Deep latent models of word representation

Wilker Aziz
University of Amsterdam

April 3, 2018

Word representation

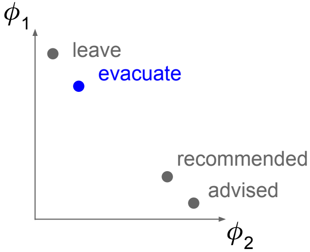
An abstraction that stands for the use of a word

In the event of a chemical spill, most children know they should evacuate as advised by people in charge.

Word representation

An abstraction that stands for the use of a word

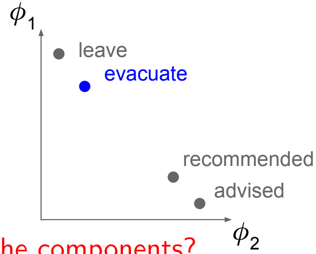
In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*



Word representation

An abstraction that stands for the use of a word

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*



How do we choose the components?

Distributional Hypothesis

Context can represent the intended use of a word

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

- success hinges on **discriminative power** of available context

Discriminative embedding models

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Discriminative embedding models

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Fit a binary classifier (Goldberg and Levy, 2014)

- positive examples
- negative examples

Discriminative embedding models

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

“Have I seen this word in this context?”

Fit a binary classifier (Goldberg and Levy, 2014)

- positive examples
- negative examples

“Far away” context expresses a form of negative correlation

Discriminative embedding models

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Fit a binary classifier (Goldberg and Levy, 2014)

- positive examples
- negative examples

"Far away" context expresses a form of negative correlation

- But data only show positive examples

Discriminative embedding models

In the event of a chemical spill, most children know they should **evacuate** *as advised by people in charge.*

ambiguity

Place words in $\mathbb{R}^d$ as to answer questions like

"Have I seen this word in this context?"

Fit a binary classifier (Goldberg and Levy, 2014)

- positive examples
- negative examples

"Far away" context expresses a form of negative correlation

- But data only show positive examples

Limitations

Meaning representation is an **unsupervised** problem

- we cannot actually observe what we are trying to learn i.e. representations

Limitations

Meaning representation is an **unsupervised** problem

- we cannot actually observe what we are trying to learn i.e. representations

Distributional hypothesis seems pretty strong

- but it fails when context is not sufficiently discriminative

EmbedAlign

Generative treatment

- model what we want to induce (i.e. representations)
- learn from **positive examples**
- learn from richer (less ambiguous) context

Rios et al. (2018)

Outline

1. Embed-Align
2. Bayesian Skip-Gram
3. Practical

Equivalence through translation

In the event of a chemical spill, most children know they should **evacuate** as advised by people in charge.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**.

Equivalence through translation

In the event of a chemical spill, most children know they should **evacuate** as advised by people in charge.

Em caso de vazamento químico, a maioria das crianças estão conscientes de que devem **deixar** o local, como sugerem autoridades.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**.

Após ingerida, a substância acelera o movimento das paredes do intestino forçando o indivíduo a se **aliviar**.

Equivalence through translation

In the event of a chemical spill, most children know they should **evacuate** as advised by people in charge.

Em caso de vazamento químico, a maioria das crianças estão conscientes de que devem **deixar** o local, como sugerem autoridades.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**.

Após ingerida, a substância acelera o movimento das paredes do intestino forçando o indivíduo a se **aliviar**.

Observation from WSD community

- foreign text as **proxy to sense supervision** (Diab and Resnik, 2002)

Embed-Align

quickly evacuate the area / deixe o local rapidamente

Embed-Align

quickly evacuate the area / deixe o local rapidamente

Z	z_1	z_2	z_3	z_4

Embed-Align

quickly evacuate the area / deixe o local rapidamente

X	quickly$_1$	evacuate$_2$	the$_3$	area$_4$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
Z	z_1	z_2	z_3	z_4

Embed-Align

quickly evacuate the area / deixe o local rapidamente

X	quickly$_1$	evacuate$_2$	the$_3$	area$_4$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
Z	z_1	z_2	z_3	z_4
A	$a_1 = 2$	$a_2 = 3$	$a_3 = 4$	$a_4 = 1$

Embed-Align

quickly evacuate the area / deixe o local rapidamente

X	quickly$_1$	evacuate$_2$	the$_3$	area$_4$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
Z	z_1	z_2	z_3	z_4
A	$a_1 = 2$	$a_2 = 3$	$a_3 = 4$	$a_4 = 1$
Z_a	z_2	z_3	z_4	z_1

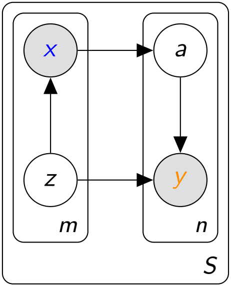
Embed-Align

quickly evacuate the area / deixe o local rapidamente

$$
\begin{array}{c|cccc}
X & \text{quickly}_1 & \text{evacuate}_2 & \text{the}_3 & \text{area}_4 \\
 & \uparrow & \uparrow & \uparrow & \uparrow \\
Z & z_1 & z_2 & z_3 & z_4 \\
 & & & & \\
A & a_1 = 2 & a_2 = 3 & a_3 = 4 & a_4 = 1 \\
Z_a & z_2 & z_3 & z_4 & z_1 \\
 & \downarrow & \downarrow & \downarrow & \downarrow \\
Y & \text{deixe}_1 & \text{o}_2 & \text{local}_3 & \text{rapidamente}_4
\end{array}
$$

Embed-Align

quickly evacuate the area / deixe o local rapidamente



X	quickly$_1$	evacuate$_2$	the$_3$	area$_4$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
Z	z_1	z_2	z_3	z_4
A	$a_1 = 2$	$a_2 = 3$	$a_3 = 4$	$a_4 = 1$
Z_a	z_2	z_3	z_4	z_1
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
Y	deixe$_1$	o$_2$	local$_3$	rapidamente$_4$

Marginalising alignments collects additional training data for z

How does it disambiguate?

