Probabilistic Transformer

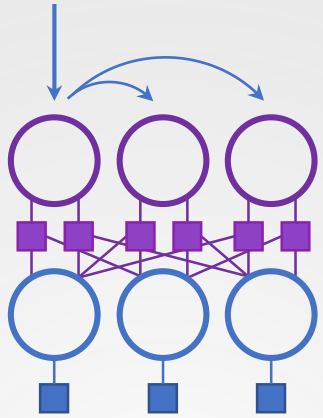
A Probabilistic Dependency Model for Contextual Word Representation

Haoyi Wu, Kewei Tu*

School of Information Science and Technology, ShanghaiTech University

Shanghai Engineering Research Center of Intelligent Vision and Imaging

上海科技大学
ShanghaiTech University



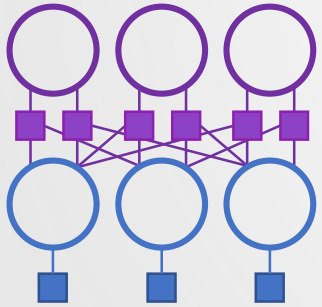
Part 1.
The Probabilistic Transformer

Nothing more than a conditional random field (CRF)

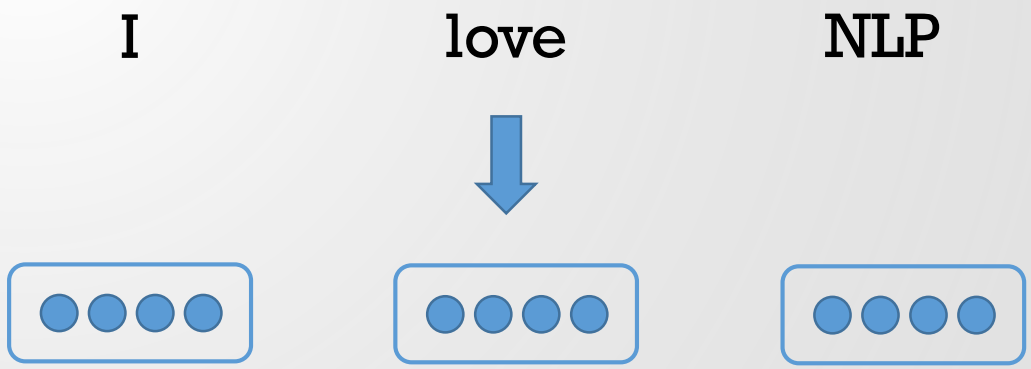
What is a Probabilistic Transformer?

