}{\sum_{v_m \in \mathcal{V}} LMI_{v_m,o_j}} \quad (10)$$

4.3 Distributional Memory

Distributional Memory (Baroni and Lenci, 2010) (DM) is a multi-purpose framework for semantic modeling. This model is more complex than the Window models because it exploits different degrees of lexicalization for each relation. Distributional information is extracted as a set of weighted <word-link-word> tuples obtained from a dependency parse of corpora. In the Window model the relation between each word pair is decided by their co-occurences within a sliding window, while in DM this relation is defined by distributional properties of the two words. These distributional properties are based on a syntactic relation or lexico-syntactic pattern that links the two words. For example, the tuple <marine, use, bomb> encodes that marine co-occurs with bomb in the corpus, and the word "use" specifies the type of the syntagmatic link.

Distributional Memory contains three different models, corresponding to different ways to construct the weighted structure through the "link". The first model, LexDM is the most heavily lexicalized model with the most variety of links, whereas the DepDM has the minimum degree of lexicalization, thus having the smallest number of links. TypeDM, which was reported to achieve the best performance in different tasks including selectional preferences, is laying somewhere in the middle of the other two models. It shares the same lexical information as in LexDM but use a different scoring function, which focuses on the variety of surface forms, rather than the frequency of a link. Hence we choose the best model, TypeDM, to learn the relationships between verbs, objects and scenes. As in the window model, we compute conditional probabilities using the LMI scores provided by the model (Equation 10).

4.4 R-LDA

To model the relationships between verbs, objects and scenes, we adapt the R-LDA model (Séaghdha, 2010) (ROOTH-LDA), which has been used for the selectional preference task in order to obtain conditional probabilities of two words. Each relation m of $< w_1, w_2 >$ is generated by picking up a distribution over topics, then both elements of the relation m share the same topic assignment z_m, which keep two different w_1-topic and w_2-topic distributions sharing the same topic (Figure 4). The models are estimated by Gibbs sampling following (Heinrich, 2004). It is also noted that these models are generative, hence they also predict the probabilities of tuples that do not occur in the corpus.

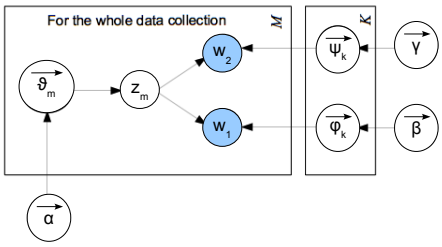

Figure 4: Generative graphical model of R-LDA: modeling the relations between two words

To model the relations between objects and verbs, we follow the data preparation in (Le et al., 2013), using the British National Corpus (BNC) which has been preprocessed and parsed using TreeTagger and Maltparser. Verbs are heads of sentences while objects are either direct or indirect objects related to those verbs by the parser. For the relations between verbs and scenes, we consider also verbs as heads

Topic 8:				Topic 14:				Topic 0:				Topic 54:			
Noun		Verb		Noun		Verb		Noun		Noun		Noun		Noun	
people	0.0208	have	0.157	year	0.0154	win	0.109	attention	0.0172	study	0.01	decision	0.02	case	.0176
job	0.0167	work	0.108	Cup	0.0093	have	0.099	model	0.0147	research	0.0123	view	0.0244	fact	.0096
work	0.0156	make	0.0262	team	0.0086	beat	0	study	0.014	work	0.0111	question	0.018	question	.0096
class	0.014	take	0.0244	race	0.00772	take	0.025	role	0.0139	chapter	0.0085	issue	0.0124	law	.0096
worker	0.0123	find	0.0194	season	0.0062	lose	0.0211	account	0.013	problem	0.0075	evidence	0.0104	decision	.0092
staff	0.0111	pay	0.0156	time	0.006	run	0.0152	analysis	0.0123	issue	0.006	point	0.0099	time	.0067
group	0.0089	say	0.0146	world	0.0058	finish	0.0135	aspect	0.012	system	0.0065	reason	0.0096	issue	.0062
way	0.0086	get	0.0136	game	0.0055	make	0.0127	problem	0.0106	area	0.006	statement	0.0086	evidence	.00617
service	0.008	leave	0.0114	champion	0.0053	lead	0.0122	effect	0.0105	process	0.006	doubt	0.008	interest	.0059
company	0.0076	run	0.0102	seat	0.0049	follow	0.0098	pattern	0.0103	policy	0.0055	attention	0.0076	point	.0058
day	0.0069	come	0.0101	match	0.0049	qualify	0.008	issue	0.0102	theory	0.00551	matter	0.00738	judge	.0055
number	0.00615	help	0.0088	place	0.0047	compete	0.0074	range	0.0095	way	0.0053	policy	0.006	statement	.005

Table 2: Random R-LDA topics with the relations between Noun-Verb (first 2 columns) and between Noun-Noun (last 2 columns)

of sentences while scenes are all nouns occurring in the same sentence. For the relations between objects and scenes as well as objects and objects, we use all nouns to capture a general model[1]. The statistics of the BNC corpus with their corresponding relations are reported in Table 3.

