Learning Semantic Similarity in a Continuous Space:

Semantic Similarity Learning Using Deep Generative Model

He Ying

Mar 17th, 2019

CONTENT

- Introduction to Semantic Similarity Learning
- VAE-Siamese Framework
  - VAE: inferring and learning intents (sentence presentations)
  - Variational Siamese Network: learning semantic similarity in a latent continuous space
- Results Analysis & Conclusion
- Brief Introduction to Seq2Seq Attention Mechanism

1. Introduction to Semantic Similarity Learning

- Semantics: used for understanding human expressions.
- Application of semantic similarity measurement between pairs of sentences:
  - Conversation systems (chatbots, FAQ)
  - Knowledge deduplication
  - Image captioning

1. Introduction to Semantic Similarity Learning

- Semantic representations of words:
  - One-hot Vector (Too sparse, lack of capability to relate to similarity in space)
  - Word2Vec (CBOW, Skip-gram)
  - GloVe
  - ELMo
  - BERT

1. Introduction to Semantic Similarity Learning

- Measuring Semantic Similarity:

  - WMD (Word Mover's Distance)

  (Measuring the dissimilarity between two bag-of-vectors (embedded words) with an optimal transport distance metric, Wasserstein 1)

  (Eg. "*why do cats like mice ?*" and "*why do mice like cats ?*" ==> Label Y: duplicate)

  - TF-IDF, Okapi BW-25, SIF etc.

  (Simple approaches to obtain a sentence representation from words' embeddings is to compute the barycenter of embedded words.)

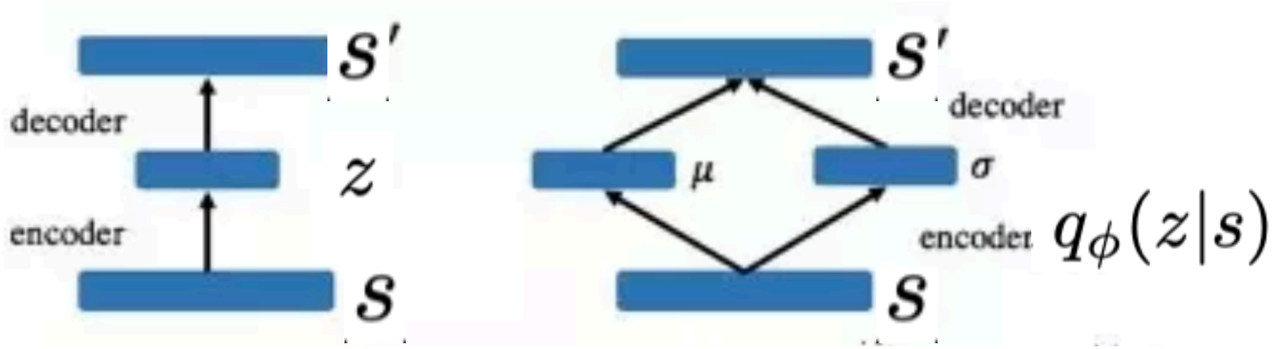
  (Still fail to capture semantics because of the bag-of-word assumption.)

2. VAE-Siamese Framework

2.1. VAE (Variational Auto-encoder)

Generative Models

- VAE in NLP
- GAN in CV

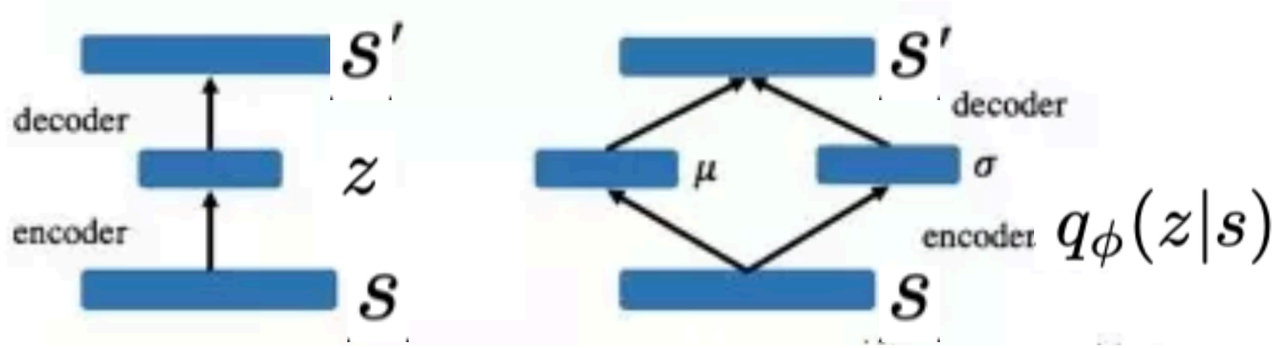


Standard Auto-encoder **Variational Auto-encoder**

VAE can be regarded as a regularized version of SAE by introducing a random latent variable.

2. VAE-Siamese Framework

2.1. VAE (Variational Auto-encoder)



Standard Auto-encoder **Variational Auto-encoder**

- The VAE learns, for any input s , a posterior distribution $q_\phi(z|s)$ over latent code z , where $z \sim N(\mu(s), \sigma(s))$, $z \in R^h$.

- The VAE's decoder parameterizes another distribution $p_\theta(s|z)$.

$$L_{\theta;\phi}(s) = E_{q_\phi(z|s)}[\log p_\theta(s|z)] - KL(q_\phi(z|s)||p(z))$$