Part 1. The Probabilistic Transformer

- A probabilistic model…
- …for contextual word representation

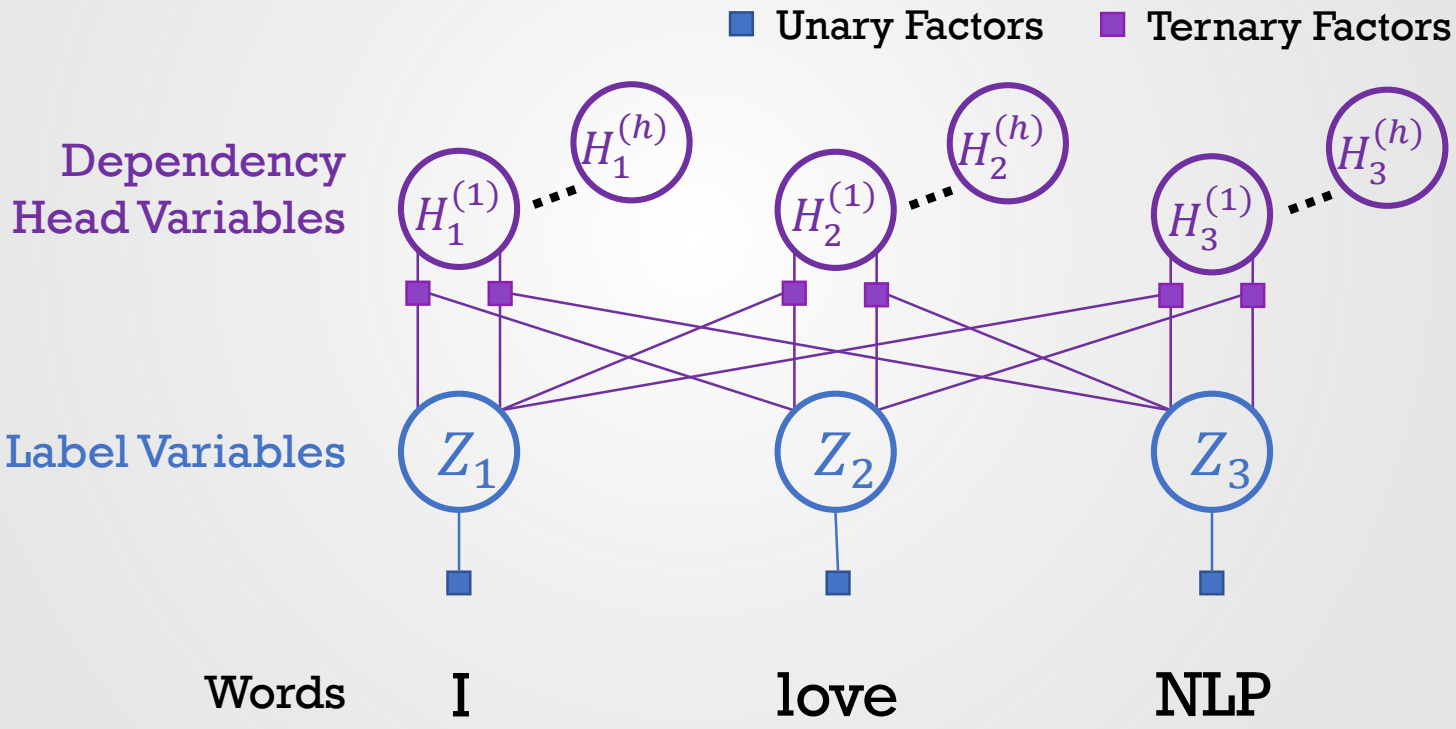


based on a CRF…



The CRF Model

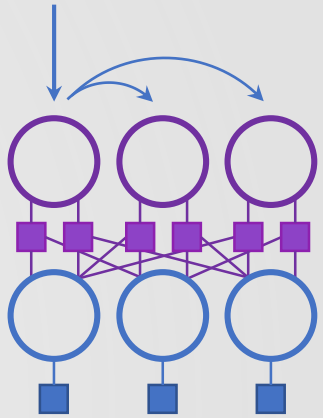
Part 1. The Probabilistic Transformer

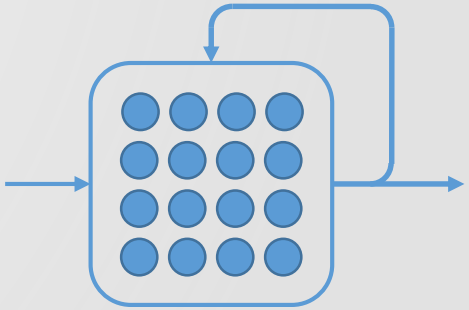


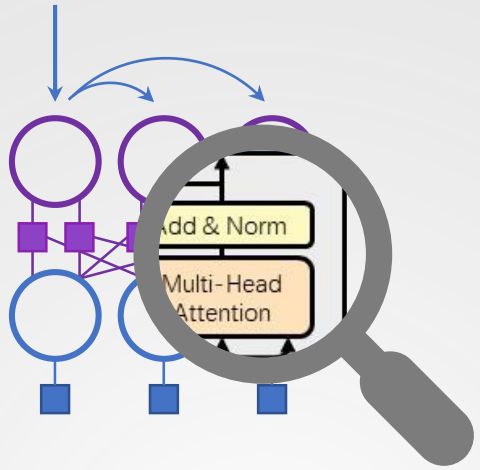
A Probabilistic Transformer

Part 1. The Probabilistic Transformer

- The CRF
  - models the discrete latent representations of words…
  - … as well as the dependency arcs between them
- Inference
  - Mean Field Variational Inference (MFVI)
- Learning
  - Differentiable! Gradient descent…