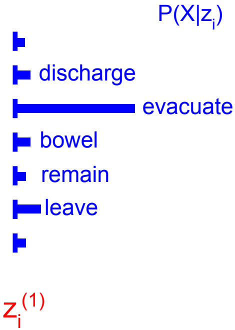
In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

How does it disambiguate?

$z_i^{(1)}$

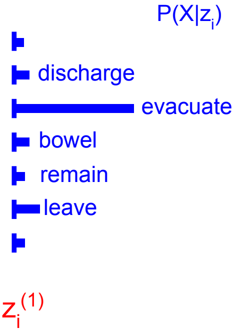
In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

How does it disambiguate?



In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

How does it disambiguate?



In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**$_i$.

How does it disambiguate?

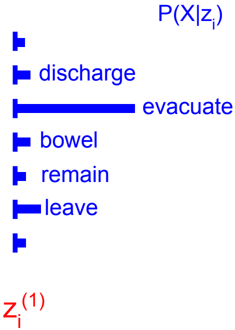
$z_i^{(1)}$

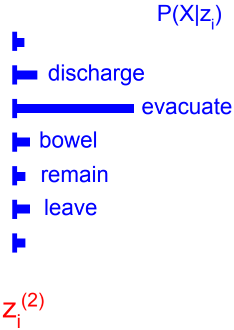
In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

$z_i^{(2)}$

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**$_i$.

How does it disambiguate?

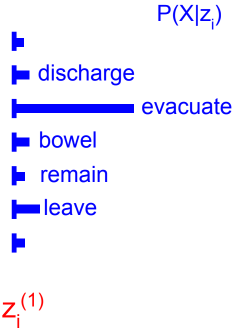


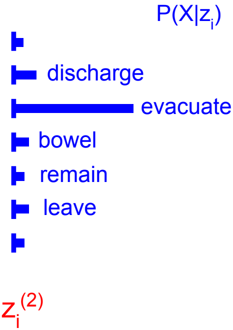


In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**$_i$.

How does it disambiguate?





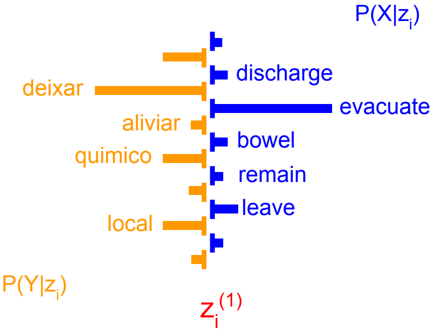
In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

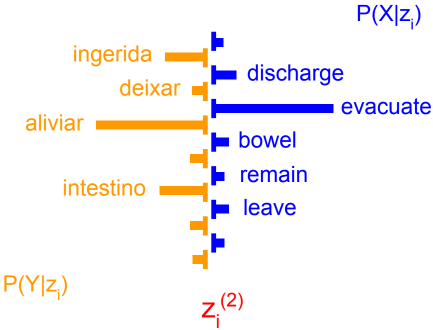
Em caso de vazamento químico, a maioria das crianças estão conscientes de que devem **deixar** o local, como sugerem autoridades.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**$_i$.

Após ingerida, a substância acelera o movimento das paredes do intestino forçando o indivíduo a se **aliviar**.

How does it disambiguate?





In the event of a chemical spill, most children know they should **evacuate**$_i$ as advised by people in charge.

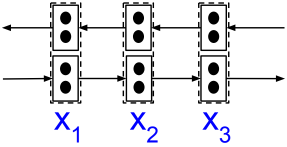
Em caso de vazamento químico, a maioria das crianças estão conscientes de que devem **deixar** o local, como sugerem autoridades.

After ingestion, the substance speeds up the movement of your bowels encouraging you to **evacuate**$_i$.

Após ingerida, a substância acelera o movimento das paredes do intestino forçando o indivíduo a se **aliviar**.

Tractable inference

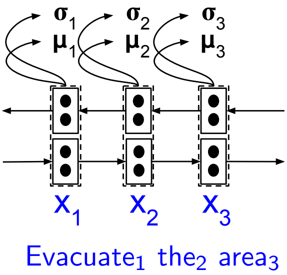
1. Read sentence

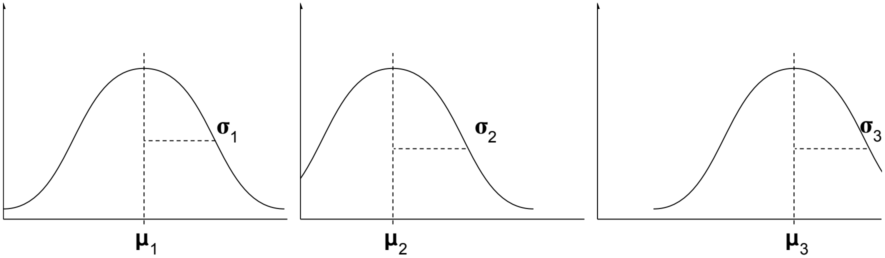


Evacuate$_1$ the$_2$ area$_3$

Tractable inference

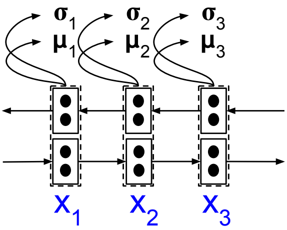
1. Read sentence
2. Predict posterior mean μ_i and std σ_i