2. VAE-Siamese Framework

2.1. VAE (Variational Auto-encoder)

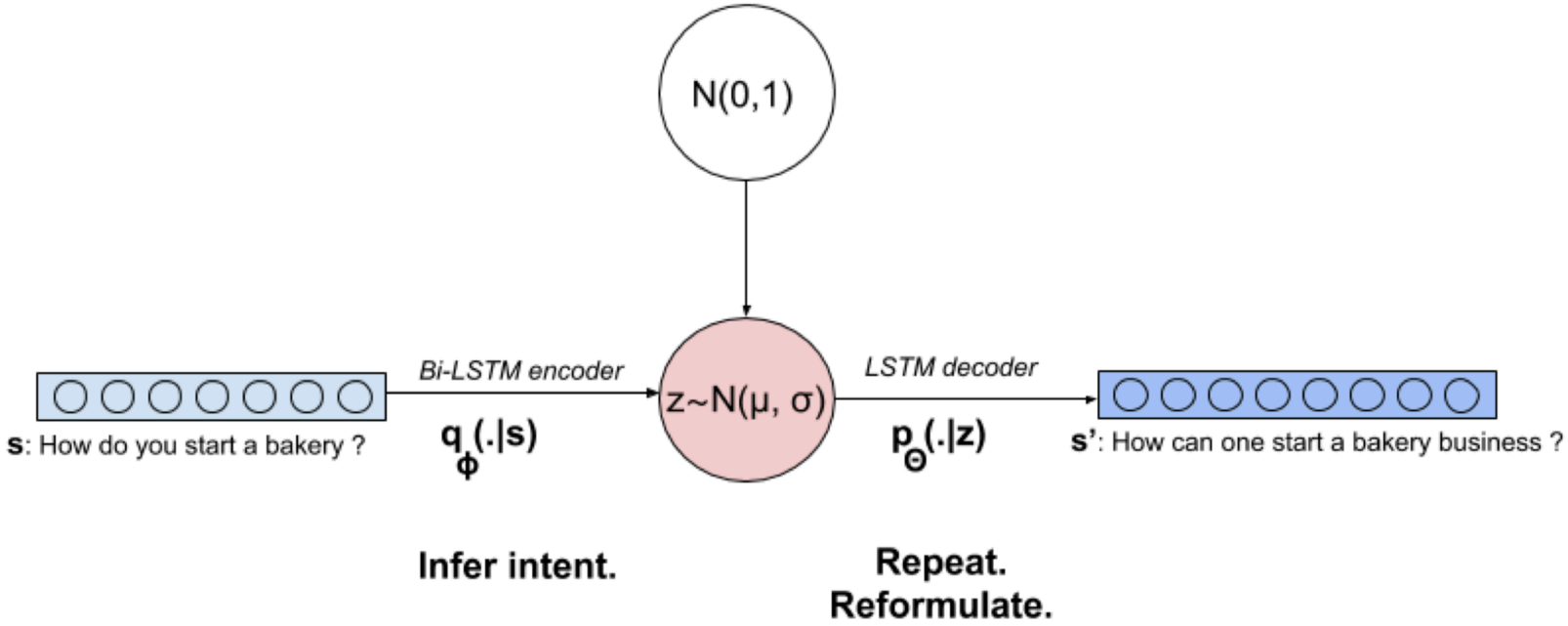


Figure 2. Bayesian framework and neural architecture to learn semantic representation of sentences.

$$-L_{\theta;\phi}(s,s') = -E_{q_\phi(z|s)}[\log p_\theta(s'|z)] + \kappa KL(q_\phi(z|s)||N(0,I))$$

$$\kappa = sigmoid(0.002(step - 2500))$$

2. VAE-Siamese Framework

2.1. VAE (Variational Auto-encoder)

$$-L_{\theta;\phi}(s, s') = -E_{q_\phi(z|s)}[\log p_\theta(s'|z)] + \kappa KL(q_\phi(z|s)||N(0, I))$$

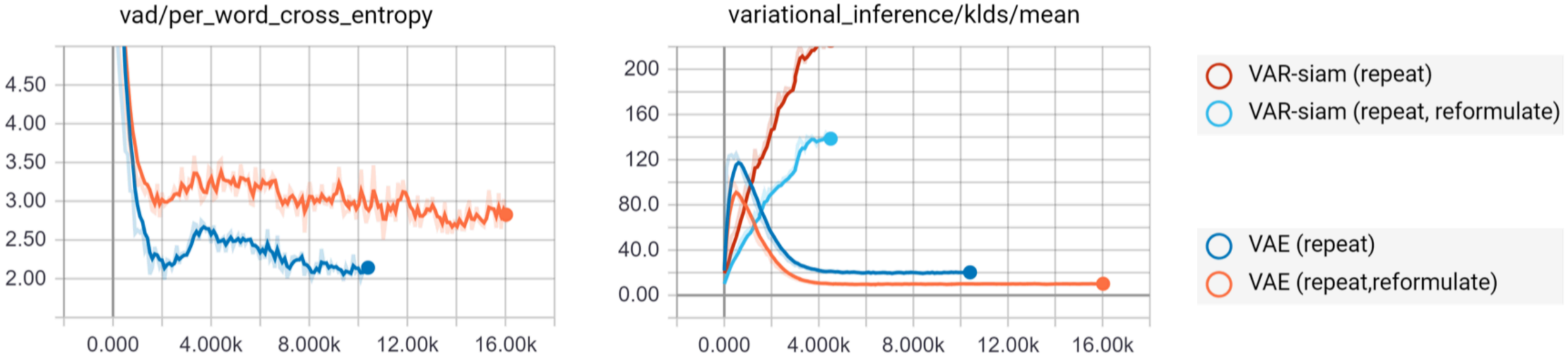


Figure 3: Results from our generative pretraining: $repeat$ (dark blue), $repeat, reformulate$ (orange). Left: Per word cross entropy. Right: KL divergence $KL(q_\phi(z|s)||N(0, I))$.

2. VAE-Siamese Framework

2.2. Variational Siamese Network

– We train our model successively on two tasks (generative and discriminative) and use the first step as a smooth initialization ($N(0, I)$ prior and semi-supervised setting) for the second one, learning similarity.

– Metrics of Similarity: Wasserstein 2 (W_2^2) , Hadamard product $\mu_1 \mu_2 \in \mathbb{R}^h$.

$$W_2^2(p_1, p_2) = \sum_{i=1}^{h} (\mu_1^i - \mu_2^i)^2 + (\sigma_1^i - \sigma_2^i)^2$$

where, $p_1 = N(\mu_1, \sigma_1) \in \mathbb{R}^h$, $p_2 = N(\mu_2, \sigma_2) \in \mathbb{R}^h$.