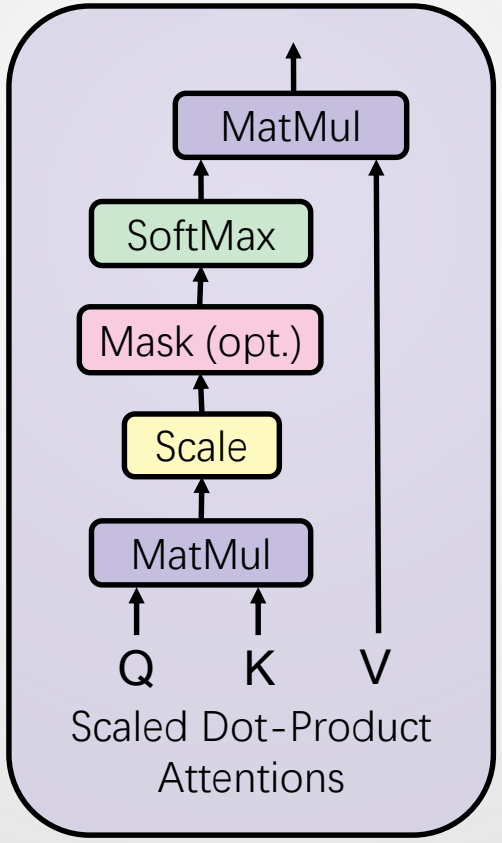


Part 2.
The Computation Graph

Strikingly similar to that of transformers

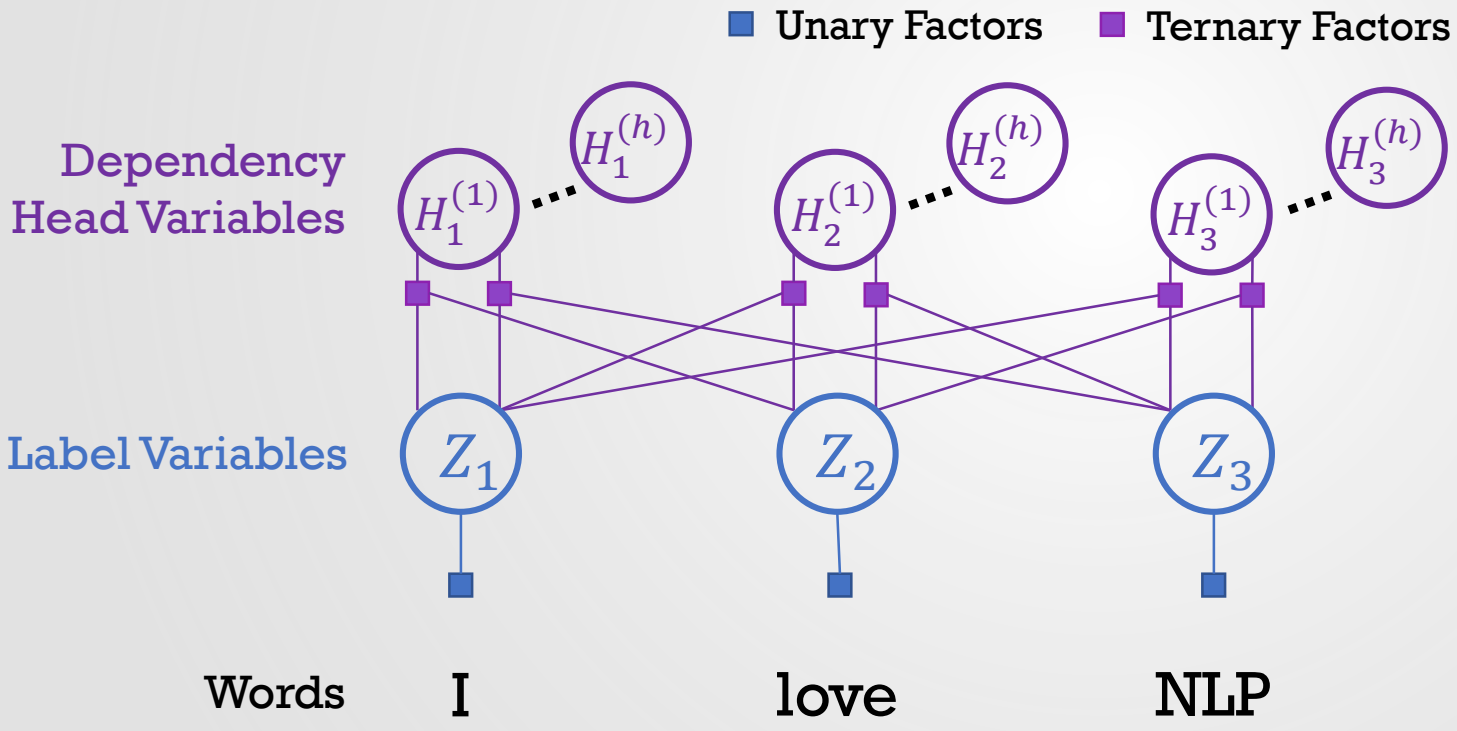
Single-Channel Update

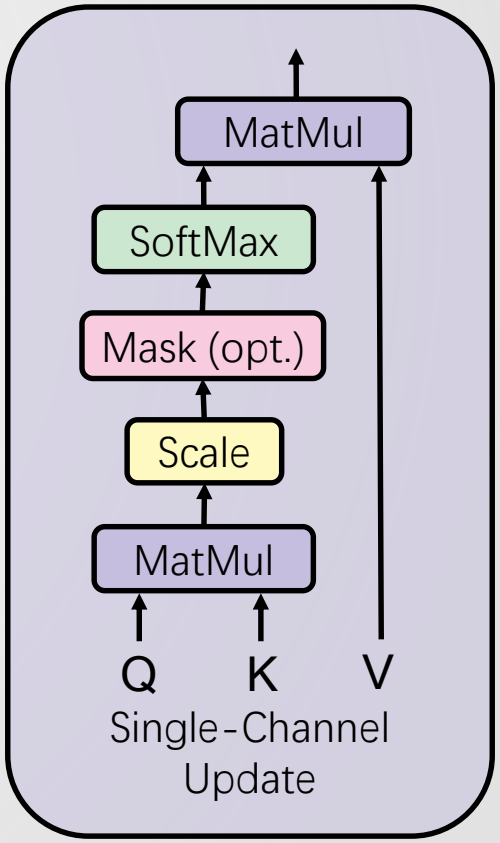
Part 2. The Computation Graph



Single-Channel Update

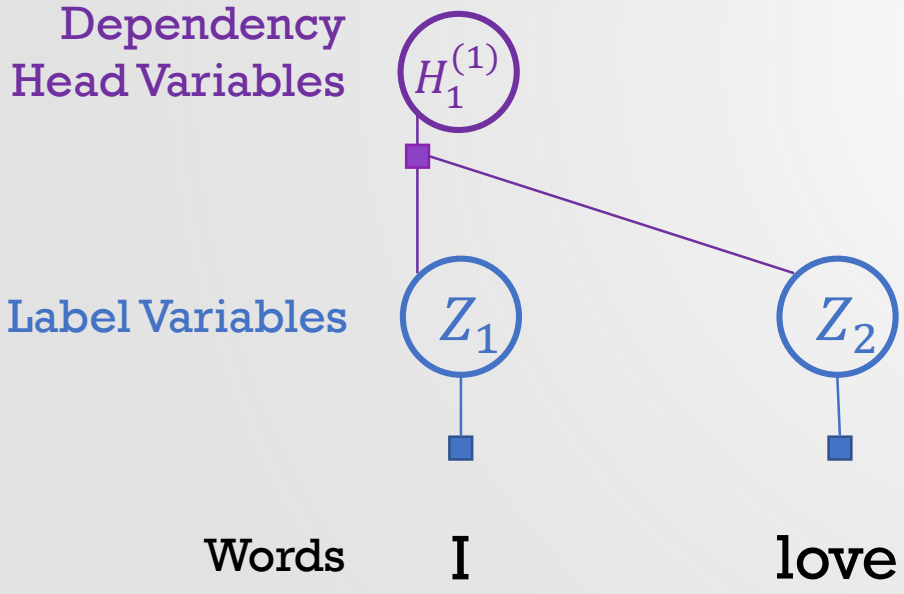
Part 2. The Computation Graph

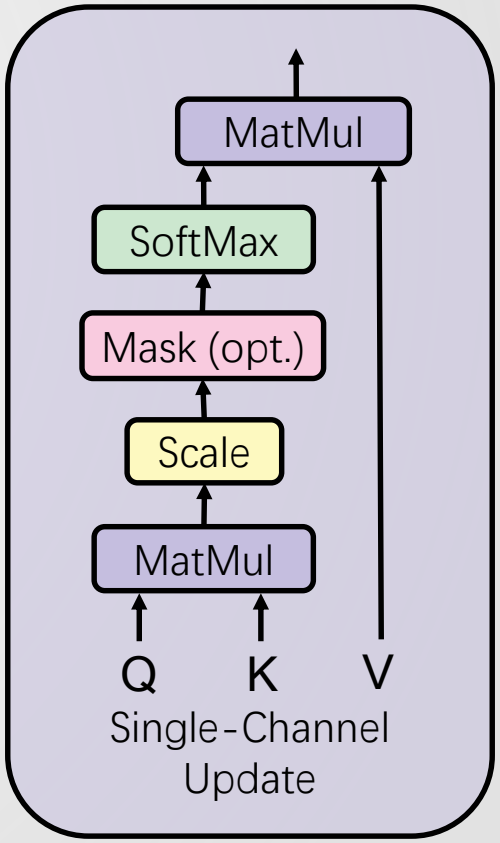