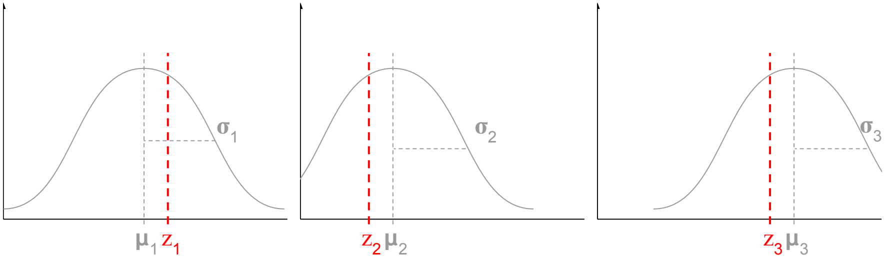


Tractable inference

1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$

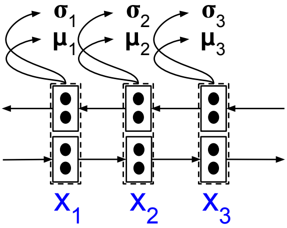


Evacuate$_1$ the$_2$ area$_3$

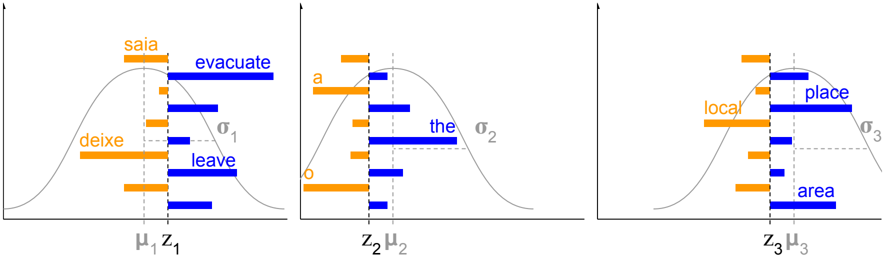


Tractable inference

1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
4. Predict categorical distributions

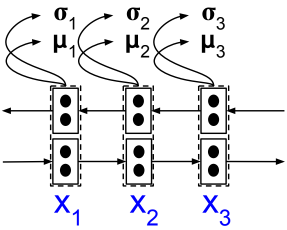


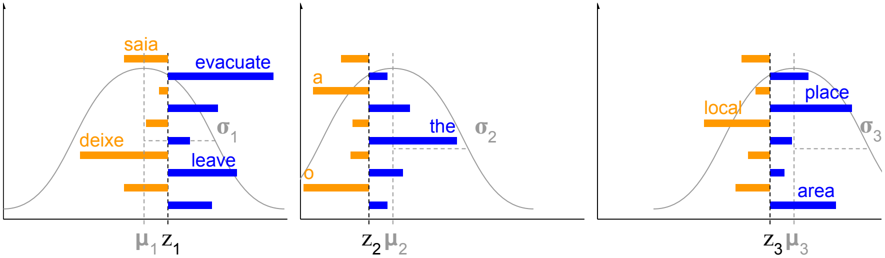
Evacuate$_1$ the$_2$ area$_3$



Tractable inference

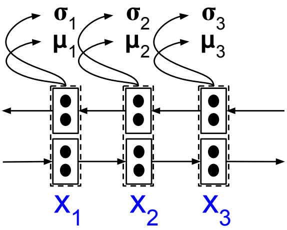
1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
4. Predict categorical distributions
5. Generate observations
   Evacuate$_1$ the$_2$ area$_3$ / Deixe$_1$ o$_2$ local$_3$

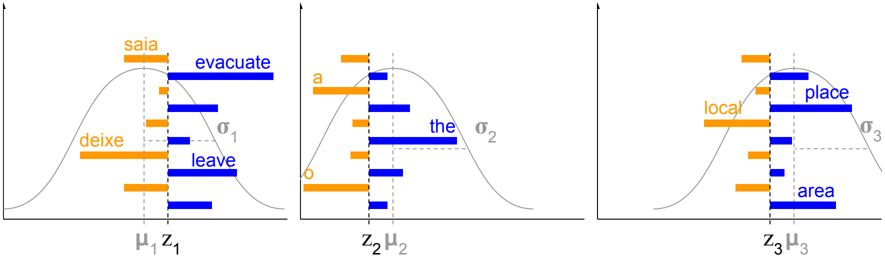




Tractable inference

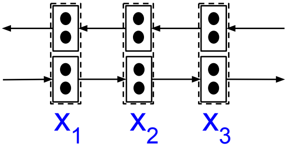
1. Read sentence
2. Predict posterior mean μ_i and std σ_i
3. Sample $Z_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
4. Predict categorical distributions
5. Generate observations
   Evacuate$_1$ the$_2$ area$_3$ / Deixe$_1$ o$_2$ local$_3$
6. Maximise a lowerbound on likelihood
   (Kingma and Welling, 2014)





What's special about it?

The model reads English text and

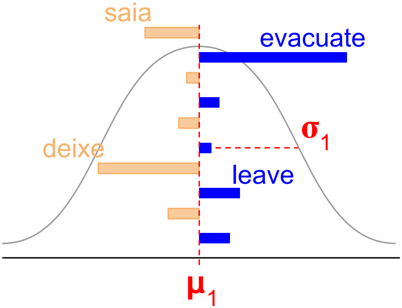


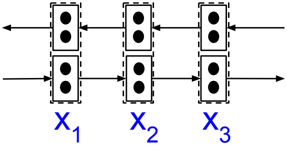
Evacuate$_1$ the$_2$ area$_3$

What's special about it?

