Tutorial:
Functional Distributional Semantics

Guy Emerson

What I'll Cover...

- Theoretical background
- Running the code (Pixie Autoencoder)
- Future/ongoing work

Background

- Distributional semantics
  - The context of a word gives us information about its meaning
  - See: “What are the goals of distributional semantics?” (ACL 2020)
- Truth-conditional semantics
  - Words are not entities

Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to

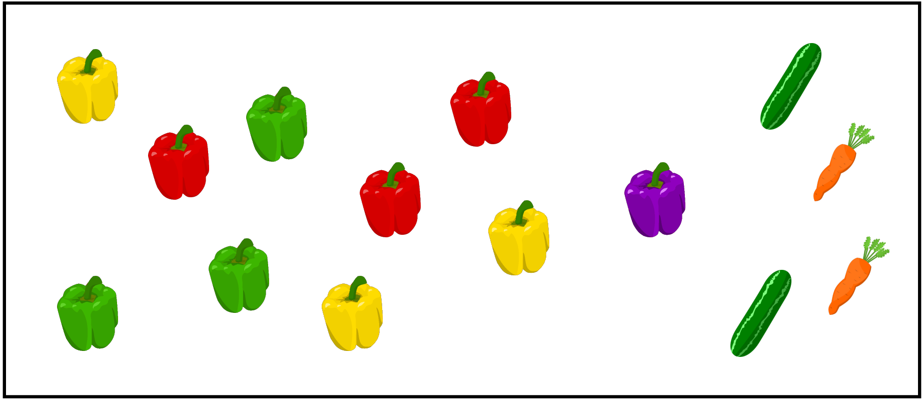
Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to
- Important for discourse: anaphora resolution, question answering, dialogue processing…

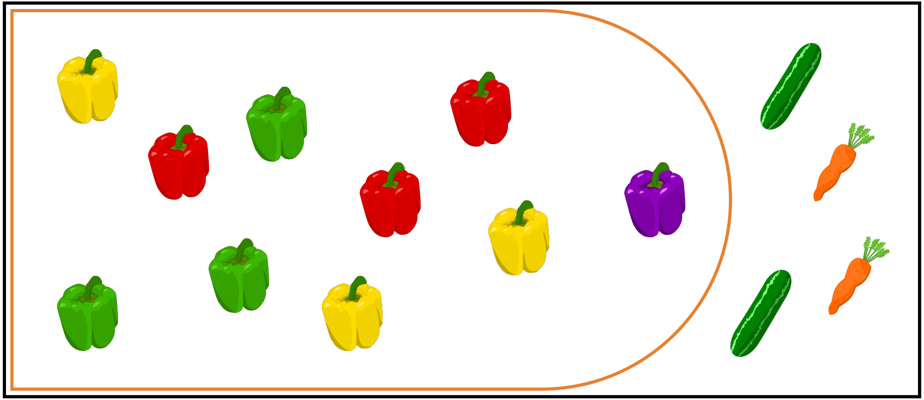
Words are not Entities

- Fundamental distinction between:
  - Words
  - Entities they refer to
- Important for discourse: anaphora resolution, question answering, dialogue processing...
- Meaning as a function over entities