	#Relations	#Tokens
Verb - Object	3.3M	6.7M
Noun - Noun	19.8M	39.7M
Verb - Noun	83.4M	166.8M

Table 3: The statistics of the dataset used for estimating R-LDA models for each relation type

Samples of topics extracted through R-LDA are illustrated in Table 3. It shows that Noun and Object share many similar terms in the same topic while Noun and Verb sharing the same topics tend to go often together (e.g., win, cup, beat, race).

5 Experiments

In this section we want to answer our two main research questions: (1) Is knowledge from language compatible with knowledge from vision? (2) Can we use knowledge extracted from language in computer vision scenarios?

5.1 Is knowledge from language and vision compatible?

In this section we compare statistics mined from texts with those mined from visual sources. Ideally, we want statistics from the language models to follow those of the image model, even though not all statistics from images can be reliably measured due to insufficient data. Therefore, we measure how well the estimated language models fit the estimated visual distributions using the the χ^2-distance:

$$\mathcal{X}^2 = \sum_{i=1}^{N} \frac{({P_I}^i - {P_L}^i)^2}{{P_I}^i} \tag{11}$$

where P_I and P_L are the probability distribution obtained from the image data and language models respectively.

For the conditional probabilities $P(V|O)$, $P(V|S)$, and $P(O|S)$ we compare language models with image statistics extracted from the 89 human action dataset. Table 4 shows the results. For the relations between verb and scene $P(V|S)$, there is not much fluctuation among different language models. For objects and scenes $P(O|S)$, R-LDA is closest to the image model. This is because R-LDA is good at measuring contextual and indirect relations by design, which is the case for object-scene relations. This also explains why TypeDM and Window20 are

Chi statistics	P(V|O)	P(V|S)	P(O|S)
R-LDA	17.8	11.6	**11.9**
Window2	11.6	**11.4**	32.6
Window20	11.7	**11.4**	23.7
TypeDM	**11.5**	13.3	23.2
ConceptNet	17.5	11.5	34.4

Table 4: $\mathcal{X}^2$ distance for relations between verbs, objects, scenes from different language models to image data

[1]Different from objects and verbs, which can be defined explicitly from the parsed corpora, scenes can only be defined from more restrained rules (e.g., followed by some prepositions), so here we take all nouns to have the most general model.

further away from the image model, followed by the Window2 model. Instead, human actions are found in language as the relation between verbs and their direct linguistic objects. Indeed, TypeDM is closest to the image model for $P(V|O)$ as it makes explicit use of this linguistic link. The Window2 and 20 models are almost as close to the image model for $P(V|O)$, while R-LDA is considerably further away due to its contextual nature. Finally, ConceptNet is the furthest away from the image model. To conclude, TypeDM is best for modelling direct verb-object relations, while R-LDA is better at capturing the more contextual object-scene relations.

To look closer at the difference between the statistics obtained from the image and language data, we give an example of the conditional probabilities of an object given a scene $P(O|S)$ in Figure 5. We see that the distribution extracted from language (TypeDM) is much smoother and contains more relations than the image model since it has been trained on general and large text corpora. The distribution from image data on the other hand is more sparse and tailored to this specific dataset. For example, given a "store", the probability that there is a "table" is 1, given "highway", the probability of a "car" is also 1 in the image dataset, while the highest conditional probability of the language model is only less than 60%.

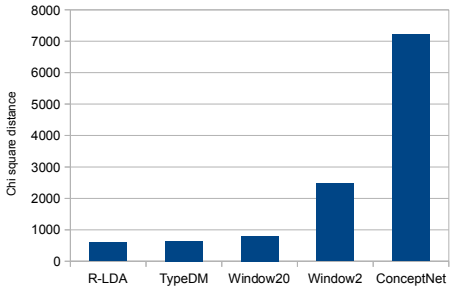

Figure 6: χ^2-distances between the tested language models and the image model for conditional probabilities of objects $P(O|O)$.