2. VAE-Siamese Framework

2.2. Variational Siamese Network

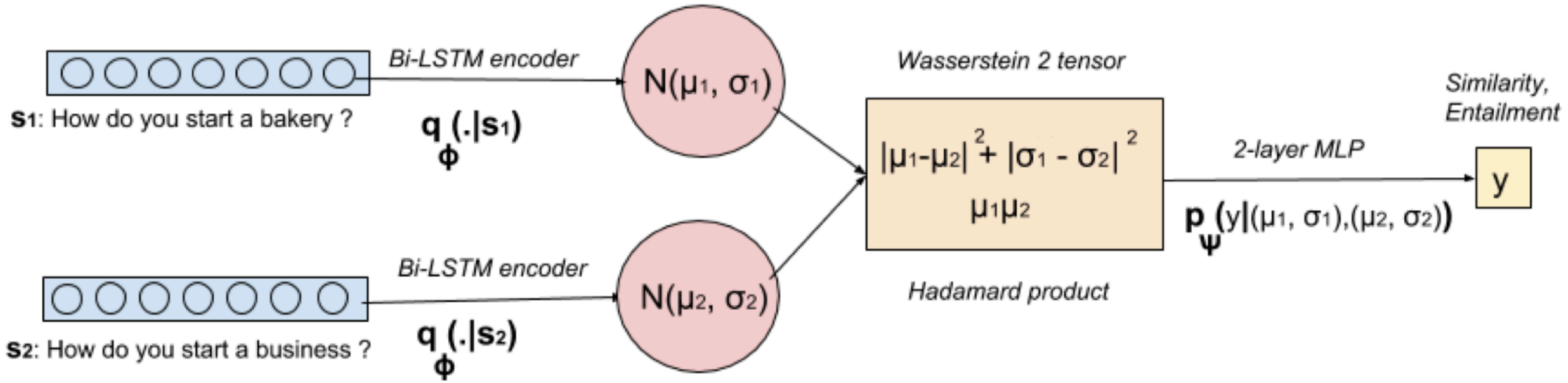


Figure 4: Our *variational siamese* network to measure and learn semantic similarity.

$$L_{\psi;\phi}(s_1, s_2) = -y \log p_\psi(y|q_\phi(.|s_1), q_\phi(.|s_2)) + \lambda||\psi||_1$$

$$\lambda = 0.00001$$

3. Results Analysis

3.1. Experimental details

- Embedding words in a 300 dimensional space using Glove vectors pre-trained on Wikipedia 2014 and Gigaword 5 as initialization.
- Variational space (μ, σ) is of dimension $h = 1000$.
- Our bi-LSTM encoder network consists of a single layer of 2h neurons and our LSTM and the LSTM decoder has a single layer with 1000 neurons.
- The MLP's inner layer has 1000 neurons.
- SGD with ADAM optimizer (learning rate = 0.001) and batches of size 256 and 128.

3. Results Analysis

3.2. Results

Table 1: Quora question pairs dataset result. The split considered is that of BiMPM [47]. Models are sorted by decreasing test accuracy. Results with † are reported in [22] and ‡ in [25]. SWEM [49] stands for Simple Word Embedding based Models.

Model	Read pairs separately	Accuracy Dev	Accuracy Test	Generative (pre)training
DIIN [22] †	False	89.44	89.06	False
VAR-Siamese (with repeat/reformulate)	**True**	**89.05**	**88.86**	**True**
pt-DECATTchar [48] †	False	88.89	88.40	False
VAR-Siamese (with repeat)	**True**	**88.18**	**88.24**	**True**
BiMPM [47] †	False	88.69	88.17	False
pt-DECATTword [48] †	False	88.44	87.54	False
L.D.C †	False	-	85.55	False
Siamese-GRU Augmented [25] ‡	True	-	85.24	False
Multi-Perspective-LSTM †	False	-	83.21	False
SWEM-concat [49]	True	-	83.03	False
SWEM-aver [49]	True	-	82.68	False
Siamese-LSTM †	True	-	82.58	False
SWEM-max [49]	True	-	82.20	False
Multi-Perspective CNN †	False	-	81.38	False
DeConv-LVM [37]	True	-	80.40	True
Siamese-CNN †	True	-	79.60	False
VAR-Siamese (w/o pretraining)	**True**	**62.16**	**62.48**	**False**
Bias in dataset (baseline)	-	62.16	62.48	-

3. Results Analysis

3.2. Results

Table 2: Question retrieval (best match, confidence) with our proposed model on Quora question pairs dataset (537 088 unique questions). All queries were retrieved in less than a second on two Tesla K80 GPU.

Query	Retrieved question 1/537088	MLP's output, Confidence
Hi! If I run the speed of the light, what would the world look like?	What would happen if I travel with a speed of light?	99.68%
How do you do a good gazpacho?	How can I make good gazpacho ?	99.52%
What can I do to save our planet?	How can I save the planet?	95.91%
What is the difference between a cat?	What is the benefit of a cat?	90.99%
Can we trust data?	Can I trust google drive with personal data?	57.13%

3. Results Analysis

3.2. Results

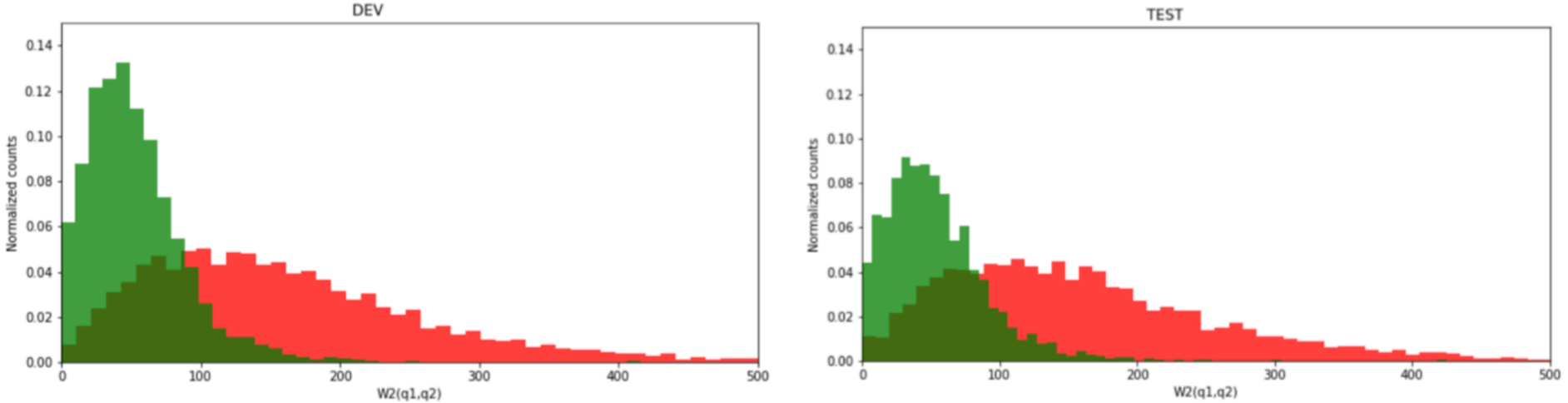


Figure 5: Empirical distribution of Wasserstein 2 distance between pairs of intents for duplicate (green) and not duplicate (red) pairs, after variational siamese training (with repeat, reformulate pretraining). Left: Quora Dev set. Right: Quora Test set.