The model reads English text and

- predicts uncertainty
- describes "sense" using Portuguese words





Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$Z_i \sim \mathcal{N}(0, I)$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$
\begin{aligned}
Z_i &\sim \mathcal{N}(0, I) \\
X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i))
\end{aligned}
$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, I) \\ X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\ \mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\ A_j | m &\sim \mathcal{U}(1/m) \end{aligned}$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$
\begin{aligned}
Z_i &\sim \mathcal{N}(0, I) \\
X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\
A_j | m &\sim \mathcal{U}(1/m) \\
Y_j | z_1^m, a_j &\sim \mathrm{Cat}(\mathbf{g}_{a_j}) \\
\mathbf{g}_{a_j} &= \mathrm{softmax}(\mathrm{affine}_\theta(z_{a_j}))
\end{aligned}
$$

Inference model: for $i = 1, \ldots, m$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$
\begin{aligned}
Z_i &\sim \mathcal{N}(0, I) \\
X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\
A_j | m &\sim \mathcal{U}(1/m) \\
Y_j | z_1^m, a_j &\sim \mathrm{Cat}(\mathbf{g}_{a_j}) \\
\mathbf{g}_{a_j} &= \mathrm{softmax}(\mathrm{affine}_\theta(z_{a_j}))
\end{aligned}
$$

Inference model: for $i = 1, \ldots, m$

$$
Z_i | x_1^m \sim \mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i))
$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$\begin{aligned}
Z_i &\sim \mathcal{N}(0, I) \\
X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\
A_j | m &\sim \mathcal{U}(1/m) \\
Y_j | z_1^m, a_j &\sim \mathrm{Cat}(\mathbf{g}_{a_j}) \\
\mathbf{g}_{a_j} &= \mathrm{softmax}(\mathrm{affine}_\theta(z_{a_j}))
\end{aligned}$$

Inference model: for $i = 1, \ldots, m$

$$\begin{aligned}
Z_i | x_1^m &\sim \mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i)) \\
\mathbf{h}_1^m &= \mathrm{enc}_\lambda(x_1^m)
\end{aligned}$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$
\begin{aligned}
Z_i &\sim \mathcal{N}(0, I) \\
X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\
A_j | m &\sim \mathcal{U}(1/m) \\
Y_j | z_1^m, a_j &\sim \mathrm{Cat}(\mathbf{g}_{a_j}) \\
\mathbf{g}_{a_j} &= \mathrm{softmax}(\mathrm{affine}_\theta(z_{a_j}))
\end{aligned}
$$

Inference model: for $i = 1, \ldots, m$

$$
\begin{aligned}
Z_i | x_1^m &\sim \mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i)) \\
\mathbf{h}_1^m &= \mathrm{enc}_\lambda(x_1^m) \\
\mathbf{u}_i &= \mathrm{affine}_\lambda(\mathrm{relu}(\mathrm{affine}_\lambda(\mathbf{h}_i)))
\end{aligned}
$$

Rios et al. (2018)

Complete specification

Generative model: for $i = 1, \ldots, m$ and $j = 1, \ldots, n$

$$\begin{aligned} Z_i &\sim \mathcal{N}(0, I) \\ X_i | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\ \mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i)) \\ A_j | m &\sim \mathcal{U}(1/m) \\ Y_j | z_1^m, a_j &\sim \mathrm{Cat}(\mathbf{g}_{a_j}) \\ \mathbf{g}_{a_j} &= \mathrm{softmax}(\mathrm{affine}_\theta(z_{a_j})) \end{aligned}$$

Inference model: for $i = 1, \ldots, m$

$$\begin{aligned} Z_i | x_1^m &\sim \mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i)) \\ \mathbf{h}_1^m &= \mathrm{enc}_\lambda(x_1^m) \\ \mathbf{u}_i &= \mathrm{affine}_\lambda(\mathrm{relu}(\mathrm{affine}_\lambda(\mathbf{h}_i))) \\ \mathbf{s}_i &= \mathrm{softplus}_\lambda(\mathrm{relu}(\mathrm{affine}_\lambda(\mathbf{h}_i))) \end{aligned}$$

Rios et al. (2018)

ELBO

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p(z_1^m)\right)$$

ELBO

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z_1^m)\right)$$

KL term

$$\mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z)\right) = \underbrace{\sum_{i=1}^{m} \mathrm{KL}\left(q_\lambda(z_i|x_1^m) \mid\mid p(z_1^m)\right)}_{\text{mean field}}$$

ELBO

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p(z_1^m)\right)$$

KL term

$$\begin{aligned}\mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p(z)\right) &= \underbrace{\sum_{i=1}^{m} \mathrm{KL}\left(q_\lambda(z_i|x_1^m) \,||\, p(z_1^m)\right)}_{\text{mean field}} \\ &= \sum_{i=1}^{m} \mathrm{KL}\left(\underbrace{\mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i))}_{\text{inference model}} \,||\, \underbrace{\mathcal{N}(0, I)}_{\text{prior}}\right)\end{aligned}$$

ELBO - L1 term

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z_1^m)\right)$$

ELBO - L1 term

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p(z_1^m)\right)$$

Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right]$$

ELBO - L1 term

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z_1^m)\right)$$

Likelihood term

$$\begin{aligned}&\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right]\\ &= \underbrace{\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m|z_1^m)\right] + \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right]}_{\text{conditional independence}}\end{aligned}$$

ELBO - L1 term

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z_1^m)\right)$$

Likelihood term

$$\begin{aligned}&\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] \\ &= \underbrace{\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m|z_1^m)\right] + \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right]}_{\text{conditional independence}}\end{aligned}$$

L1 term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m|z_1^m)\right] = \sum_{i=1}^{m} \mathbb{E}_{q_\lambda(z_i|x_1^m)}\left[\log P_\theta(x_i|z_i)\right]$$

ELBO - L1 term

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p(z_1^m)\right)$$

Likelihood term

$$\begin{aligned}&\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m, y_1^n|z_1^m)\right]\\ &= \underbrace{\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m|z_1^m)\right] + \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right]}_{\text{conditional independence}}\end{aligned}$$

L1 term

$$\begin{aligned}\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(x_1^m|z_1^m)\right] &= \sum_{i=1}^{m} \mathbb{E}_{q_\lambda(z_i|x_1^m)}\left[\log P_\theta(x_i|z_i)\right]\\ &= \sum_{i=1}^{m} \mathbb{E}_{q_\lambda(z_i|x_1^m)}\left[\log \mathrm{Cat}(x_i|\mathbf{f}_i)\right]\end{aligned}$$

