Learning Functional Distributional Semantics with Visual Data

Yinhong Liu, Guy Emerson

Department of Computer Science and Technology,
University of Cambridge

Summary

- 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.

Functional Distributional Semantics

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**.

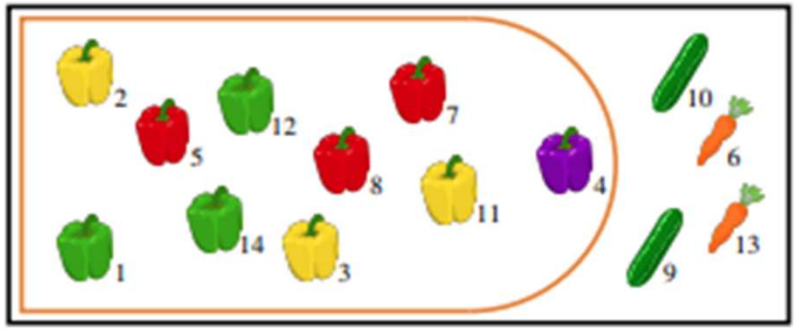


Fig. from [Emerson, 2020]

- Each item is an individual.
- The red box represents the semantic function of predicate ‘pepper’.

Functional Distributional Semantics

- 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.

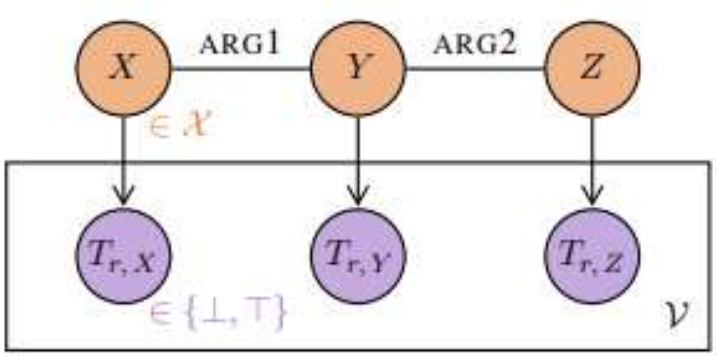


Fig. from [Emerson, 2020]

- **Motivations** of the framework:
  - FDS is **interpretable** in formal semantic terms and supports first-order logic.

Visual Grounding

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.

Figure 1: An example image in Visual Genome, annotated with the relation ['Computer', 'ON', 'Desk']

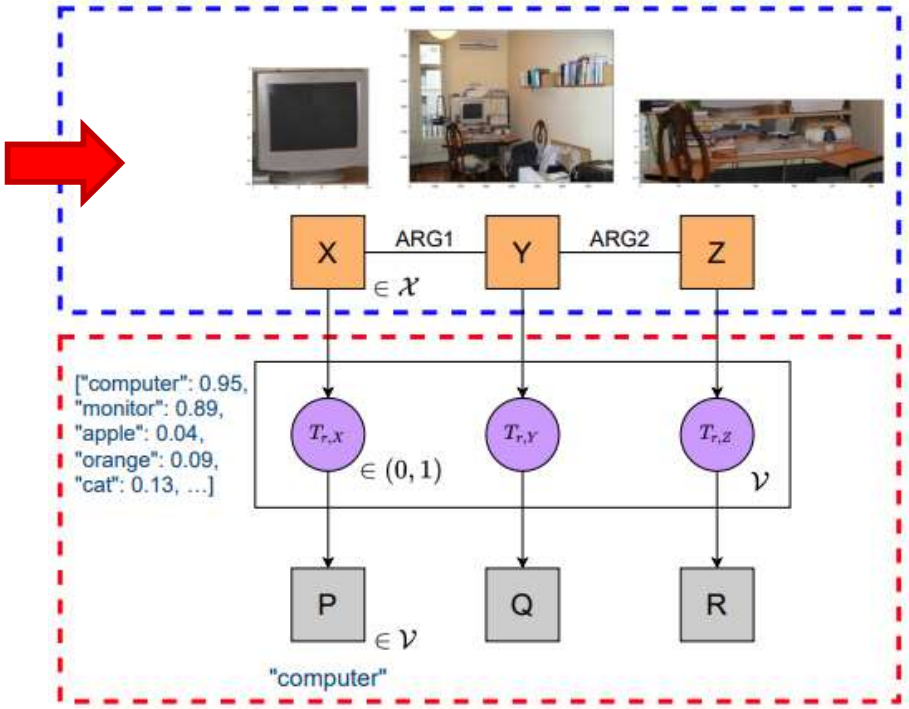
Visual Genome datasets

- Visual Genome contains over 108K images.
- Only use the relation set, formulated as predicate triples: [Subject, Relation, Object].
- The objects are identified with bounding boxes.