For the relations between objects and objects, we use the SUN dataset, which is much bigger and more general than the action dataset. As shown in Figure 6, R-LDA is most similar to the image model, closely followed by TypeDM and the Window20 model. All these three models are good at capturing broad contextual relations. The Window2 model has a significantly larger distance to the image model as it captures a narrow context of 2 words, which is apparently not enough to find co-occurences of objects. ConceptNet is the most inconsistent with this image data since not enough objects and their relations are extracted from it. To sum up, R-LDA achieves the best performance in modeling the relations between objects and objects among all language models.

5.2 Language Models for Visual Recognition

To measure the performance of the two visual recognition scenarios, we use the position p^i of the correct action found in the ranked list for each image i.

We report the average ranking over all images (AR_I) and over all objects or actions (AR_O, AR_A):

$$AR_I = \frac{\sum_{i=0}^{N} p^i}{N}; AR_O = \frac{\sum_{j=0}^{N_o} p_o^j}{N_o} \quad (12)$$

where N is the number of images, N_o is the number of objects and p_o^j is the average rank of all images having object j. The average rank over all actions AR_A is defined similarly to AR_O. The average rank over all image measures the performance over the image dataset, but infrequent objects/events have little impact on this performance. The average rank over objects or actions gives more weight to rare examples.

5.2.1 Human Action Recognition

We evaluate the performance of human action recognition in images based on objects and scenes individually, and then study the integration of them. The training set contains 1,104 images (for training the image relations) and the test set has 710 images. First, we test how the model predicts an action knowing the actual object and/or scene appearing in an image (given object/scene gold standard), i.e., O_{gs}, S_{gs} and $O_{gs}S_{gs}$ in the settings. After that, we test a complete model which is based on the output of our object recognizer and our scene recognizer (O_{rec}, S_{rec}, $O_{rec}S_{rec}$).

For each setting, we try different action models, either learnt from the training images (Image), or from each of the language models (TypeDM, R-LDA, Window2, Window20, ConceptNet).

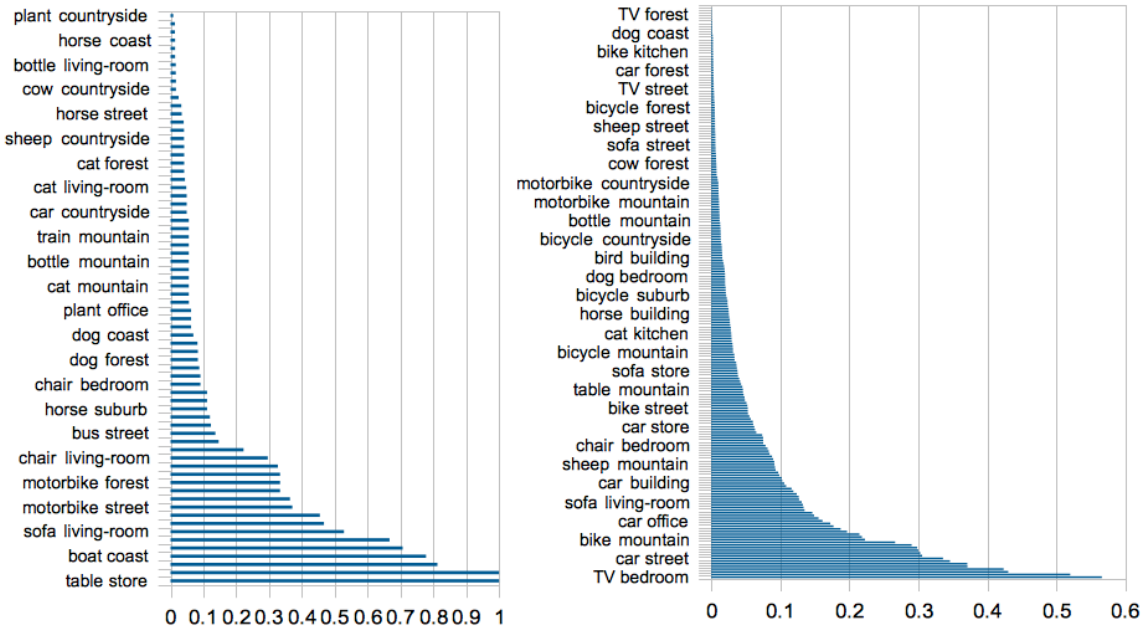

Figure 5: Probability distributions of scene over object extracted from: (left) image dataset; (right) TypeDM model (as there are many <object - scene> relations, only a few are shown on the Y-axises). The number of relations in the TypeDM is much bigger than in the image model, which shows a more general model than the image one.