3. Results Analysis

3.2. Results

Table 3: Discriminative power of Wasserstein 1 and 2 on Quora dataset.

Encoding	Wasserstein	Area under ROC curve		Training
		Dev set	Test set	labels
VAR-siamese (with repeat)	W2	87.11	86.74	Positive
VAR-siamese (with repeat/reformulate)	W2	86.88	86.27	& Negative
VAE (repeat/reformulate)	W2	77.70	77.44	Positive
Word Mover's Distance [1] (bag-of-word)	W1	73.47	73.10	None
VAE (repeat)	W2	70.67	69.91	None

4. Conclusion

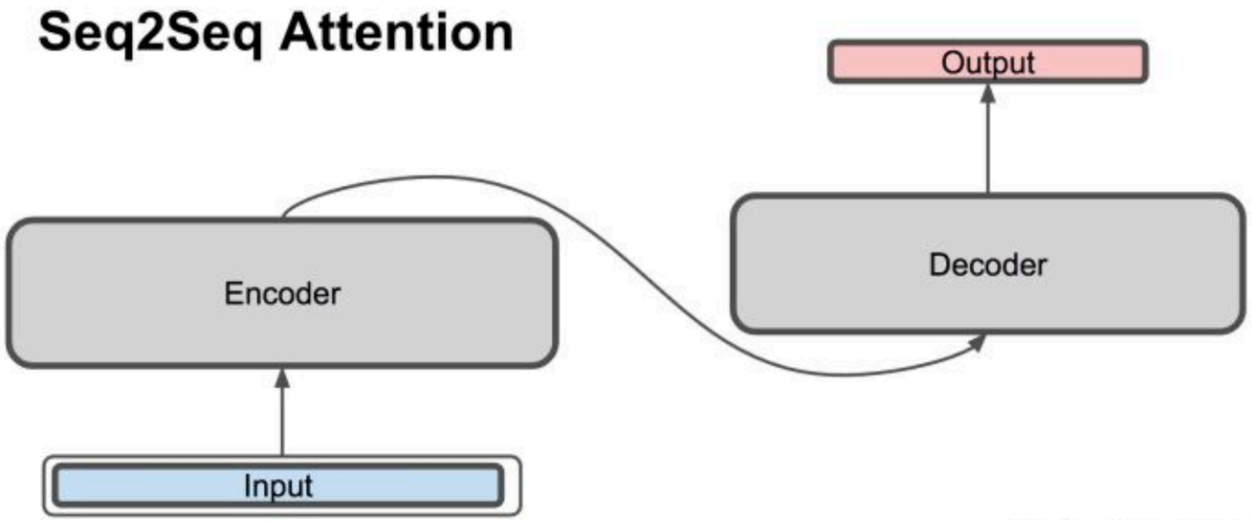
- Decompose the representation of a sentence in a mean vector and a diagonal covariance matrix to account for uncertainty and ambiguity in language.

- Encoding semantic information in an explicit sentence representation by learning to repeat, reformulate with the generative framework.

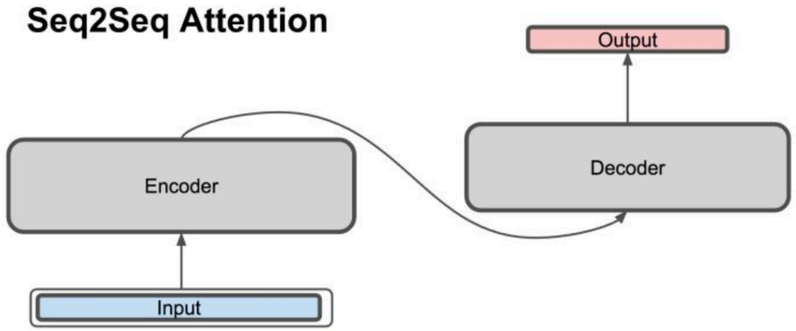
- The variational siamese network extends Word Mover's Distance to continuous representation of sentences in measuring semantic similarity.

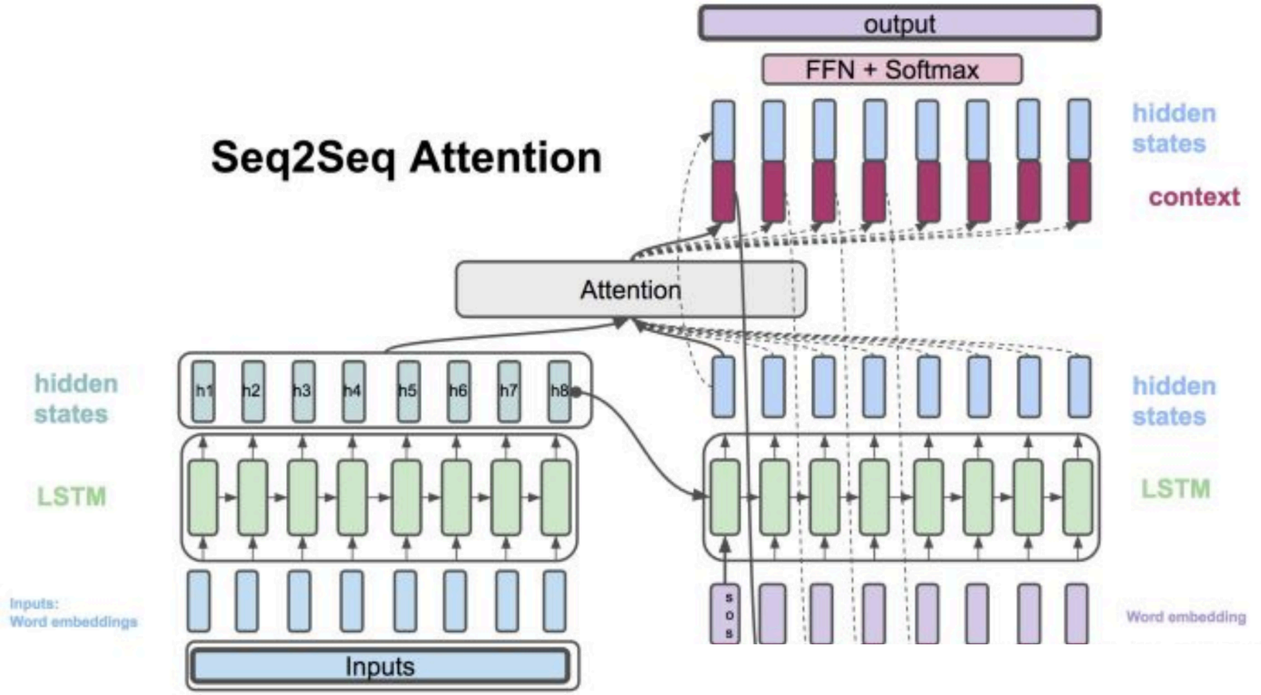
- Our approach performs strongly on Quora question pairs dataset and experiments show its effectiveness for question retrieval in knowledge databases with a million entries.