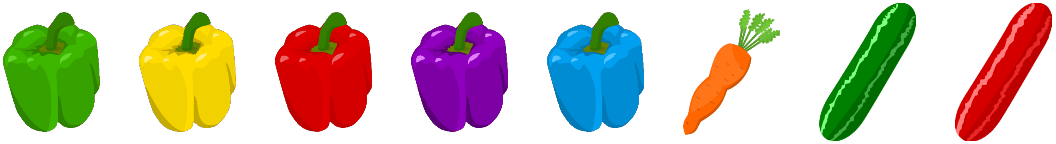
Truth-Conditional Semantics



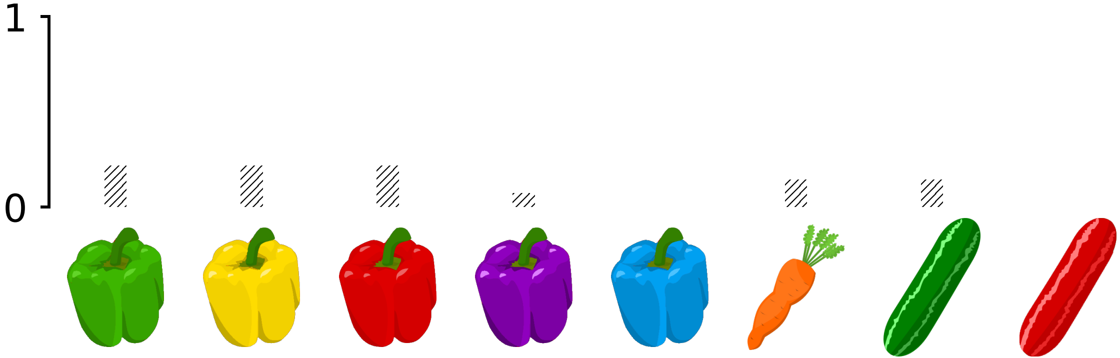
Truth-Conditional Semantics



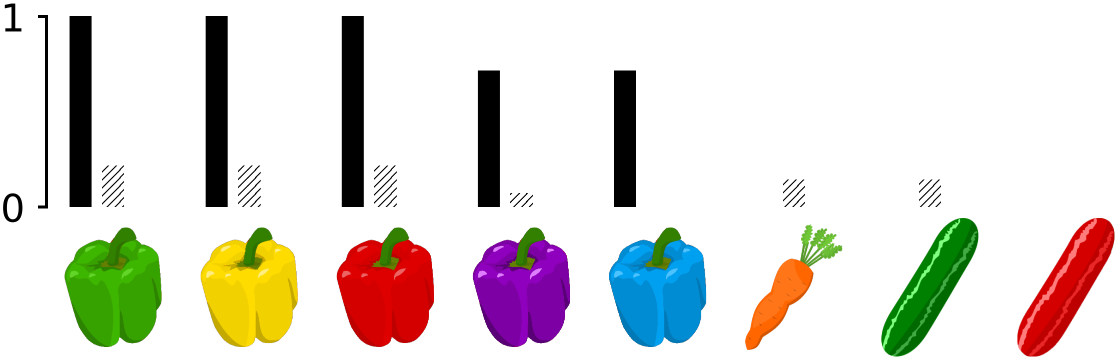
Truth-Conditional Functions



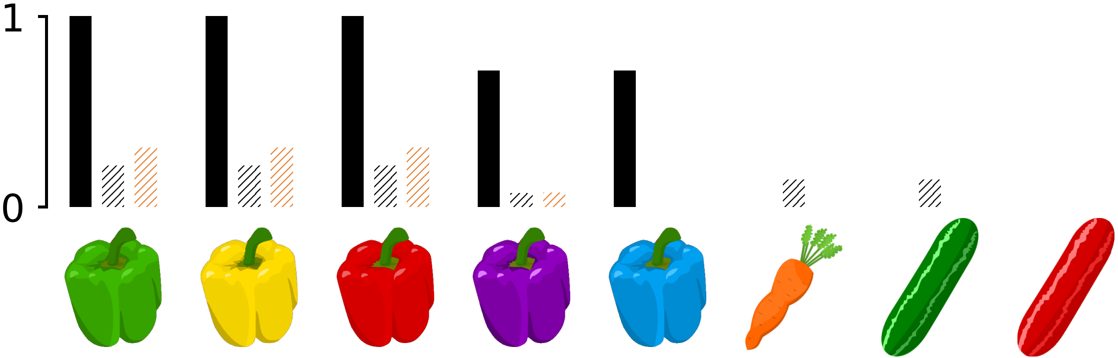
Truth-Conditional Functions



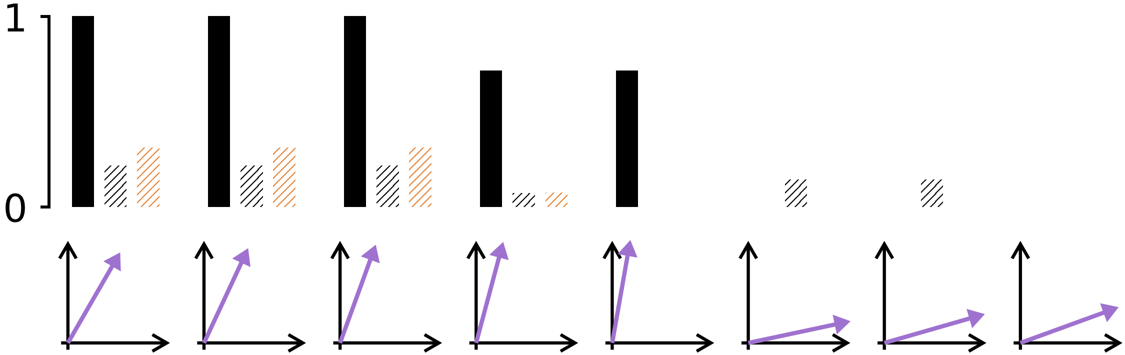
Truth-Conditional Functions



Truth-Conditional Functions



Truth-Conditional Functions



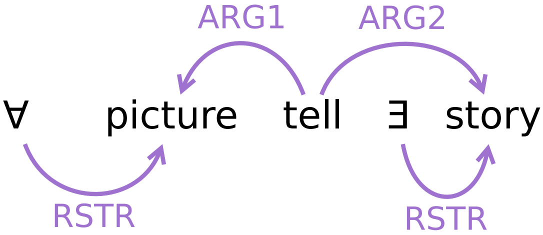
Summary of What’s New

- Pixie: entity representation
- Word meanings as functions:
  pixie $\mapsto$ probability of truth

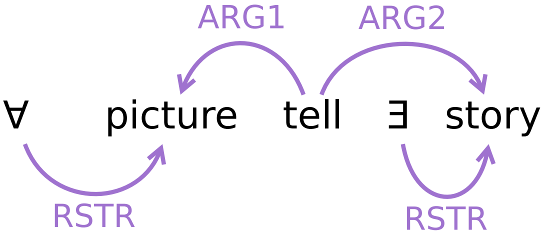
DMRS

Every picture tells a story

DMRS

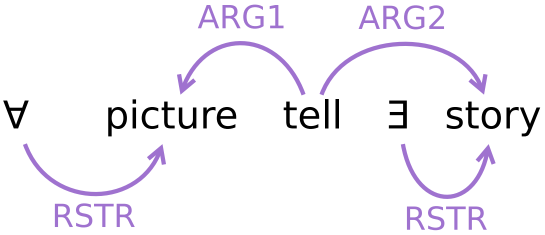


DMRS