Table 5 presents the average ranking over all images. Results show that the action model learnt directly from the training images achieves the best performance in all settings, even if we give more weight to infrequent actions by taking the average ranking over all actions, as presented in Table 6. One explanation may be that the action dataset has a limited domain of only 19 objects, while the language models were learnt from broad knowledge (See Figure 5). Another possibility is that verbs used for describing actions in images are more specific than verbs used in language. For example, in language one would "use the car", while in images such action would be labelled "drive a car".

	Image	**TypeDM**	**R-LDA**	**Window2**	**Window20**	**C.Net**
O_{gs}	0.3	16.1	63.4	16.4	18.3	86.1
O_{rec}	14.9	26.9	66.7	44.7	54.9	115.6
S_{gs}	35.7	181.7	174.9	168.5	174.8	252.5
S_{rec}	46.8	250.5	348	190.2	189.8	241.2
$O_{gs}S_{gs}$	**0.28**	**10.2**	**15.2**	**13.8**	**13.6**	**81.9**
$O_{rec}S_{rec}$	13.6	26.9	66.7	44.7	54.9	115.6

Table 5: Average rank over all images AR_I of the human action recognition using different settings: O_{gs}, O_{rec} use only objects (gold standard and object recognizer); S_{gs}, S_{rec} use only scenes, $O_{gs}S_{gs}$ and $O_{rec}S_{rec}$ integrate both objects and scenes together

If we look at the performance of the language models, TypeDM performs best by a significant margin. This makes sense, as the most powerful term for predicting an action is obviously $P(V|O)$, and we saw earlier that TypeDM produces probabilities $P(V|O)$ which are closest to the image model. For the same reason, the second and third best language model are the Window2 and Window20 models, although their performance is significantly lower when using the predictions for objects and/or scenes. This is somewhat surprising considering that TypeDM, Window2 and Window20 are all very close in distance to the image model. Of course, the distance is just an indication. R-LDA performs poorly because it is much more contextual. Finally, ConceptNet performs the worst.

Another observation is that using the scene identity should theoretically help in human action recognition: Using TypeDM, the use of the gold standard object identity yields an average ranking over all images of 16.1, while using both the scene and object identity yields an average ranking of 10.2, which is significantly better. It means that the use of the scene can disambiguate some actions (e.g. "ride a horse" vs. "feed a horse"). However, when using the recognition system, using the scene does not increase the overall performance. This shows that the

visual recognition system may not be strong enough for recognizing these 15 scenes. Another problem may be the limitation of 15 scenes only: while annotating we frequently found that it was hard for numerous images to put them into one of the 15 scenes. So a bigger scene database may help.

The main problem with most available annotated human action datasets is that they are very restricted and domain-specific. For example, in this dataset with 19 objects and 15 scenes, there are many photos of a person riding a motorbike on rocky mountains as a kind of sport. Consequently, the probability of "riding" given "mountain" learnt from the image dataset is high according to the image data (78%) but is uncommon in general. So the image dataset might be too restricted or biased for general knowledge to work well. In the next section we therefore use a more general dataset.

	Image	**TypeDM**	**R-LDA**	**Window2**	**Window20**	**C.Net**
AR_I	13.6	26.9	66.7	44.7	54.9	115.6
AR_A	16.4	30.8	64.7	45.3	51.9	131.7

Table 6: Average rank over all images vs. actions of the human action recognition using the $O_{rec}S_{rec}$ setting

5.2.2 Objects in Context

For every object in every image in the test set of the SUN database, we guess the identity of an object given the identity of all other objects in the image. In total, there are 78,306 object predictions within 10,652 images.

As shown in Figure 7, the R-LDA model outperforms all other models for both average rank over images and over objects. Interestingly, both R-LDA and TypeDM are better at predicting the correct objects in images than the model learnt from the image training set itself. It shows that for many cases, the relation statistics learnt from language data can help in visual recognition. These language models are even better than the information extracted from general, relatively unbiased image datasets, where annotation is limited. For the limited annotation, this hypothesis is further supported by looking at the average rank over objects, which gives more weight to rarely occurring objects. As seen in Figure 7, all language models except ConceptNet outperform the image model. We conclude that language models

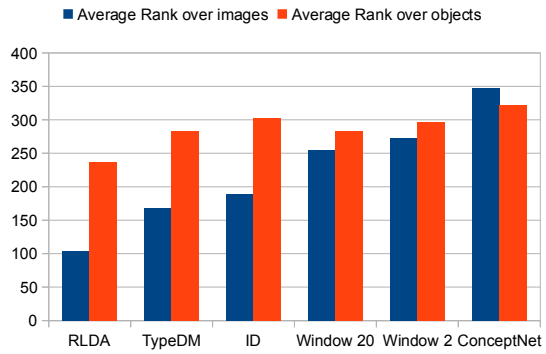

Figure 7: Average rank over all images and objects using different language models and ID (image data)

can aid visual models in large-scale visual recognition problems which use co-occurrence of objects as their context, especially when the annotation is limited, as is often the case.