5. Seq2Seq Attention Mechanism

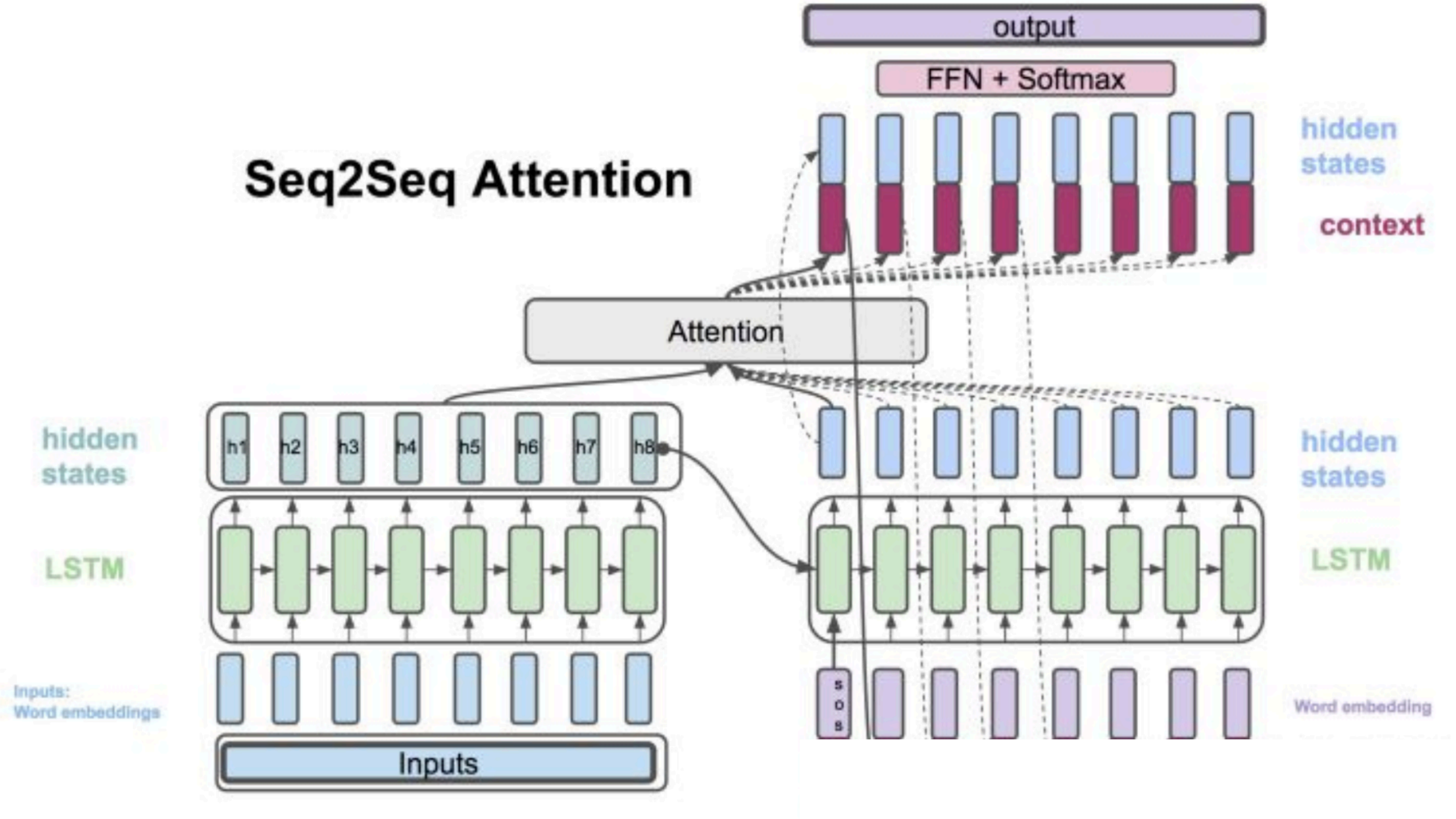


5. Seq2Seq Attention Mechanism

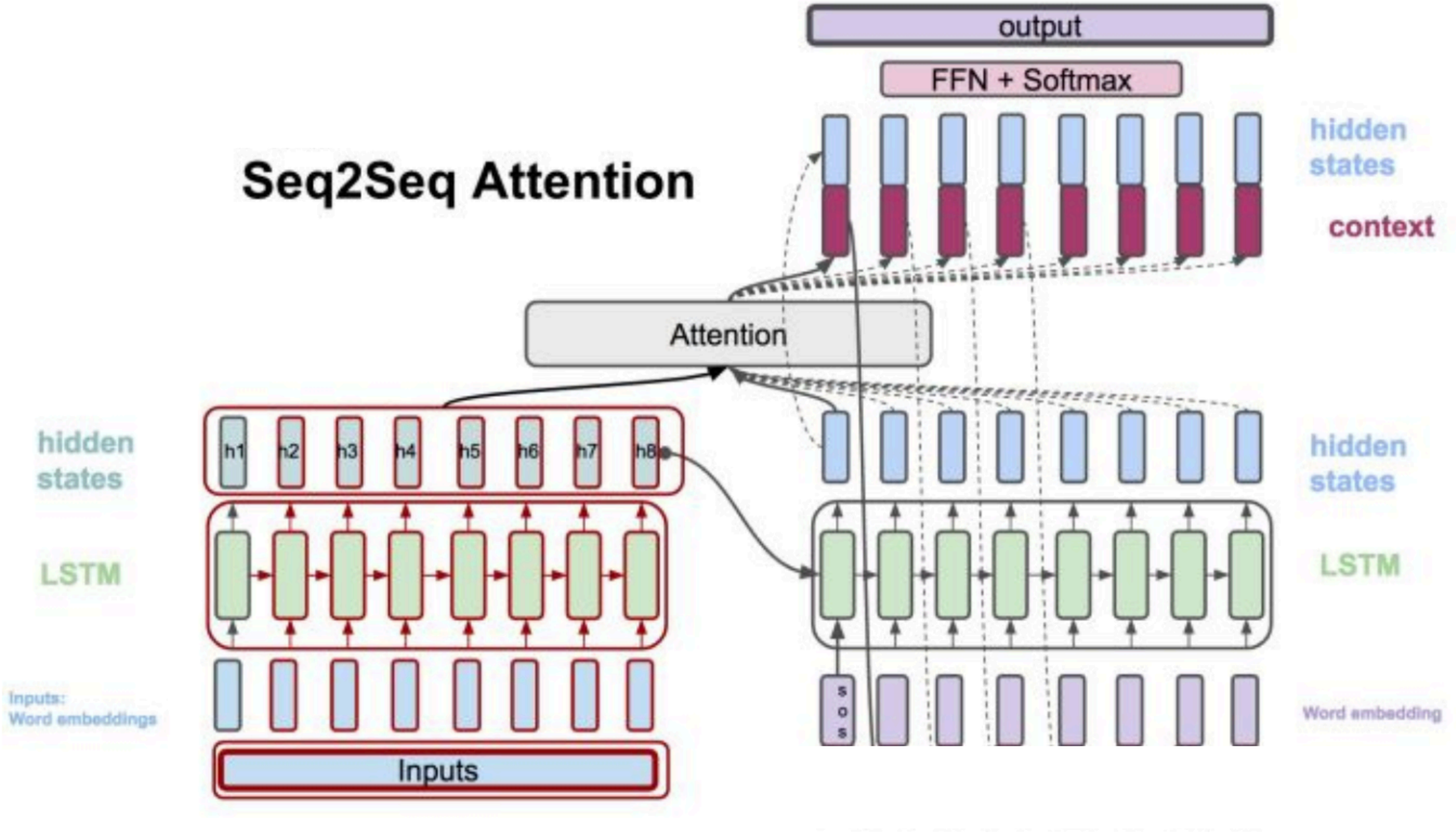




