Learning Functional Distributional Semantics with Visual Data

Yinhong Liu, Guy Emerson

Email: {yl535,gete2}@cam.ac.uk

Website: https://github.com/williamLyh/PixieVGModel

UNIVERSITY OF CAMBRIDGE

Abstract

Functional Distributional Semantics is a recently proposed framework for learning distributional semantics that provides linguistic interpretability. It models the meaning of a word as a binary classifier rather than a numerical vector. In this work, we propose a method to train a Functional Distributional Semantics model with grounded visual data. We train it on the Visual Genome dataset, which is closer to the kind of data encountered in human language acquisition than a large text corpus. On four external evaluation datasets, our model outperforms previous work on learning semantics from Visual Genome.

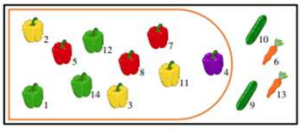

Fig. 1 (from [Emerson, 2020a]): An example model structure with 14 individuals. Subscripts distinguish individuals with identical features, but are otherwise arbitrary. The pepper predicate is true of individuals inside the orange line, but the positions of individuals are otherwise arbitrary.

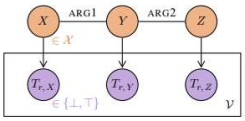

Fig. 2 (from [Emerson, 2020b]): A situation composed of three individuals. Top row: the *X*, *Y* and *Z* lie in a semantic space. They are jointly distributed according to the semantic roles ARG1 and ARG2. Bottom row: each predicate c in the vocabulary V has a stochastic truth value for each individual.

Fig. 3: An example image in Visual Genome, annotated with the relation ['Computer', 'ON', 'Desk']. 'Computer' and 'Desk' are labelled by yellow bounding boxes.

Method

World model

- The world model learns the distribution of situation, or in another words, the joint distribution of pixies, with a Gaussian Markov Random Field (MRF).
- The individuals are grounded by images, so the pixie vectors are obtained by extracting visual features with the pretrained **ResNet101**, from their corresponding images and reducing dimension with **PCA**.
- The world model is optimized to maximize the **log-likelihood of generating the observed situation.** The Maximum Likelihood Estimate (MLE) of the Gaussian parameters has a closed-form solution.

Lexicon Model

- The lexicon model learns a list of **semantic functions**, each corresponds to a word in predicate vocabulary.
- Each predicate is a **logistic regression** classifier over the pixie space. In another words, a single neural net layer with a sigmoid activation function. **All semantic functions are applied to each pixie.** A single predicate is generated according to the truth values. The more likely a predicate is to be true, the more likely it is to be generated.
- The lexicon model is optimized to maximize the **log-likelihood of generating the predicates given the pixies**. This can be done by gradient descent.

Variational Inference

- We provide an inference model to infer **latent pixie distributions given observed predicates**.
- The posterior distribution is intractable, so we use a **variational inference** algorithm to approximate it. The approximate distribution is optimized to maximize the Evidence Lower Bound (ELBO).
- Each pixie is **jointly inferred** based on all predicates in the triple. For example, as shown in Fig.5, the truth of 'horse' for X also depends on the observed predicate 'tail' or 'paw'. This is not a direct dependence between words, but rather relies on three intermediate representations (the three pixies).

Evaluation

External Dataset (subset)

- Lexical similarity datasets: MEN and SimLex-999.
- Contextual datasets: RELPRON and GS2011.
- Evaluation metrics: Spearman correlation and Mean Average Precision.

Baselines

- Large corpus baselines: Word2vec models and Glove.
- Visual Genome baselines: a Count-based model, a skip-gram model trained on VG (EVA) [Herbelot, 2020] and an image-retrieval baseline.

Results:

- We achieve a new state of the art on learning lexical semantics from Visual Genome. Our model can understand more semantics because it learns from the visual information and leverages textual cooccurrence.
- Compared to other VG baselines, our model is less affected by data sparsity and has advantage of learning similarity (compared to relatedness) from visual features.
- Our model can use parameters and data more effectively and efficiently than Word2vec and Glove, achieving acceptable performance with less training data and fewer parameters.

Conclusions

- Functional Distributional Semantic is a framework that provide interpretability.
- We demonstrate an approach to train the Functional Distributional Semantics framework with visual data.
- Our framework achieves SOTA performance on learning semantics from Visual Genome dataset.
- Our model can use parameters and data more efficiently than Word2vec and Glove.