Single-Channel Update

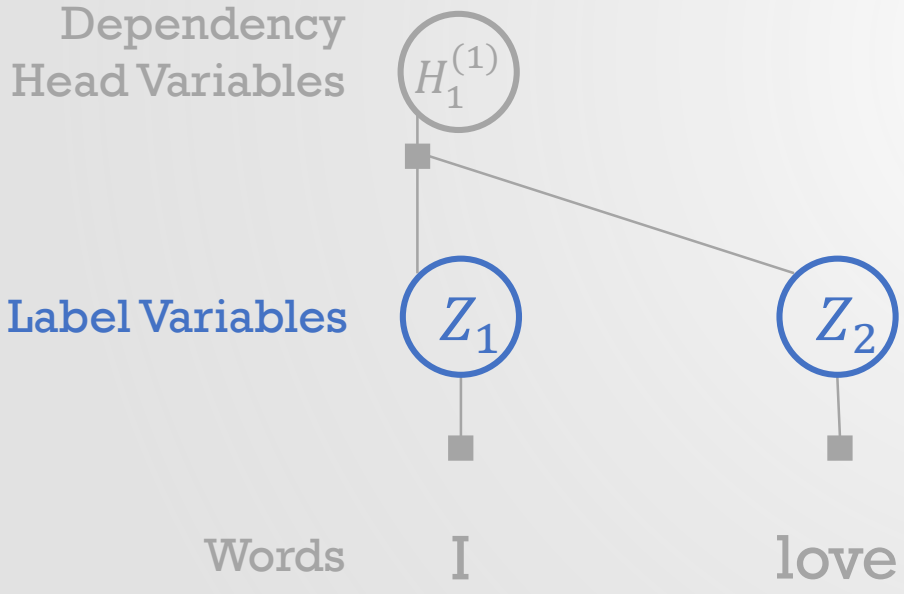
Part 2. The Computation Graph

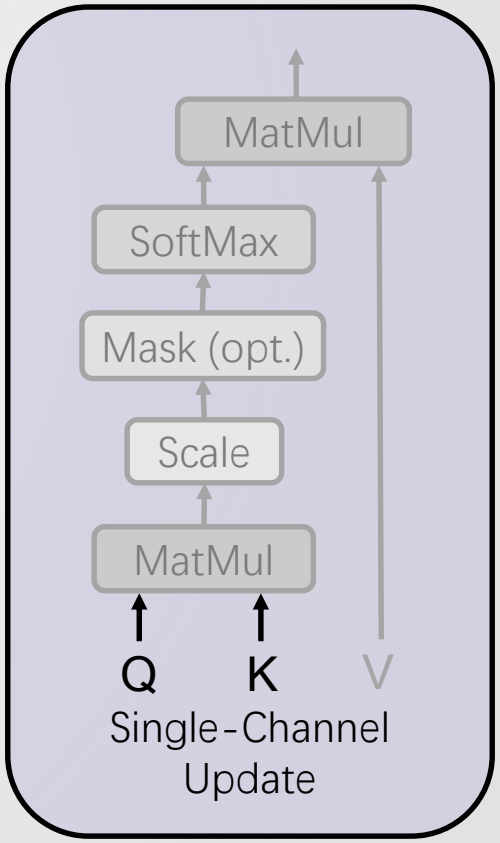




Single-Channel Update

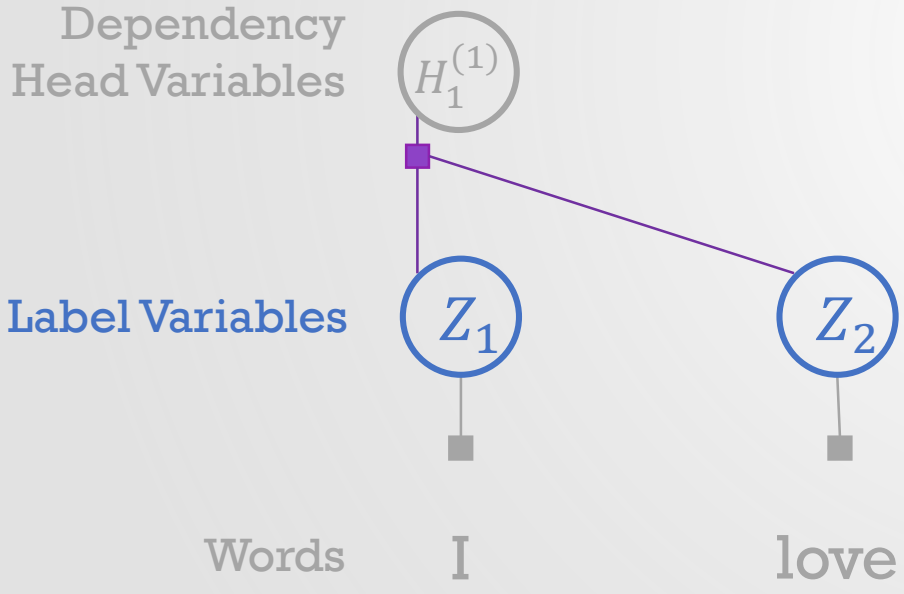
Part 2. The Computation Graph

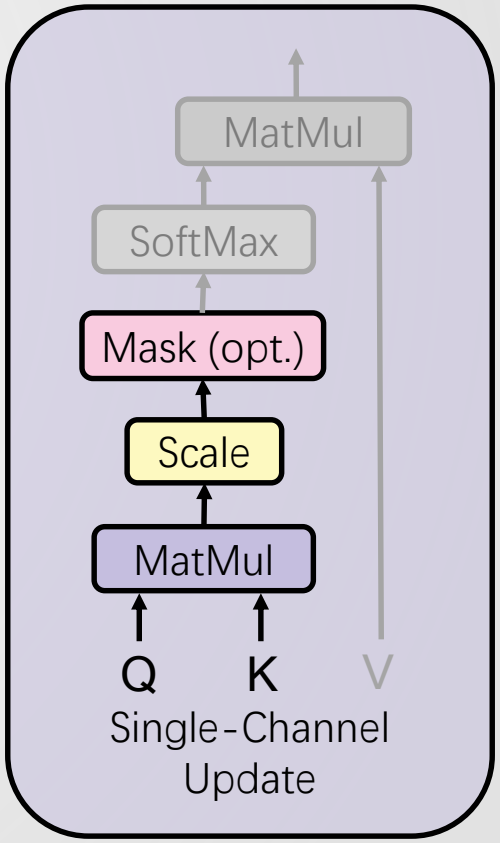




Single-Channel Update

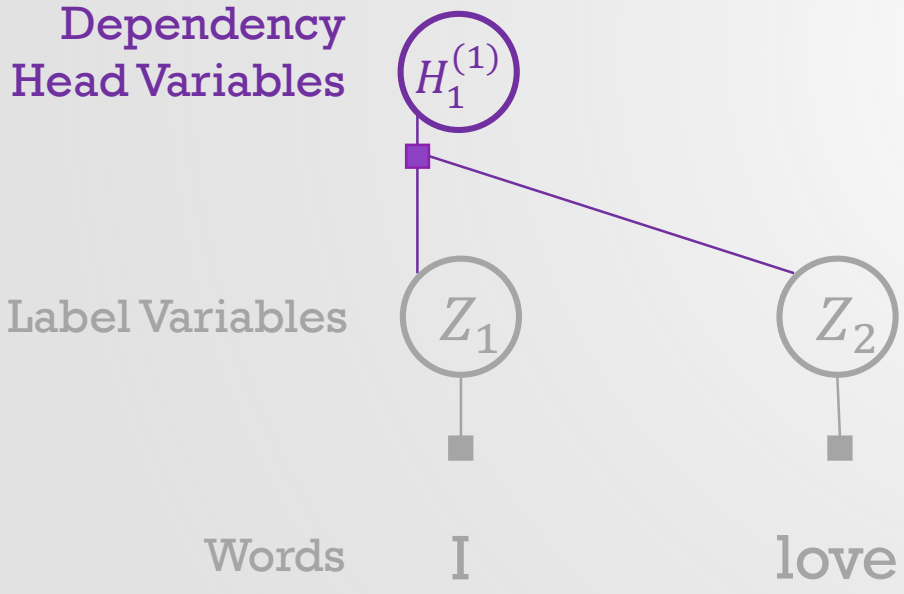
Part 2. The Computation Graph