$$\forall x \exists y \exists z \ \text{picture}(x) \Rightarrow \big[\, \text{story}(z) \wedge \text{tell}(y) \wedge \text{ARG1}(y,x) \wedge \text{ARG2}(y,z) \,\big]$$

DMRS



$$\forall x \exists y \exists z \; \text{picture}(x) \Rightarrow \left[\text{story}(z) \wedge \text{tell}(y) \wedge \text{ARG1}(y, x) \wedge \text{ARG2}(y, z) \right]$$

- See: “Linguists Who Use Probabilistic Models Love Them: Quantification in Functional Distributional Semantics” (PaM2020)

Functional Distributional Semantics

$$\text{dog} \xleftarrow{\text{ARG1}} \text{chase} \xrightarrow{\text{ARG2}} \text{cat}$$

Functional Distributional Semantics

$$x \xleftarrow{\text{ARG1}} y \xrightarrow{\text{ARG2}} z$$

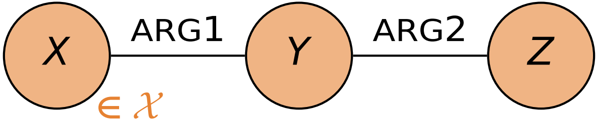
$$\text{dog}(x) \qquad \text{chase}(y) \qquad \text{cat}(z)$$