ELBO - L2 term

L2 term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right]$$

ELBO - L2 term

L2 term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] = \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{j=1}^{n} P_\theta(y_j|m, z_1^m)\right]$$

ELBO - L2 term

L2 term

$$\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{j=1}^{n} P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log P_\theta(y_j|m, z_1^m)\right]
\end{aligned}$$

ELBO - L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{j=1}^{n} P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P_\theta(y_j, a_j|m, z_1^m)\right]
\end{aligned}
$$

ELBO - L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{j=1}^{n} P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P_\theta(y_j, a_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right]
\end{aligned}
$$

ELBO - L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{j=1}^{n} P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log P_\theta(y_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P_\theta(y_j, a_j|m, z_1^m)\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} \mathcal{U}(a_j|1/m)\, \mathrm{Cat}(y_j|\mathbf{g}_{a_j})\right]
\end{aligned}
$$

ELBO - Lowerbound on L2 term

L2 term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] = \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right]$$

You may skip over this slide.

ELBO - Lowerbound on L2 term

L2 term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] = \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right]$$

$$= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \mathbb{E}_{P(a_j|m)}\left[P_\theta(y_j|z_{a_j})\right]\right]$$

You may skip over this slide.

ELBO - Lowerbound on L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j} P(a_j|m) P_\theta(y_j|z_{a_j})\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \mathbb{E}_{P(a_j|m)}\left[P_\theta(y_j|z_{a_j})\right]\right] \\
&\overset{\text{JI}}{\geq} \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \mathbb{E}_{P(a_j|m)}\left[\log P_\theta(y_j|z_{a_j})\right]\right]
\end{aligned}
$$

You may skip over this slide.

ELBO - Lowerbound on L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \mathbb{E}_{P(a_j|m)}\left[P_\theta(y_j|z_{a_j})\right]\right] \\
&\overset{\text{JI}}{\geq} \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \mathbb{E}_{P(a_j|m)}\left[\log P_\theta(y_j|z_{a_j})\right]\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \mathbb{E}_{\mathcal{U}(a_j|m)}\left[\log \mathrm{Cat}(y_j|\mathbf{g}_{a_j})\right]\right]
\end{aligned}
$$

You may skip over this slide.

ELBO - Lowerbound on L2 term

L2 term

$$
\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log P_\theta(y_1^n|m, z_1^m)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \sum_{a_j}^{m} P(a_j|m) P_\theta(y_j|z_{a_j})\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \log \mathbb{E}_{P(a_j|m)}\left[P_\theta(y_j|z_{a_j})\right]\right] \\
&\overset{\text{JI}}{\geq} \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \mathbb{E}_{P(a_j|m)}[\log P_\theta(y_j|z_{a_j})]\right] \\
&= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{j=1}^{n} \mathbb{E}_{\mathcal{U}(a_j|m)}[\log \mathrm{Cat}(y_j|\mathbf{g}_{a_j})]\right]
\end{aligned}
$$

Alignment model $P(a_j|m) = \frac{1}{m}$ is not a function of θ

- we can make an MC estimate by sampling candidate alignments uniformly

---

You may skip over this slide.

The softmax problem

Categorical parameters are expensive to compute

$$
\begin{aligned}
X &\sim \mathrm{Cat}(\mathbf{f}) \\
\mathbf{f} &= \mathrm{softmax}(\hat{\mathbf{f}}) \\
f_x &= \frac{\exp\left(\hat{f}_x\right)}{\sum_{x' \in \mathcal{X}} \exp\left(\hat{f}_{x'}\right)}
\end{aligned}
$$

$v_1 = |\mathcal{X}|$ is the size of the vocabulary of L_1; $v_2 = |\mathcal{Y}|$ is the size of the vocabulary of L_2

The softmax problem

Categorical parameters are expensive to compute

$$X \sim \mathrm{Cat}(\mathbf{f})$$

$$\mathbf{f} = \mathrm{softmax}(\hat{\mathbf{f}})$$

$$f_x = \frac{\exp\left(\hat{f}_x\right)}{\sum_{x' \in \mathcal{X}} \exp\left(\hat{f}_{x'}\right)}$$

- $\mathbf{f}_1^m$ requires normalising m distributions over the vocabulary of L_1, thus it takes time $O(m \times v_1)$
- $\mathbf{g}_1^m$ requires normalising m distributions over the vocabulary of L_2, thus it takes time $O(m \times v_2)$

$v_1 = |\mathcal{X}|$ is the size of the vocabulary of L_1; $v_2 = |\mathcal{Y}|$ is the size of the vocabulary of L_2

Efficient softmax

Logistic regression

$$P(X = x|z) = \frac{\exp(s(z, x))}{\sum_{x' \in \mathcal{X}} \exp(s(z, x'))}$$

Efficient softmax

Logistic regression

$$P(X = x|z) = \frac{\exp(s(z, x))}{\sum_{x' \in \mathcal{X}} \exp(s(z, x'))}$$

Define

- a set $\mathcal{C}(x)$ such that $x \in \mathcal{C}$
- a set $\mathcal{N}(x)$ such that $\mathcal{C}(x) \cap \mathcal{N}(x) = \emptyset$

Efficient softmax

Logistic regression

$$P(X = x|z) = \frac{\exp(s(z, x))}{\sum_{x' \in \mathcal{X}} \exp(s(z, x'))}$$

Define

- a set $\mathcal{C}(x)$ such that $x \in \mathcal{C}$
- a set $\mathcal{N}(x)$ such that $\mathcal{C}(x) \cap \mathcal{N}(x) = \emptyset$

Re-express normaliser for $P(X = x|z)$

$$\sum_{x' \in \mathcal{X}} \exp(s(z, x')) = \sum_{x' \in \mathcal{C}(x)} \exp(s(z, x')) + \sum_{x' \in \mathcal{N}(x)} \kappa(x') \exp(s(z, x'))$$