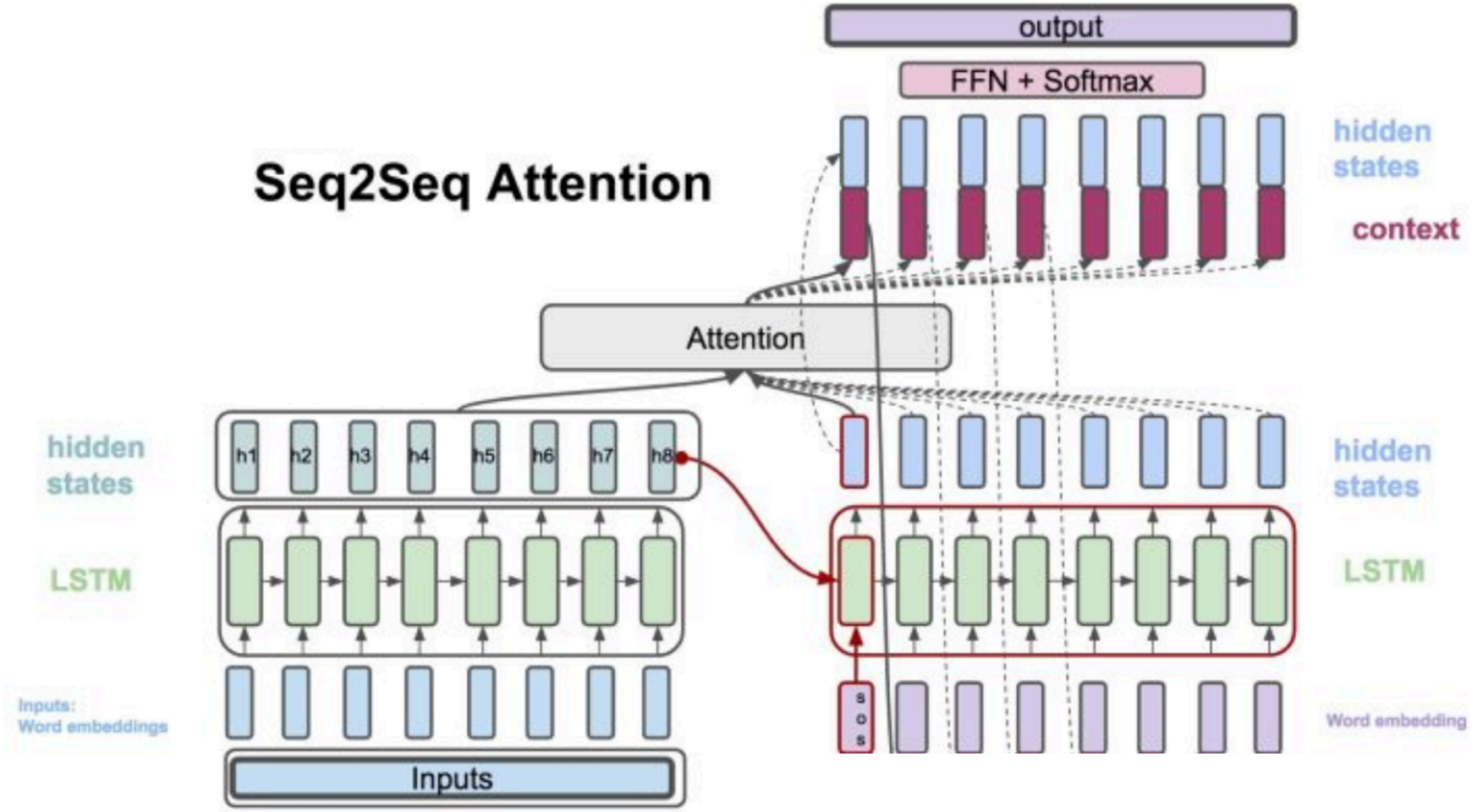
5. Seq2Seq Attention Mechanism



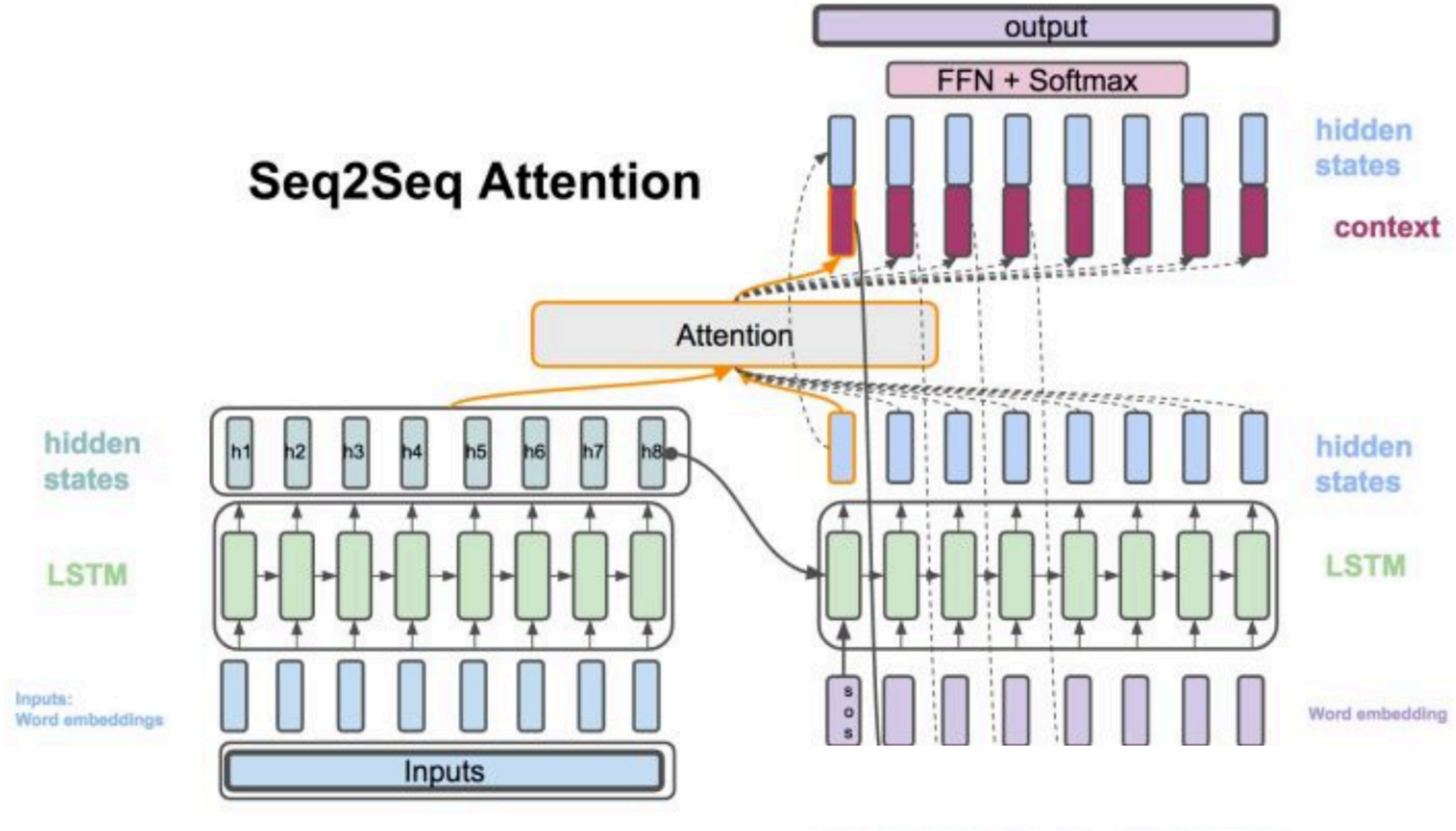