Functional Distributional Semantics

$$x \xleftarrow{\text{ARG1}} y \xrightarrow{\text{ARG2}} z$$

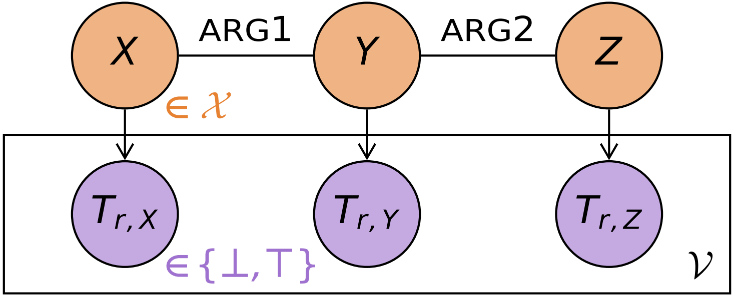
dog(x)	chase(y)	cat(z)
animal(x)	pursue(y)	animal(z)
chase(x)	dog(y)	chase(z)
pursue(x)	cat(y)	pursue(z)
cat(x)	animal(y)	dog(z)

Functional Distributional Semantics

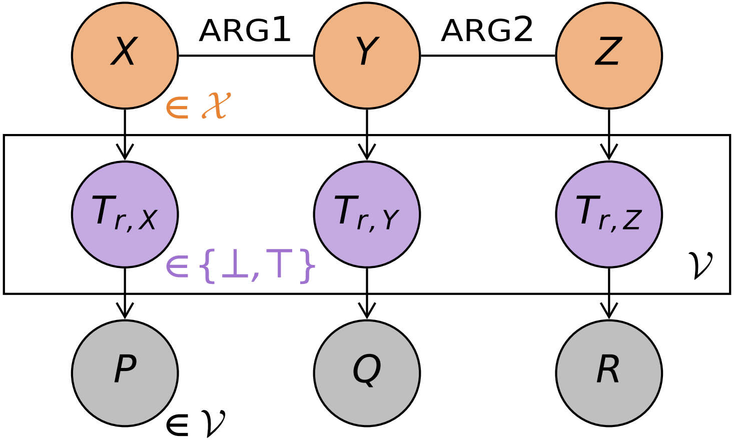


dog(X)	chase(Y)	cat(Z)
animal(X)	pursue(Y)	animal(Z)
chase(X)	dog(Y)	chase(Z)
pursue(X)	cat(Y)	pursue(Z)
cat(X)	animal(Y)	dog(Z)

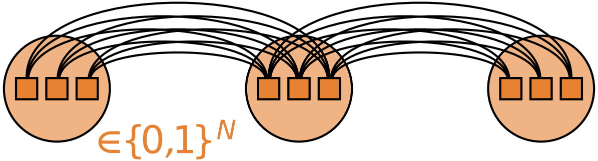
Functional Distributional Semantics



Functional Distributional Semantics

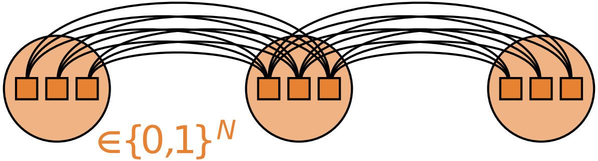


World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

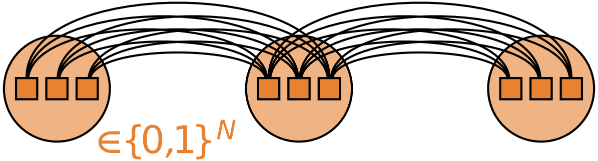
World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

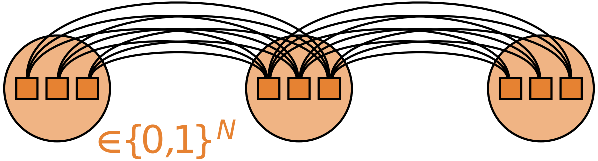
- (Work in progress: real-valued version)

World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

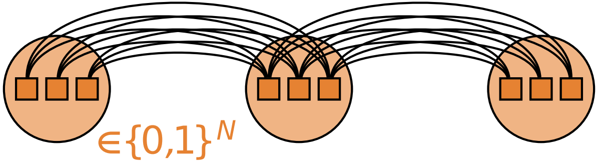
World Model



- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)

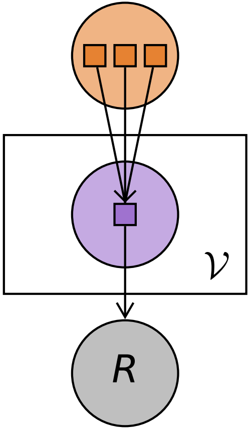
- $\mathbb{P}(s) \propto \exp\left(-E(s)\right)$