Our approach

World model

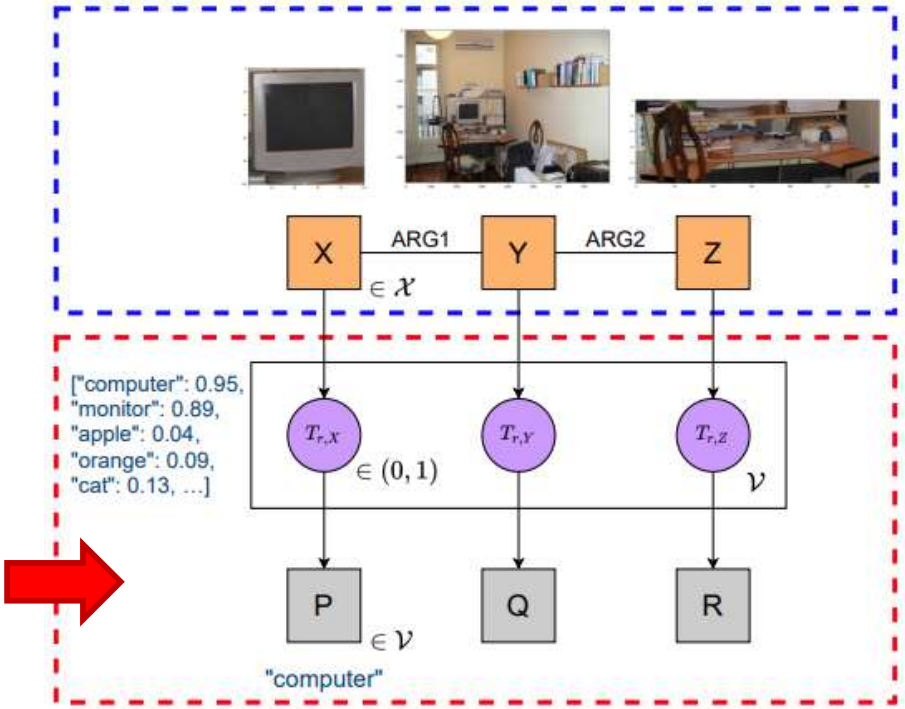
- The world model models the joint distribution of pixies with a Gaussian Markov Random Field (MRF).
- We obtain the pixie vectors by extracting visual features with the pretrained **ResNet101**, from their corresponding images and reducing dimension with **PCA**.
- The world model is optimized to maximize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



Our approach

Lexicon Model

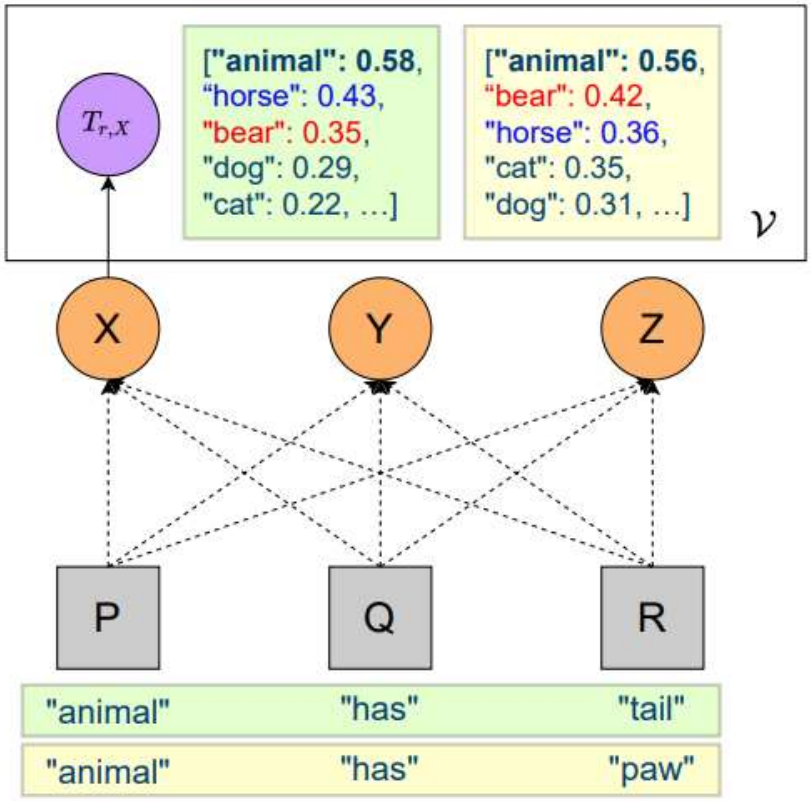
- The lexicon model learns a list of **semantic functions**, each corresponds to a word in predicate vocabulary.
- Each predicate is a **logistic regression** classifier over the pixie space. In another words, a single neural net layer with a sigmoid activation function.
- **All semantic functions are 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.



Our approach

Variational Inference

- We provide an inference model to infer **latent pixie distributions given observed predicates**.
- However, 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, 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
  - SimLex-999
- Contextual datasets:
  - RELPRON
  - 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) and an image-retrieval baseline.

Evaluation – External Dataset

	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.

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**.