- $\kappa(x') = \frac{1}{q(x')}$ and $q(x')$ is an importance distribution

Approximate $P_{X|Z}$

Logistic regression

$$P_\theta(x|z) = \frac{\exp(s_\theta(z,x))}{\sum_{x' \in \mathcal{X}} \exp(s_\theta(z,x'))}$$

Approximate $P_{X|Z}$

Logistic regression

$$P_\theta(x|z) = \frac{\exp(s_\theta(z, x))}{\sum_{x' \in \mathcal{X}} \exp(s_\theta(z, x'))}$$

Build

- a set $\mathcal{C}$ containing all L_1 words in batch
- a set $\mathcal{N}$ sampling uniformly without replacement from $\mathcal{X} \setminus \mathcal{C}$

Approximate $P_{X|Z}$

Logistic regression

$$P_\theta(x|z) = \frac{\exp(s_\theta(z,x))}{\sum_{x' \in \mathcal{X}} \exp(s_\theta(z,x'))}$$

Build

- a set $\mathcal{C}$ containing all L_1 words in batch
- a set $\mathcal{N}$ sampling uniformly without replacement from $\mathcal{X} \setminus \mathcal{C}$

Approximate normaliser for $P_\theta(x|z)$

$$\sum_{x' \in \mathcal{X}} \exp(s_\theta(z,x')) \approx \sum_{x' \in \mathcal{C}} \exp(s_\theta(z,x')) + \sum_{x' \in \mathcal{N}} \frac{|\mathcal{X} \setminus \mathcal{C}|}{|\mathcal{N}|} \exp(s_\theta(z,x'))$$

- $s_\theta(z,x) = z^\top \mathbf{c}_x + b_x$
  $\mathbf{c}_x = \mathrm{lookup}_\theta(x)$ is a deterministic embedding, b_x a bias term

Approximate $P_{Y|Z}$

Logistic regression

$$P_\theta(y|z) = \frac{\exp(u_\theta(z, y))}{\sum_{y' \in \mathcal{Y}} \exp(u_\theta(z, y'))}$$

Approximate $P_{Y|Z}$

Logistic regression

$$P_\theta(y|z) = \frac{\exp(u_\theta(z, y))}{\sum_{y' \in \mathcal{Y}} \exp(u_\theta(z, y'))}$$

Build

- a set $\mathcal{C}$ containing all L_2 words in batch
- a set $\mathcal{N}$ sampling uniformly without replacement from $\mathcal{Y} \setminus \mathcal{C}$

Approximate $P_{Y|Z}$

Logistic regression

$$P_\theta(y|z) = \frac{\exp(u_\theta(z, y))}{\sum_{y' \in \mathcal{Y}} \exp(u_\theta(z, y'))}$$

Build

- a set $\mathcal{C}$ containing all L_2 words in batch
- a set $\mathcal{N}$ sampling uniformly without replacement from $\mathcal{Y} \setminus \mathcal{C}$

Approximate normaliser for $P_\theta(y|z)$

$$\sum_{y' \in \mathcal{X}} \exp(s_\theta(z, y')) \approx \sum_{y' \in \mathcal{C}} \exp(s_\theta(z, y')) + \sum_{y' \in \mathcal{N}} \frac{|\mathcal{Y} \setminus \mathcal{C}|}{|\mathcal{N}|} \exp(s_\theta(z, y'))$$

- $s_\theta(z, y) = z^\top \mathbf{c}_y + b_y$
  $\mathbf{c}_y = \mathrm{lookup}_\theta(y)$ is a deterministic embedding, b_y a bias term

Complementary Sum Sampling (CSS) - Summary

The approximation effectively reduces the size of the support of the categorical variable

- in each batch, the support is made of the word types in the batch
- along with a random subset of "negative words"
- this is similar to "negative sampling" but improves on asymptotic behaviour
- it only affects the softmax: the model remains generative

Botev et al. (2017)

Outline

1. Embed-Align
2. Bayesian Skip-Gram
3. Practical

Bayesian Skip-Gram

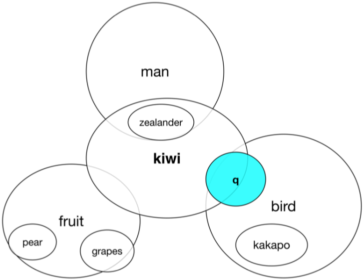


Figure 1: An idealized illustration of density embeddings. Unshaded ellipsoids encode prior densities of Gaussians. The shaded ellipsoid corresponds to the posterior for the word 'kiwi' when it appears in a context indicating that 'kiwi' refers to a bird.

"Representing a word as a distribution provides many potential benefits. For example, such embeddings let us encode generality of terms (e.g., 'kakapo' is a type of 'bird'), characterize uncertainty about semantic properties of the corresponding referent (e.g., a proper noun, such as 'John', encodes little about the person it refers to) or represent polysemy (e.g., 'kiwi' may refer to a fruit, a bird or a New Zealander)."

Bražinskas et al. (2017)

Bayesian Skip-Gram

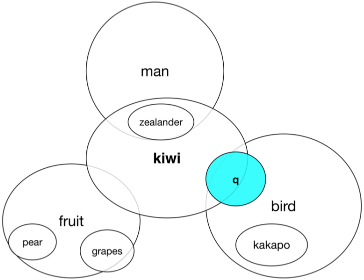


Figure 1: An idealized illustration of density embeddings. Unshaded ellipsoids encode prior densities of Gaussians. The shaded ellipsoid corresponds to the posterior for the word 'kiwi' when it appears in a context indicating that 'kiwi' refers to a bird.