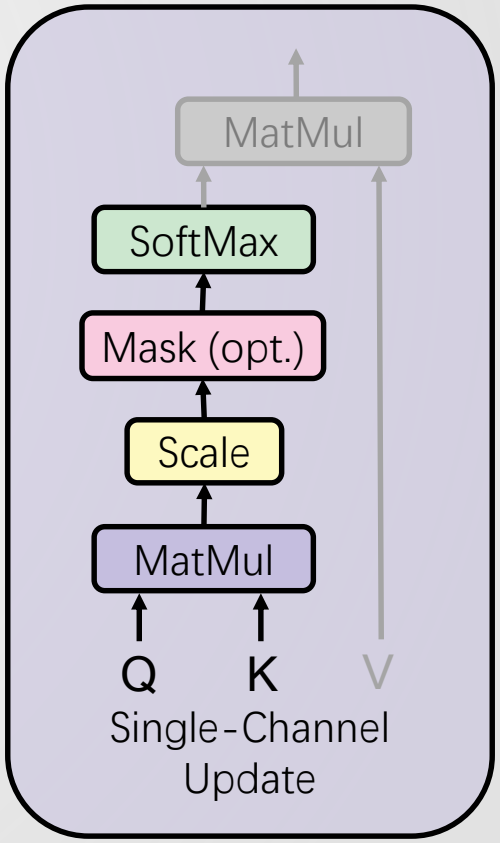




Single-Channel Update

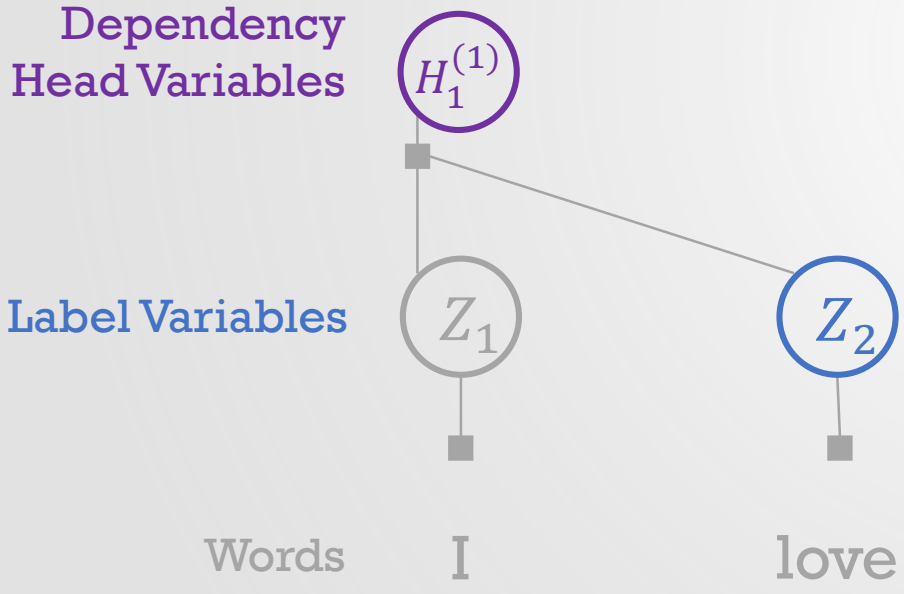
Part 2. The Computation Graph

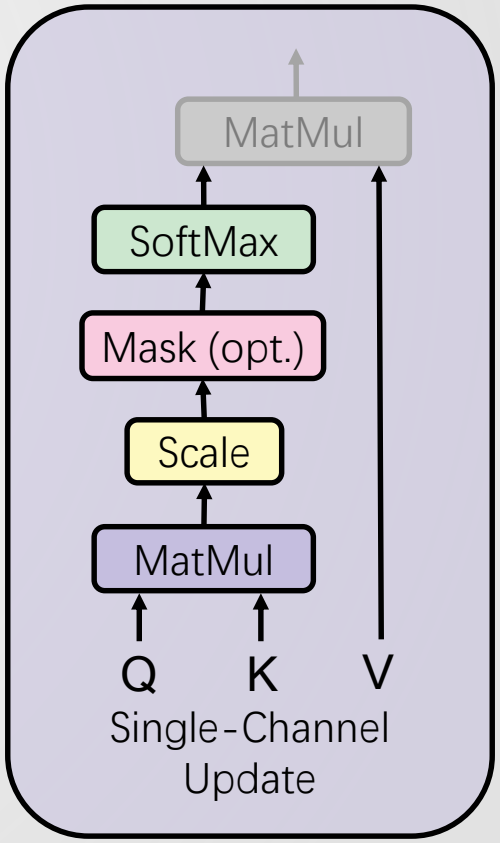




Single-Channel Update

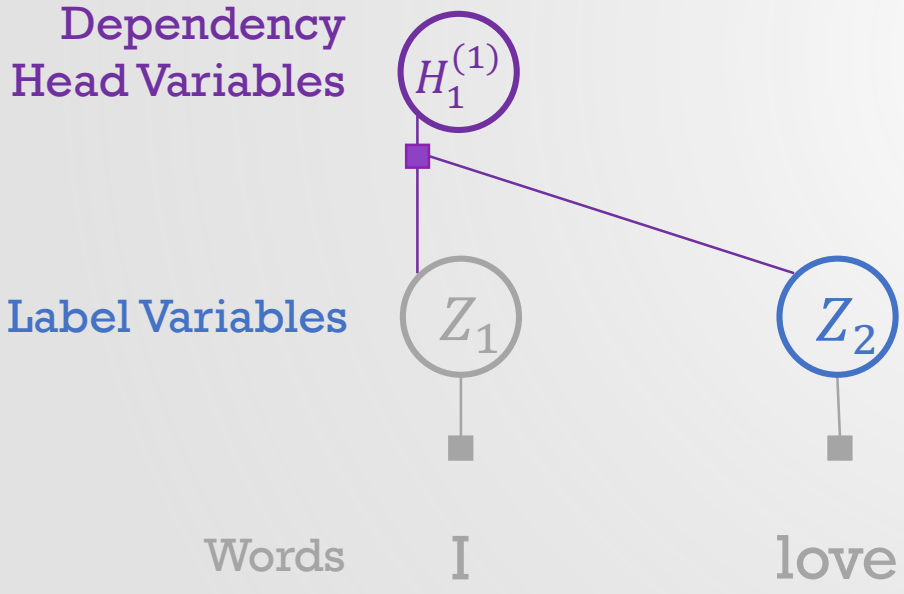
Part 2. The Computation Graph

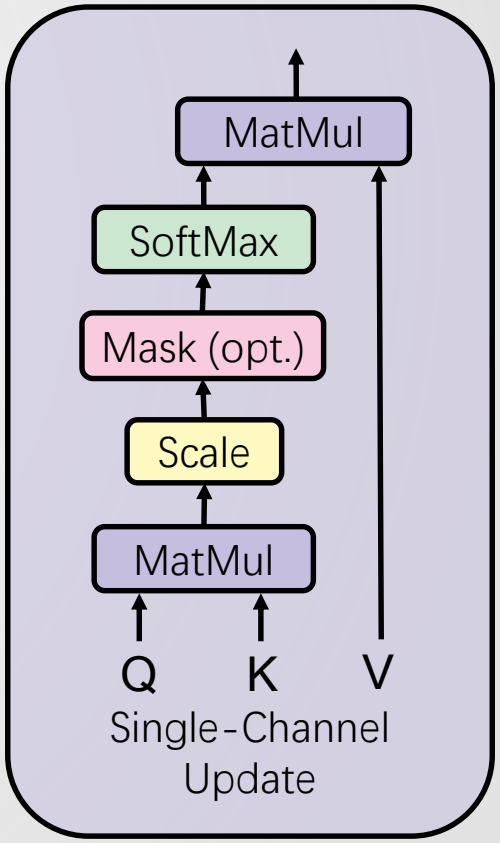




Single-Channel Update

Part 2. The Computation Graph