World Model



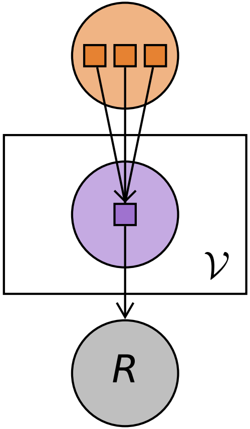
- Cardinality Restricted Boltzmann Machine
  (CaRBM; Swersky et al., 2012)
- $\mathbb{P}(s) \propto \exp\left(\sum_{x \xrightarrow{L} y \text{ in } s} w_{ij}^{(L)} x_i y_j \right)$

Lexical Model



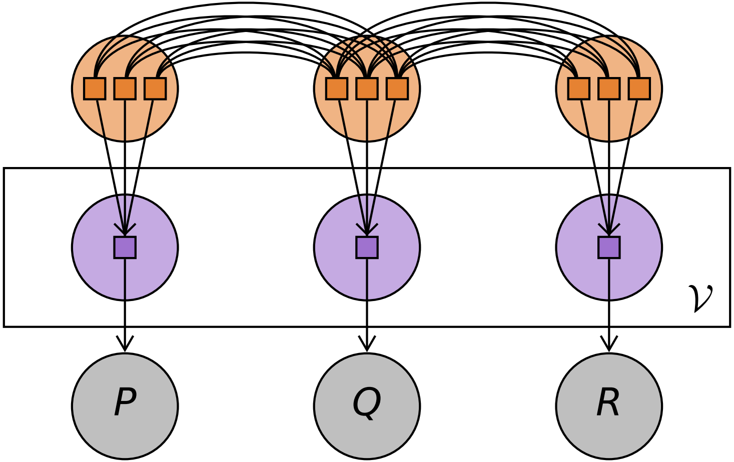
- Feedforward networks
- $t^{(r)}(x) = \sigma\left(v_i^{(r)} x_i\right)$

Lexical Model



- Feedforward networks
- $t^{(r)}(x) = \sigma\left(v_i^{(r)} x_i\right)$
- $\mathbb{P}(r \mid x) \propto t^{(r)}(x)$

Functional Distributional Semantics



Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial\theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial\theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

Gradient Descent

$$\frac{\partial}{\partial \theta} \log \mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s \right) \left[\frac{\partial}{\partial \theta} \left(-E(s) \right) \right] + \mathbb{E}_{s|g} \left[\frac{\partial}{\partial \theta} \log \mathbb{P}(g \mid s) \right]$$

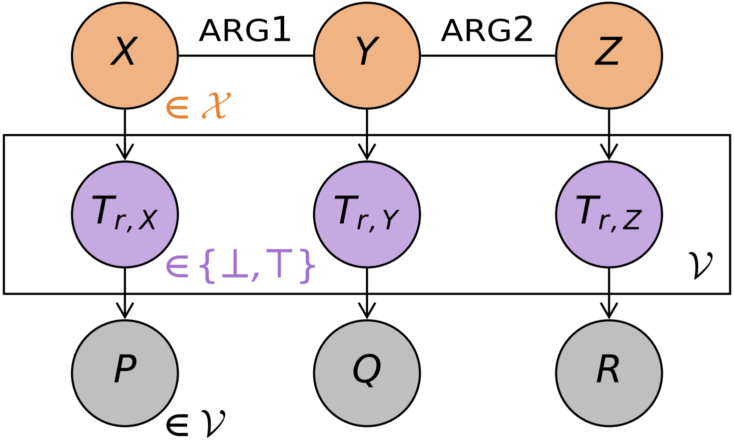
- Latent variables necessary but inconvenient