5. Seq2Seq Attention Mechanism



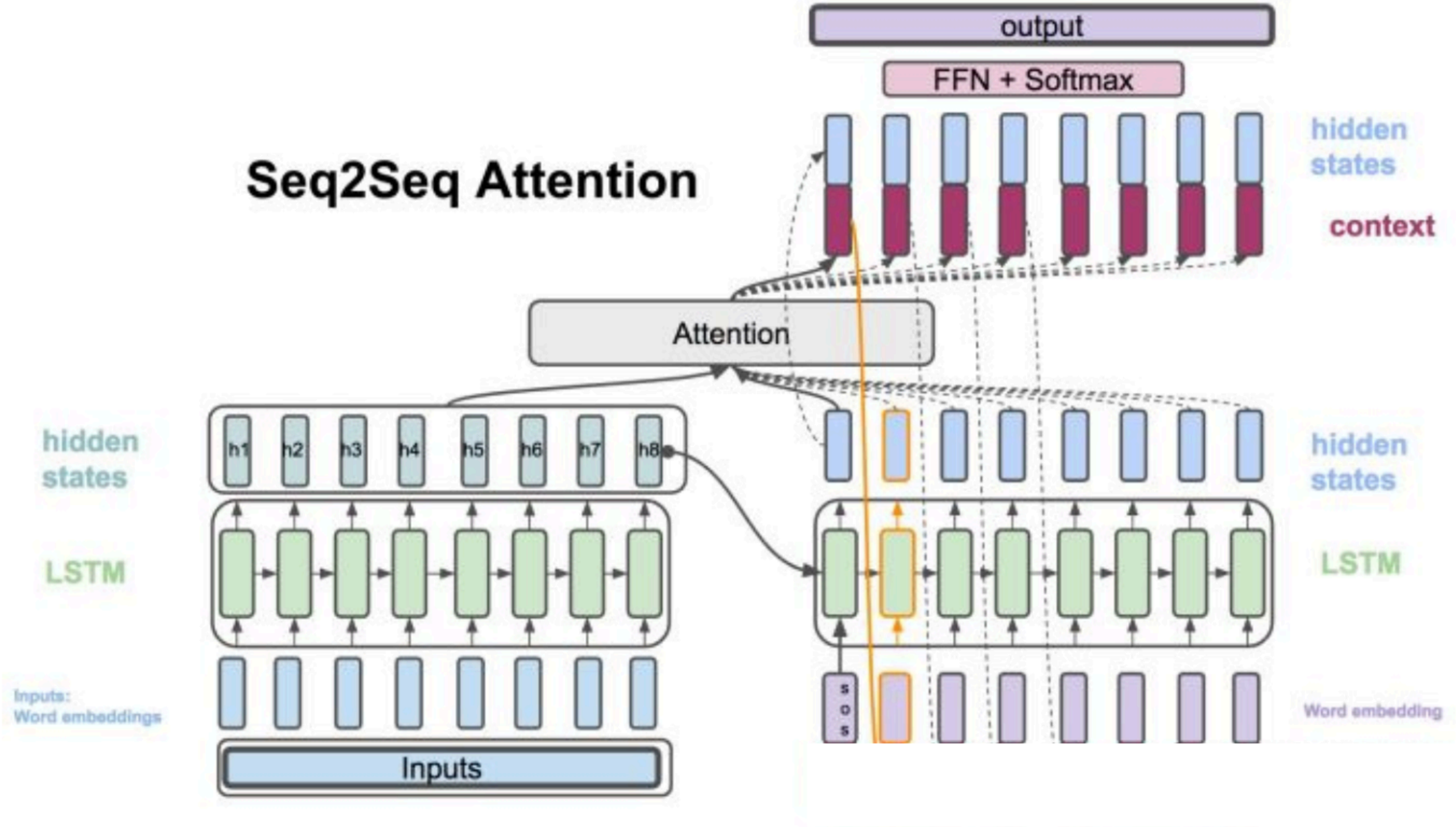
5. Seq2Seq Attention Mechanism



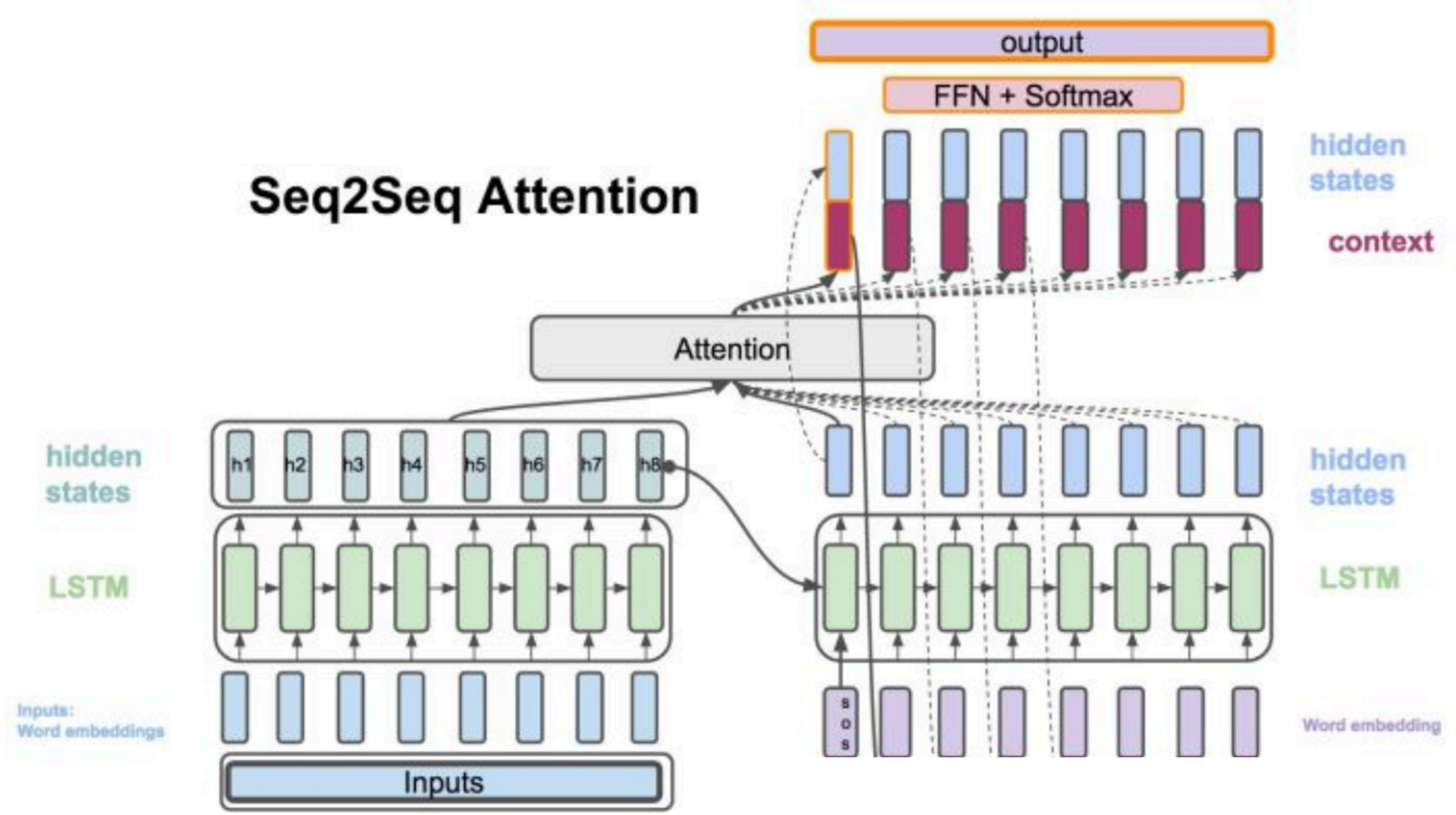
5. Seq2Seq Attention Mechanism



5. Seq2Seq Attention Mechanism



5. Seq2Seq Attention Mechanism



Thank You !