GLoMo: Unsupervised Learning of Transferable Relational Graphs

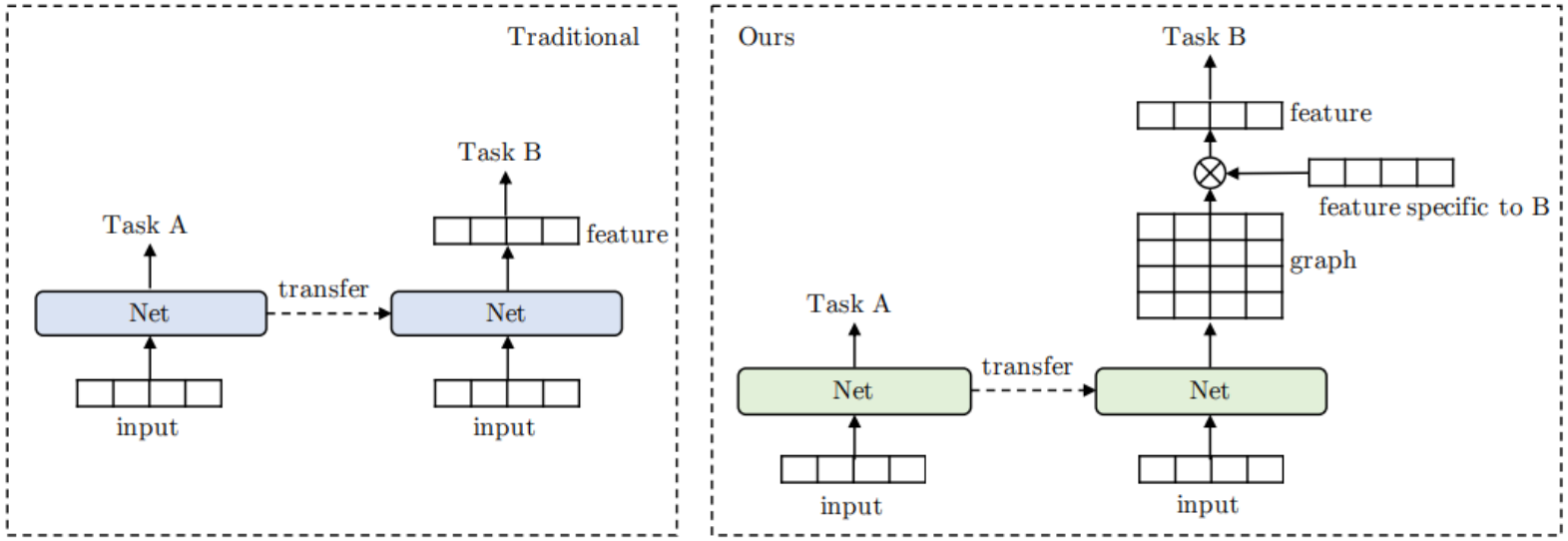
Tae Yeon Kim

Seoul National University, Department of Mechanical and Aerospace engineering

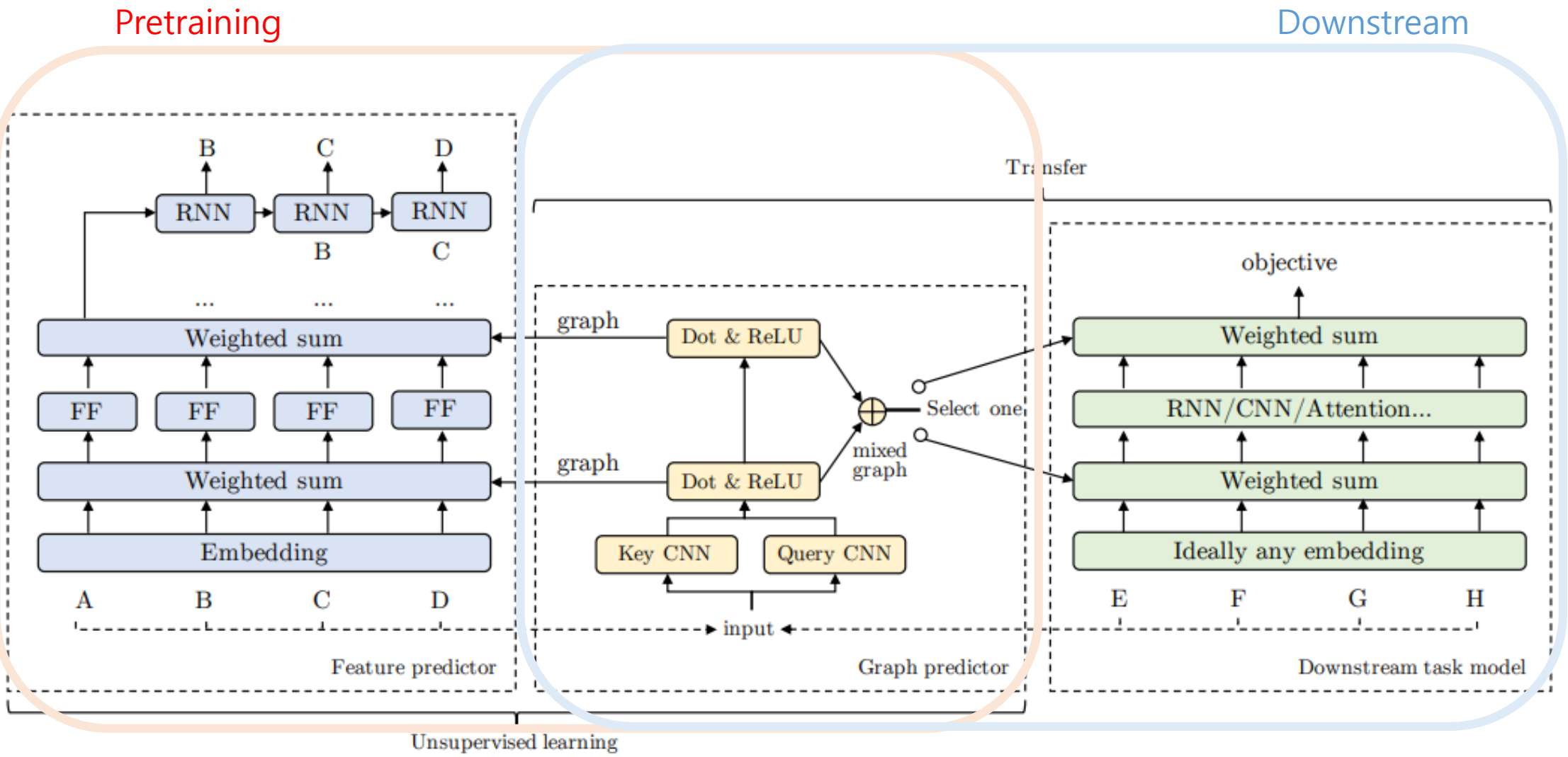
Zhilin Yang*[1], Jake (Junbo) Zhao*[23], Bhuwan Dhingra[1]
Kaiming He[3], William W. Cohen[1], Ruslan Salakhutdinov[1], Yann LeCun[23]

Summary of GLoMo

- Break away from the standardized norm of feature based deep transfer learning.
- Learn versatile structures in the data with data driven approach.
- Decouple features and graphs
- GLoMo learns generic latent relational graphs between pairs of data units from large scale data.
- Transfer the relational graphs to downstream tasks.