"In principle, using densities to represent words provides a natural way of encoding entailment: the decision regarding entailment relation can be made by testing the level sets of the distributions for 'soft inclusion'. For example, in Figure 1, the ellipse for 'kakapo' lies within the ellipse for 'bird'. "

Bražinskas et al. (2017)

Complete specification

Generative model: for $i = 1, \ldots, m$

$$
\begin{aligned}
Z_i | x_i &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \mathrm{diag}(\boldsymbol{\sigma}_{x_i} \odot \boldsymbol{\sigma}_{x_i})) \\
\boldsymbol{\mu}_{x_i} &= \mathrm{lookup}_\theta(x_i) \\
\boldsymbol{\sigma}_{x_i} &= \mathrm{softplus}(\mathrm{lookup}_\theta(x_i))
\end{aligned}
$$

Complete specification

Generative model: for $i = 1, \ldots, m$

$$Z_i | x_i \sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \mathrm{diag}(\boldsymbol{\sigma}_{x_i} \odot \boldsymbol{\sigma}_{x_i}))$$
$$\boldsymbol{\mu}_{x_i} = \mathrm{lookup}_\theta(x_i)$$
$$\boldsymbol{\sigma}_{x_i} = \mathrm{softplus}(\mathrm{lookup}_\theta(x_i))$$

for $k \in \mathcal{K}_i = \{\underbrace{i-n, \ldots, i-1, i+1, \ldots, i+n}_{n \text{ words on each side of } x_i}\}$

Complete specification

Generative model: for $i = 1, \ldots, m$

$$
\begin{aligned}
Z_i | x_i &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \mathrm{diag}(\boldsymbol{\sigma}_{x_i} \odot \boldsymbol{\sigma}_{x_i})) \\
\boldsymbol{\mu}_{x_i} &= \mathrm{lookup}_\theta(x_i) \\
\boldsymbol{\sigma}_{x_i} &= \mathrm{softplus}(\mathrm{lookup}_\theta(x_i))
\end{aligned}
$$

for $k \in \mathcal{K}_i = \{\underbrace{i-n, \ldots, i-1, i+1, \ldots, i+n}_{n \text{ words on each side of } x_i}\}$

$$
\begin{aligned}
X_k | z_i &\sim \mathrm{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \mathrm{softmax}(\mathrm{affine}_\theta(z_i))
\end{aligned}
$$

Complete specification

Generative model: for $i = 1, \ldots, m$

$$
\begin{aligned}
Z_i | x_i &\sim \mathcal{N}(\boldsymbol{\mu}_{x_i}, \operatorname{diag}(\boldsymbol{\sigma}_{x_i} \odot \boldsymbol{\sigma}_{x_i})) \\
\boldsymbol{\mu}_{x_i} &= \operatorname{lookup}_\theta(x_i) \\
\boldsymbol{\sigma}_{x_i} &= \operatorname{softplus}(\operatorname{lookup}_\theta(x_i))
\end{aligned}
$$

for $k \in \mathcal{K}_i = \{\underbrace{i-n, \ldots, i-1, i+1, \ldots, i+n}_{n \text{ words on each side of } x_i}\}$

$$
\begin{aligned}
X_k | z_i &\sim \operatorname{Cat}(\mathbf{f}_i) \\
\mathbf{f}_i &= \operatorname{softmax}(\operatorname{affine}_\theta(z_i))
\end{aligned}
$$

Inference model

$$
\begin{aligned}
Z_i | x_{i-n}^{i+n} &\sim \mathcal{N}(\mathbf{u}_i, \operatorname{diag}(\mathbf{s}_i \odot \mathbf{s}_i)) \\
\mathbf{h}_i &= \sum_{k \in \mathcal{K}_i} \operatorname{relu}(\operatorname{affine}_\lambda([x_k, x_i])) \\
\mathbf{u}_i &= \operatorname{affine}_\lambda(h_i) \\
\mathbf{s}_i &= \operatorname{softplus}(\operatorname{affine}_\lambda(h_i))
\end{aligned}
$$

ELBO

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p_\theta(z_1^m|x_1^m)\right)$$

- observation model is trained discriminatively
  latent variables generate overlapping subsets of observations

ELBO - KL term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

KL term

$$\mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

Empirical Bayes: point estimate prior parameters

ELBO - KL term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

KL term

$$\mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$
$$= \sum_{i=1}^{m} \mathrm{KL}\left(q_\lambda(z_i|x_{i-n}^{i+n}) \mid\mid p_\theta(z_i|x_i)\right)$$

Empirical Bayes: point estimate prior parameters

ELBO - KL term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

KL term

$$\begin{aligned}
&\mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right) \\
&= \sum_{i=1}^{m} \mathrm{KL}\left(q_\lambda(z_i|x_{i-n}^{i+n}) \mid\mid p_\theta(z_i|x_i)\right) \\
&= \sum_{i=1}^{m} \mathrm{KL}\left(\underbrace{\mathcal{N}(\mathbf{u}_i, \mathrm{diag}(\mathbf{s}_i \odot \mathbf{s}_i))}_{\text{inference model}} \mid\mid \underbrace{\mathcal{N}(\boldsymbol{\mu}_{x_i}, \mathrm{diag}(\boldsymbol{\sigma}_{x_i} \odot \boldsymbol{\sigma}_{x_i}))}_{\text{prior}}\right)
\end{aligned}$$

Empirical Bayes: point estimate prior parameters

ELBO - Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

ELBO - Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right]$$

ELBO - Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] = \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \log P_\theta(x_k|z_i)\right]$$

ELBO - Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \mid\mid p_\theta(z_1^m|x_1^m)\right)$$

Likelihood term

$$\begin{aligned}\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \log P_\theta(x_k|z_i)\right] \\ &= \sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \mathbb{E}_{q_\lambda(z_i|x_{i-n}^{i+n})}\left[\log P_\theta(x_k|z_i)\right]\end{aligned}$$

ELBO - Likelihood term

$$\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] - \mathrm{KL}\left(q_\lambda(z_1^m|x_1^m) \,||\, p_\theta(z_1^m|x_1^m)\right)$$