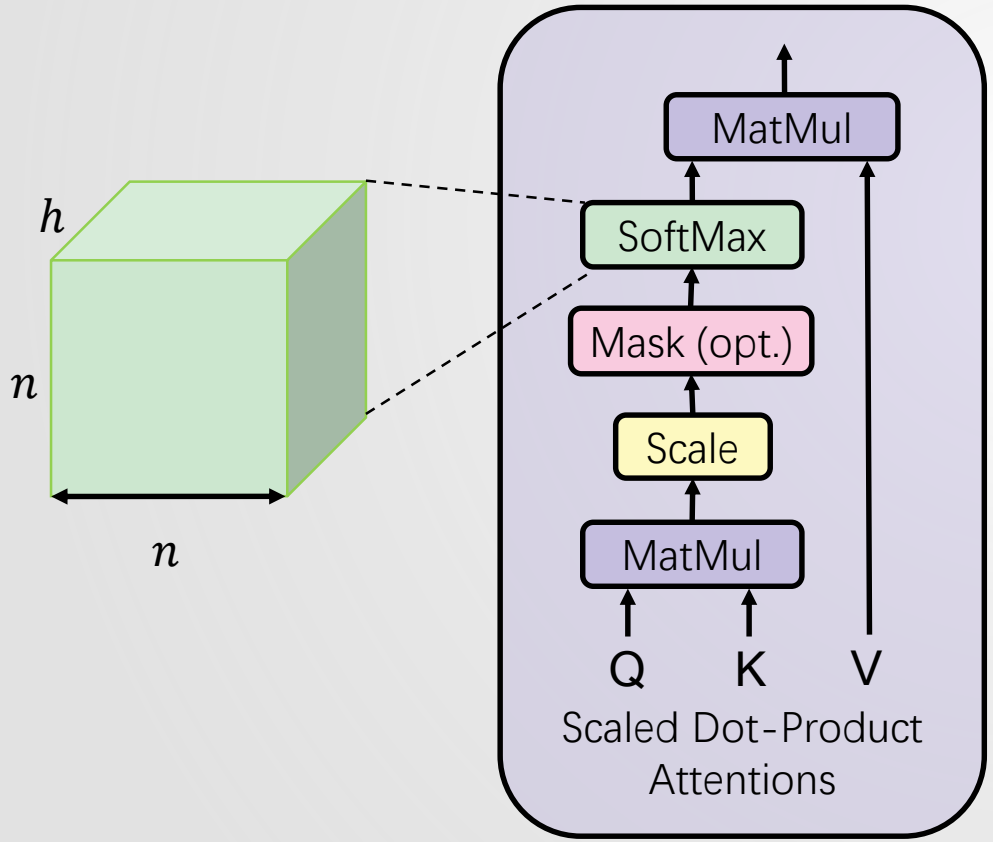




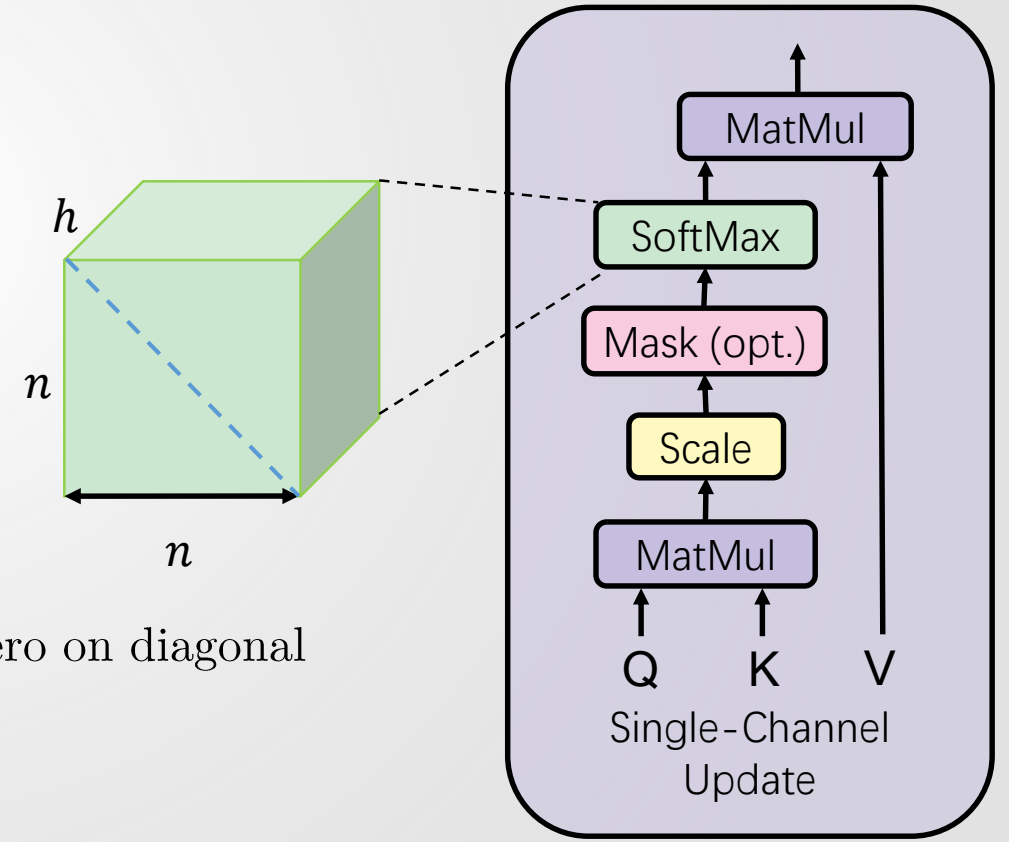
Single-Channel Update

Part 2. The Computation Graph

Transformers



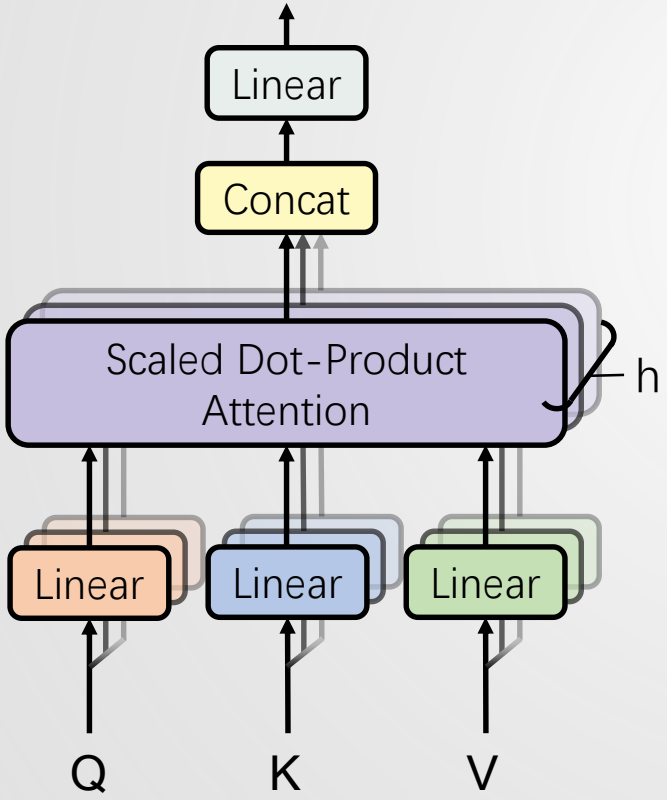
Probabilistic Transformers



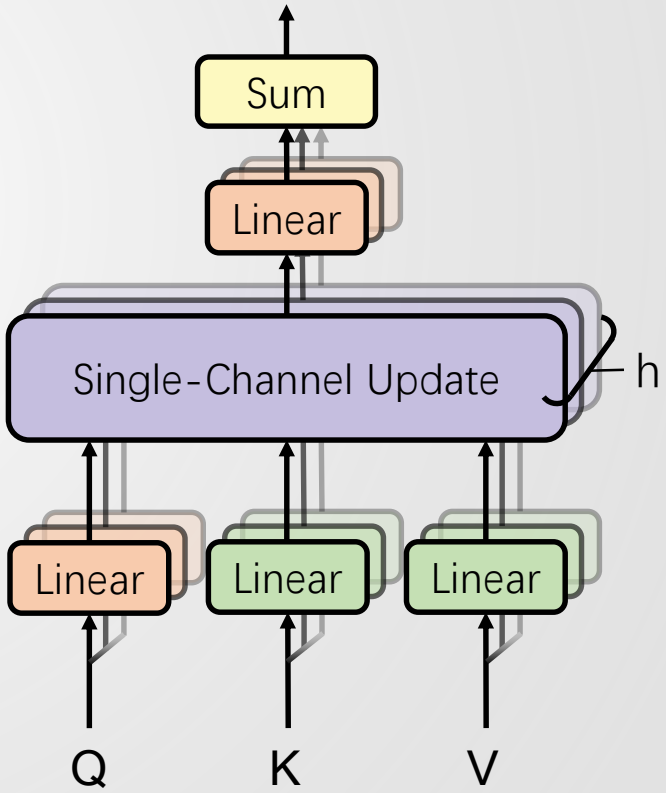
Multi-Channel Update

Part 2. The Computation Graph

Transformers



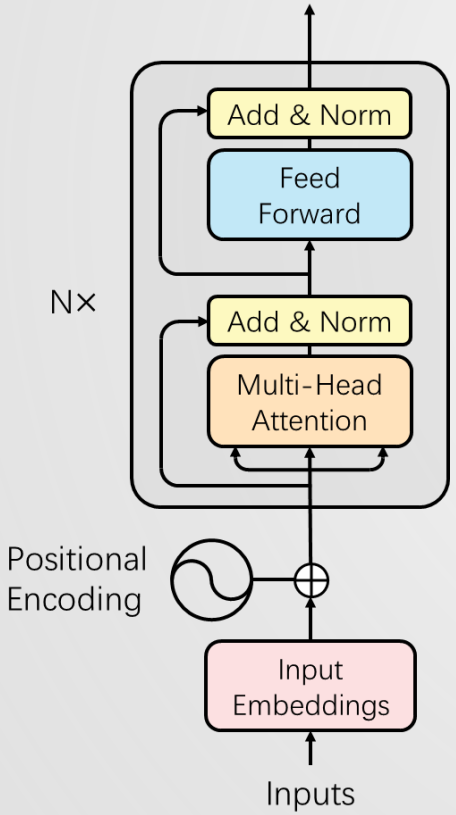