Background

Functional Distributional Semantics (FDS)

- Functional Distributional Semantics was first proposed by [Emerson, 2016]. FDS separates the modeling of words and individuals, and it defines meaning in terms of truth.
- An individual is represented in a high-dimensional feature space. The representation is called **pixie**. The meaning of a content word is called **predicate**. The predicate is formalized as a **binary classifier** over pixies. It assigns true if an individual could be described by it, and false otherwise. Such a classifier is called a **semantic function**.
- Therefore, the model is separated into a **world model** and a **lexicon model**. The world model defines a distribution over situations. Each situation consists of a set of individuals, connected by semantic roles (ARG1 and ARG2). The lexicon model consists of semantic functions of all predicates in the vocabulary.
- The **motivations** of this framework is that FDS is **interpretable** in formal semantic terms and supports first-order logic.

Motivations of visual grounding

- Grounding connect the model to the physical world, which provides more interpretability.
- Grounding the individuals in the FDS is more accurate than grounding words.
- The Visual Genome dataset is considered similar to the data encountered during language acquisition.

Visual Genome dataset

- The Visual Genome dataset contains over 108K images.
- We only use the relation annotations, which is formulated as predicate triples: [Subject, Relation, Object]. In Fig. 3 we shown an example of a situation with pixie ['Computer', 'ON', 'Desk']. 'Computer' and 'Desk' are ARG1 and ARG2 to 'ON' respectively.

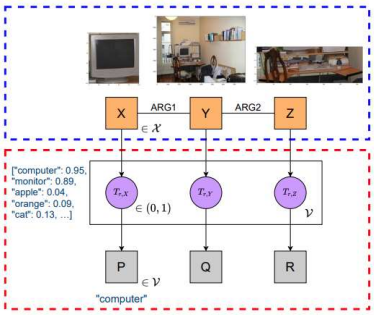

Fig. 4: Our probabilistic graphical model. The top blue box contains the world model, which learns the joint distribution of the observed pixies X, Y and Z from their corresponding images. The bottom red box shows the lexicon model, where each semantic function in the vocabulary V is applied to each pixie. For each pixie, one predicate is generated, with probability proportional to the probability of truth.

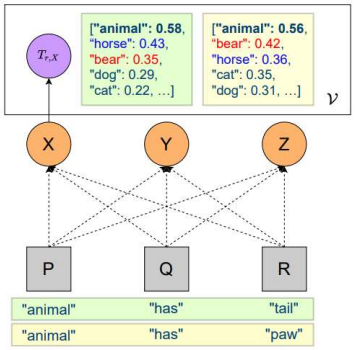

Fig. 5: : Graphical inference model: The pixies X, Y and Z in the middle row are jointly inferred from the observed predicates P, Q and R in the bottom row, using variational inference. Semantic functions are applied to X, to give probabilities of truth. As well as the observed predicate 'animal', the model predicts that other predicates may also be true. Two examples are given, in green and yellow, showing how the predicted truth values for 'tail' and 'paw' depend on all observed predicates. All probability values are inferred from our trained model.

	Models	Lexical datasets		Contextual datasets	
		MEN	**SimLex-999**	**GS2011**	**RELPRON**
Large corpus baselines	Word2vec-1B	0.641	0.384	0.265	0.381
	Word2vec-6B	0.652	0.397	0.278	0.401
	Glove-6B	**0.717**	0.409	**0.293**	**0.432**
VG baselines	VG-count	0.336	0.224	0.063	0.038
	VG-retrieval	0.420	0.190	0.072	0.045
	EVA	0.543	0.390	0.068	0.032
Proposed approach	Our model	0.639	**0.431**	0.171	0.117

Table 1: Evaluation results. For MEN, SimLex-999 and GS2011, the metric is Spearman correlation; for RELPRON, mean average precision. All models are evaluated on subsets of the data covered by the VG vocabulary.

References (selected)

1. Guy Emerson. 2020a. *Autoencoding pixies: Amortised variational inference with graph convolutions for Functional Distributional Semantics.*
2. Guy Emerson. 2020b. *Linguists who use probabilistic models love them: Quantification in Functional Distributional Semantics.*
3. Guy Emerson and Ann Copestake. 2016. *Functional distributional semantics.*
4. Elizaveta Kuzmenko and Aurélie Herbelot. 2019. *Distributional semantics in the real world: Building word vector representations from a truth-theoretic model.*
5. Aurélie Herbelot. 2020. *Re-solve it: simulating the acquisition of core semantic competences from small data.*
6. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. *ImageNet: A large-scale hierarchical image database.*