Gradient Descent

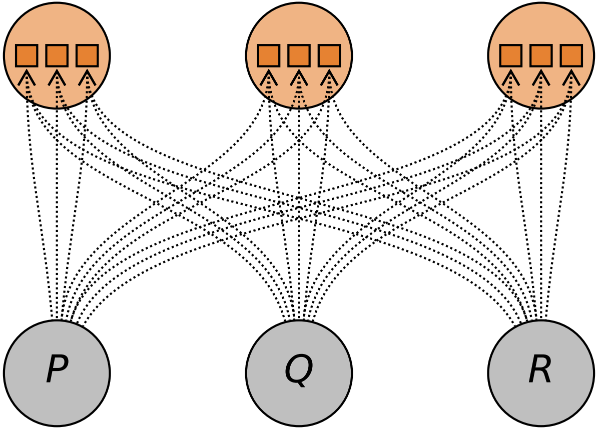
$$\frac{\partial}{\partial \theta} \log \mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial \theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial \theta} \log \mathbb{P}(g \mid s)\right]$$

- Latent variables necessary but inconvenient
- Approximate distribution: variational inference
  (Jordan et al., 1999; Attias, 2000)

Functional Distributional Semantics



Variational Inference



Amortised Variational Inference

- Variational distribution must be optimised *for each input graph*

Amortised Variational Inference

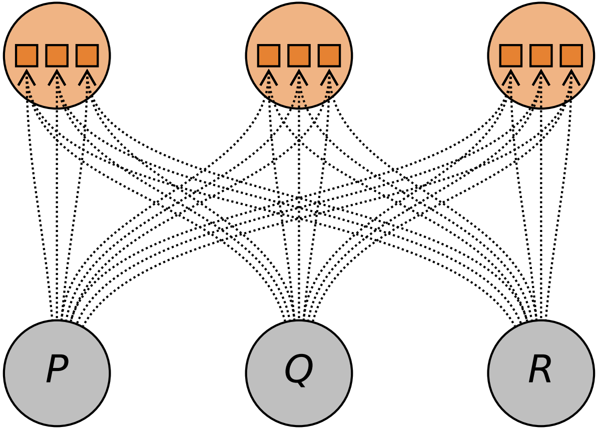
- Variational distribution must be optimised *for each input graph*

- Amortisation: train a network to predict the variational distribution (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014)

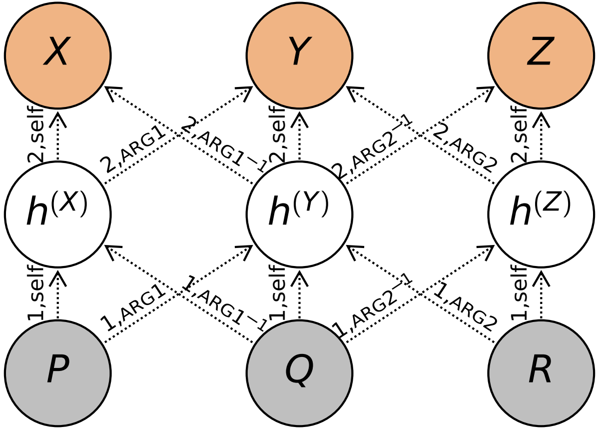
Amortised Variational Inference

- Variational distribution must be optimised *for each input graph*
- Amortisation: train a network to predict the variational distribution (Kingma and Welling, 2014; Rezende et al., 2014; Mnih and Gregor, 2014)
- Input graphs of different topologies: share network weights with graph convolutions (Duvenaud et al., 2015; Marcheggiani and Titov, 2017)

Variational Inference



Amortised Variational Inference



Amortised Variational Inference

$$
\begin{aligned}
\frac{\partial}{\partial \phi} D(\mathbb{Q} | \mathbb{P}) = &- \frac{\partial}{\partial \phi} \mathbb{E}_{\mathbb{Q}(s)} \left[\log \mathbb{P}(s) \right] \\
&- \frac{\partial}{\partial \phi} \mathbb{E}_{\mathbb{Q}(s)} \left[\log \mathbb{P}(g \mid s) \right] \\
&- \frac{\partial}{\partial \phi} H(\mathbb{Q})
\end{aligned}
$$

Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial\theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

- Latent variables: amortised variational inference

Gradient Descent

$$\frac{\partial}{\partial\theta}\log\mathbb{P}(g) = \left(\mathbb{E}_{s|g} - \mathbb{E}_s\right)\left[\frac{\partial}{\partial\theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial\theta}\log\mathbb{P}(g\,|\,s)\right]$$

- Latent variables: amortised variational inference

- Additional details... regularisation, dropout, β-VAE weighting, negative sampling, probit approximation, learning rate, warm start, soft constraints, belief propagation for $\mathbb{E}_s$...