6 Conclusion

In this paper, we investigated the problem of applying knowledge learnt from language corpora to visual recognition. We compared statistics of various language models mined on general corpora with statistics observed in image datasets. It shows that the generative R-LDA model is good at relating contextual relations (e.g., object - object, object - scene), while the syntactic based distributional model TypeDM is good at representing direct relations such as verb - object in images.

We have evaluated the performance of the language models in two visual scenarios: human action recognition and object prediction. It suggests that the language models need some tailoring when applied to restricted datasets, but for a bigger and more general dataset, the language models even outperform the model learnt from annotated images itself. This shows that language models built from available text corpora can be used for visual recognition instead of expensive annotated image data.

In the future, we want to further investigate the problem of domain adaptation when applying general language models to a new image dataset. This problem can be integrated into the energy-based model during the training phase. We plan to extend work on human action recognition by including the relative position between the human and object in the images.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, ACL. ACL.

Vincent Delaitre, Ivan Laptev, and Josef Sivic. 2010. Recognizing human actions in still images: a study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference*. BMVA Press.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.

I. Everts, J. van Gemert, and T. Gevers. 2013. Evaluation of color stips for human action recognition. In *CVPR*.

Abhinav Gupta and Larry S. Davis. 2008. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *Proceedings of the 10th European Conference on Computer Vision*.

Gregor Heinrich. 2004. Parameter estimation for text analysis. Technical report.

H. Kuehne, D. Gehrig, T. Schultz, and R. Stiefelhagen. 2012. On-line action recognition from sparse feature flow. In *VISAPP*.

C.H. Lampert, H. Nickisch, and S. Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Conference on Computer Vision and Pattern Recognition CVPR*.

S. Lazebnik, C. Schmid, and J. Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Conference on Computer Vision and Pattern Recognition CVPR*, volume 2, pages 2169–2178.

Dieu Thu Le, Raffaella Bernardi, and Jasper Uijlings. 2013. Exploiting language models to recognize unseen actions. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval ICMR*. ACM.

Yann Lecun, Sumit Chopra, Raia Hadsell, Fu J. Huang, G. Bakir, T. Hofman, B. Schölkopf, A. Smola, and B. Taskar Eds. 2006. A tutorial on energy-based learning. In *Predicting Structured Data*.

Kevin Lund and Curt Burgess. 1996. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*.

K. Reddy and M. Shah. 2012. Recognizing 50 human action categories of web videos. In *Machine Vision and Applications*.

Diarmuid Ó. Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 435–444. Association for Computational Linguistics.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*. Springer Berlin Heidelberg.

Munirathnam Srikanth, Joshua Varner, Mitchell Bowden, and Dan Moldovan. 2005. Exploiting ontologies for automatic image annotation. In *Special Interest Group on Information Retrieval SIGIR*. ACM.

C.L. Teo, Yezhou Yang, H. Daume, C. Fermuller, and Y. Aloimonos. 2012. Towards a watson that sees: Language-guided action recognition for robots. In *2012 IEEE International Conference on Robotics and Automation ICRA*.

J R R Uijlings, A W M Smeulders, and R J H Scha. 2010. Real-time Visual Concept Classification. *IEEE Transactions on Multimedia*, 12.

J R R Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders. 2013. Selective search for object recognition. *International Journal of Computer Vision*.

Yoshitaka Ushiku, Tatsuya Harada, and Yasuo Kuniyoshi. 2012. Efficient image annotation for automatic sentence generation. In *ACM International Conference on Multimedia ACM MM*.

H. Wang, A. Kläser, C. Schmid, and C. Liu. 2013. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103:60–79.

Jianxiong Xiao, J. Hays, K.A. Ehinger, A. Oliva, and A. Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition CVPR*.

Yezhou Yang, Ching Lik Teo, Hal Daumé, III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Conference on Empirical Methods in Natural Language Processing EMNLP*.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Li Fei-Fei. 2011. Action recognition by learning bases of action attributes and parts. In *International Conference on Computer Vision ICCV*.

Xiaodong Yu, Cornelia Fermuller, Ching Lik Teo, Yezhou Yang, and Yiannis Aloimonos. 2011. Active scene recognition with vision and language. In *Proceedings of the 2011 International Conference on Computer Vision*, International Conference on Computer Vision ICCV.