Probabilistic Transformers



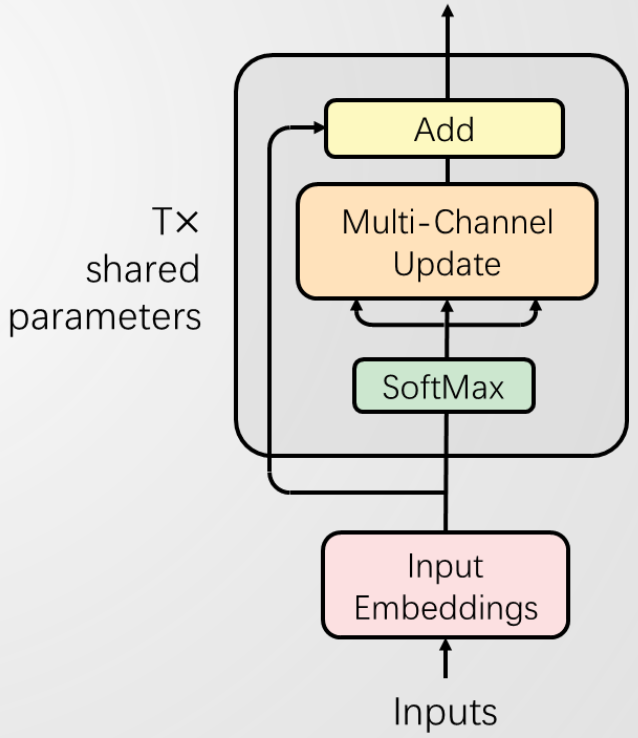
The Full Model

Part 2. The Computation Graph

Transformers



Probabilistic Transformers



What's More…

Probabilistic Transformer: A Probabilistic Dependency Model for Contextual Word Representation

- The variants of PTs also have interesting correspondences to the variants of transformers
  - RealFormer
  - All-attention Layer
  - …
- Performance is competitive to transformers
  - on small to medium sized datasets

Thanks!

Probabilistic Transformer:
A Probabilistic Dependency Model for Contextual Word Representation


wuhy1@shanghaitech.edu.cn


github.com/whyNLP/ Probabilistic-Transformer


faculty.sist.shanghaitech.edu.cn/faculty/tukw/

上海科技大学
ShanghaiTech University