Gradient Descent

$$\frac{\partial}{\partial \theta} \log \mathbb{P}\left(g\right) = \left(\mathbb{E}_{s|g} - \mathbb{E}_{s}\right)\left[\frac{\partial}{\partial \theta}\left(-E(s)\right)\right] + \mathbb{E}_{s|g}\left[\frac{\partial}{\partial \theta} \log \mathbb{P}\left(g \,|\, s\right)\right]$$

- Latent variables: amortised variational inference
- See: "Autoencoding Pixies: Amortised Variational Inference with Graph Convolutions for Functional Distributional Semantics" (ACL 2020)

Pixie Autoencoder

- Generative model & inference network

Pixie Autoencoder

- Generative model & inference network
- Unique selling point:
    - Truth-conditional distributional semantics

Pixie Autoencoder

- Generative model & inference network
- Unique selling point:
  - Truth-conditional distributional semantics
- `https://gitlab.com/guyemerson/pixie/`

Training Needs Graphs

- Training needs dependency graphs, not raw text

Training Needs Graphs

- Training needs dependency graphs, not raw text
- WikiWoods
  - English Wikipedia, parsed into DMRS graphs
  - 31 million graphs (after preprocessing)

Training Needs Graphs

- Training needs dependency graphs, not raw text
- WikiWoods
  - English Wikipedia, parsed into DMRS graphs
  - 31 million graphs (after preprocessing)
  - (So far, only verbs with ARG1 & ARG2 nouns)

Ongoing/Future Work

- Joint learning with grounded data
- Joint learning with lexical resources
- More efficient model (continuous pixies)
- Latent variable for “topic”
- Covariance of truth values (for pragmatics)
- Deeper networks (for polysemy)
- Semi-compositional idioms

Ongoing/Future Work

- Joint learning with grounded data
- Joint learning with lexical resources
- More efficient model (continuous pixies)
- Latent variable for “topic”
- Covariance of truth values (for pragmatics)
- Deeper networks (for polysemy)
- Semi-compositional idioms

Ongoing/Future Work

- Joint learning with grounded data
- Joint learning with lexical resources
- More efficient model (continuous pixies)
- Latent variable for “topic”
- Covariance of truth values (for pragmatics)
- Deeper networks (for polysemy)
- Semi-compositional idioms

Joint Learning with Grounded Data

- Fundamental distinction between words and entities

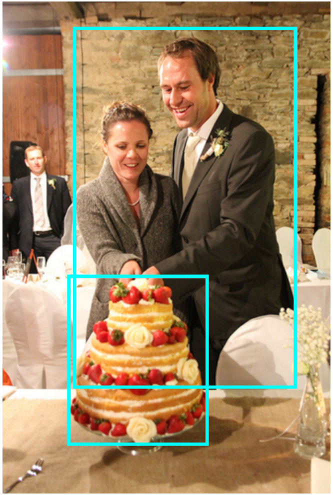
Joint Learning with Grounded Data

- Fundamental distinction between words and entities
- Vector space models:
  - Early fusion, late fusion, cross-modal maps...

Joint Learning with Grounded Data

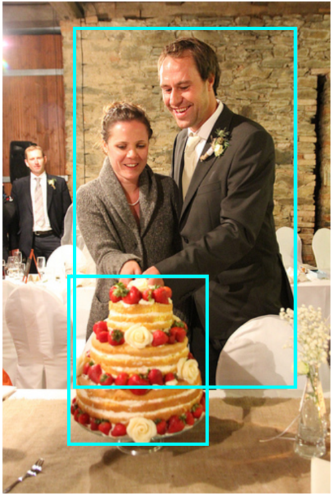
- Fundamental distinction between words and entities
- Vector space models:
  - Early fusion, late fusion, cross-modal maps...
- Functional Distributional Semantics:
  - Text → pixies are latent
  - Grounded data → pixies are observed

Visual Genome Dataset



"couple cutting cake"

Visual Genome Dataset



"couple cutting cake"