Likelihood term

$$\begin{aligned}
\mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\log \prod_{i=1}^{m} \prod_{k \in \mathcal{K}_i} P_\theta(x_k|z_i)\right] &= \mathbb{E}_{q_\lambda(z_1^m|x_1^m)}\left[\sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \log P_\theta(x_k|z_i)\right] \\
&= \sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \mathbb{E}_{q_\lambda(z_i|x_{i-n}^{i+n})}\left[\log P_\theta(x_k|z_i)\right] \\
&= \sum_{i=1}^{m} \sum_{k \in \mathcal{K}_i} \mathbb{E}_{q_\lambda(z_i|x_{i-n}^{i+n})}\left[\log \mathrm{Cat}(x_k|\mathbf{f}_i)\right]
\end{aligned}$$

The softmax problem

To circumvent an expensive softmax, change the likelihood term

$$P_\theta(x|z) = \frac{u_\theta(z, x)}{\sum_{x' \in \mathcal{X}} u_\theta(z, x')} \qquad \text{with } u_\theta(\cdot, \cdot) > 0$$

The softmax problem

To circumvent an expensive softmax, change the likelihood term

$$P_\theta(x|z) = \frac{u_\theta(z, x)}{\sum_{x' \in \mathcal{X}} u_\theta(z, x')} \qquad \text{with } u_\theta(\cdot, \cdot) > 0$$

and re-write the likelihood term

$$\mathbb{E}_{q_\lambda(z)}\left[\log P_\theta(x|z)\right] = \mathbb{E}_{q_\lambda(z)}\left[\log u_\theta(z, x) - \log \sum_{x' \in \mathcal{X}} u_\theta(z, x')\right]$$

The softmax problem

To circumvent an expensive softmax, change the likelihood term

$$P_\theta(x|z) = \frac{u_\theta(z,x)}{\sum_{x' \in \mathcal{X}} u_\theta(z,x')} \qquad \text{with } u_\theta(\cdot,\cdot) > 0$$

and re-write the likelihood term

$$\begin{aligned}\mathbb{E}_{q_\lambda(z)}\left[\log P_\theta(x|z)\right] &= \mathbb{E}_{q_\lambda(z)}\left[\log u_\theta(z,x) - \log \sum_{x' \in \mathcal{X}} u_\theta(z,x')\right] \\ &= \mathbb{E}_{q_\lambda(z)}\left[\log u_\theta(z,x)\right] - \mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} u_\theta(z,x')\right]\end{aligned}$$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z, x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z | \boldsymbol{\mu}_x, \text{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z, x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

And bound $\mathbb{E}_{q_\lambda(z)} \left[\log \sum_{x' \in \mathcal{X}} u_\theta(z, x')\right]$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z,x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

And bound $\mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} u_\theta(z, x')\right]$

$$= \mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} P(x)\mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]$$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z,x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

And bound $\mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x'\in\mathcal{X}} u_\theta(z,x')\right]$

$$= \mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x'\in\mathcal{X}} P(x)\mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]$$

$$= \mathbb{E}_{q_\lambda(z)}\left[\log \mathbb{E}_{P(x)}\left[\mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]\right]$$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z,x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

And bound $\mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} u_\theta(z,x')\right]$

$$\begin{aligned}
&= \mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} P(x)\mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right] \\
&= \mathbb{E}_{q_\lambda(z)}\left[\log \mathbb{E}_{P(x)}\left[\mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]\right] \\
&\overset{\text{JI}}{\geq} \mathbb{E}_{q_\lambda(z)}\left[\mathbb{E}_{P(x)}\left[\log \mathcal{N}(z|\boldsymbol{\mu}_x, \mathrm{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]\right]
\end{aligned}$$

Lowerbound on likelihood term

Then (by design) let

$$u_\theta(z, x) = \underbrace{P(x)}_{\text{fixed}} \mathcal{N}(z|\boldsymbol{\mu}_x, \operatorname{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))$$

And bound $\mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} u_\theta(z, x')\right]$

$$\begin{aligned}
&= \mathbb{E}_{q_\lambda(z)}\left[\log \sum_{x' \in \mathcal{X}} P(x)\mathcal{N}(z|\boldsymbol{\mu}_x, \operatorname{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right] \\
&= \mathbb{E}_{q_\lambda(z)}\left[\log \mathbb{E}_{P(x)}\left[\mathcal{N}(z|\boldsymbol{\mu}_x, \operatorname{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]\right] \\
&\overset{\text{JI}}{\geq} \mathbb{E}_{q_\lambda(z)}\left[\mathbb{E}_{P(x)}\left[\log \mathcal{N}(z|\boldsymbol{\mu}_x, \operatorname{diag}(\boldsymbol{\sigma}_x \odot \boldsymbol{\sigma}_x))\right]\right]
\end{aligned}$$

- $P(x)$ does not depend on θ, we can compute an MC estimate e.g. empirical (unigram) distribution

Outline

Practical

- Skip-gram (Mikolov et al., 2013)
- Bayesian skip-gram (Bražinskas et al., 2017)
- Embed-Align (Rios et al., 2018)

Comparison

	SkipGram	Bayesian SkipGram	EmbedAlign
LVM		✓	✓
Generative training			✓
Prior		type-specific Gaussian	$\mathcal{N}(0, I)$
Inference model		FFNN	BiLSTM
Softmax	negative sampling	JI	CSS

Literature I

Aleksandar Botev, Bowen Zheng, and David Barber. Complementary sum sampling for likelihood approximation in large scale classification. In *Artificial Intelligence and Statistics*, pages 1030–1038, 2017.

Arthur Bražinskas, Serhii Havrylov, and Ivan Titov. Embedding words as distributions with a bayesian skip-gram model. *arXiv preprint arXiv:1711.11027*, 2017.

Yoav Goldberg and Omer Levy. word2vec explained: Deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Literature II

Miguel Rios, Wilker Aziz, and Khalil Sima'an. Deep generative model for joint alignment and word representation. *arXiv preprint arXiv:1802.05883*, 2018.