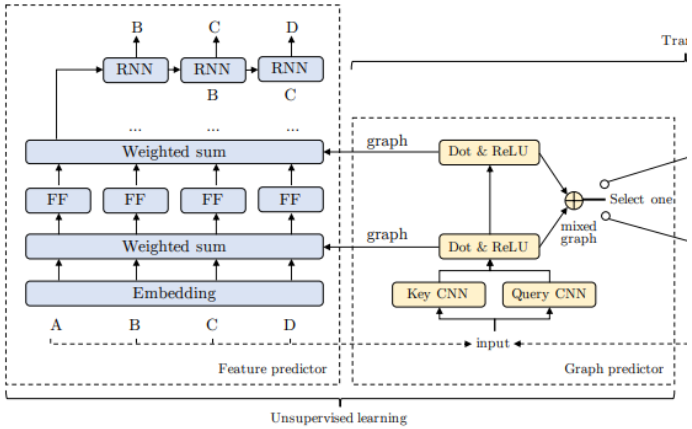
Summary of GLoMo



Example of GLoMo

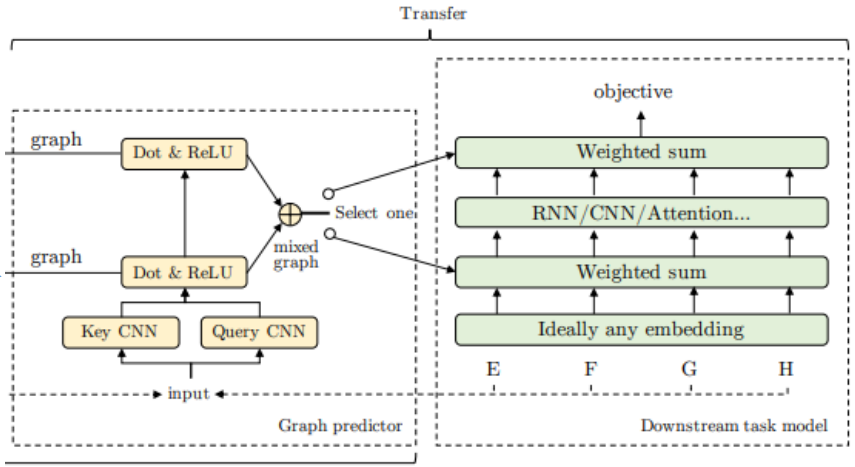
Over 700 million tokens

training

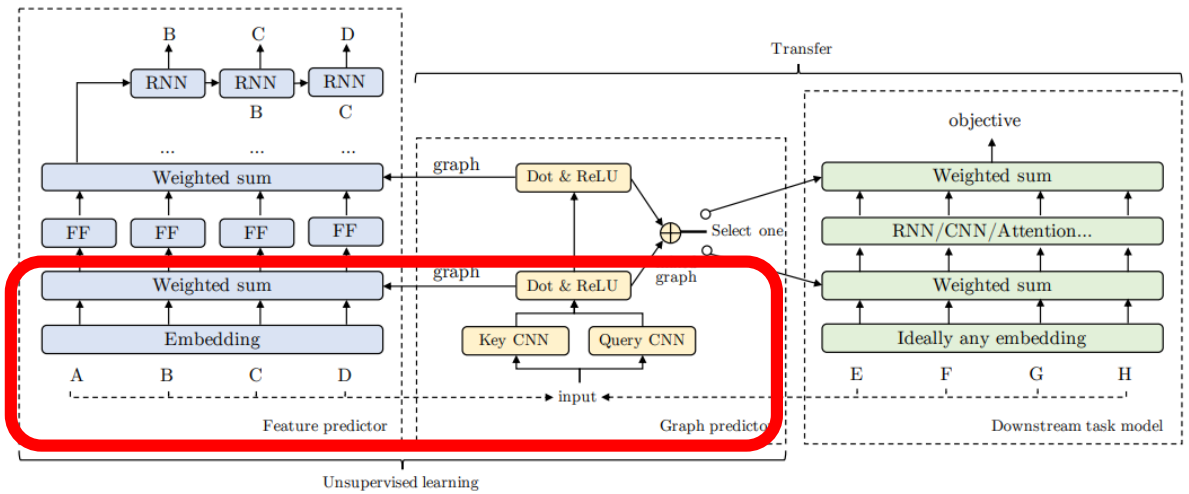


Transfer graph predictor

Downstream task



Pretraining part



Input: sequence of units $\boldsymbol{x} = (\boldsymbol{x_1}, \cdots, \boldsymbol{x_T})$

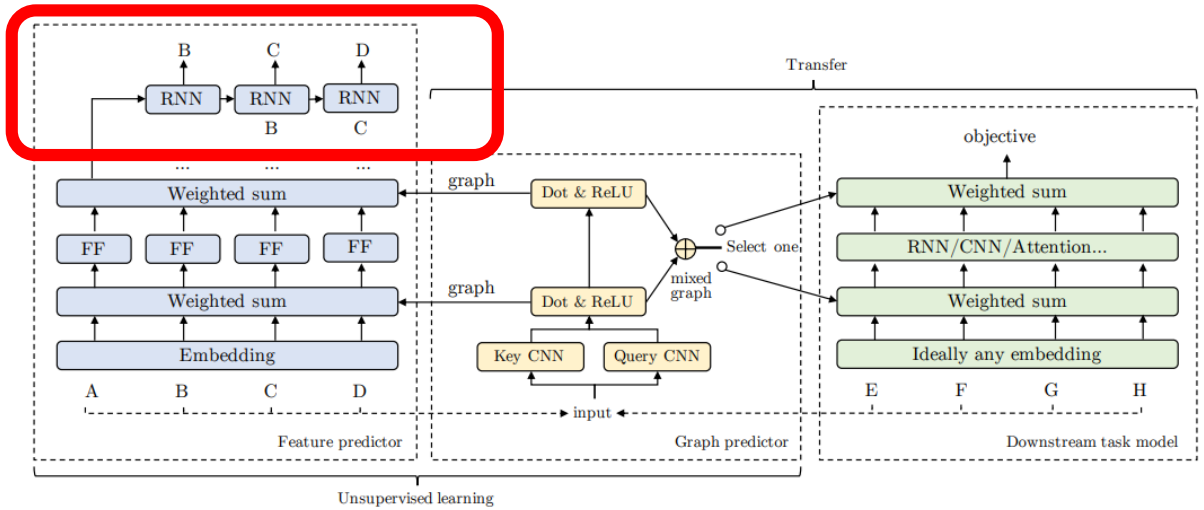
Key CNN output k

Query CNN output q

Learning affinity matrix **G** $\boldsymbol{G_{ij}} = \frac{(\boldsymbol{ReLU}(\boldsymbol{k_i q_j + b}))^2}{\sum_{\boldsymbol{i'}}(\boldsymbol{ReLU}(\boldsymbol{k_{i'} q_j + b}))^2}$

After graph prediction, combine unary features of previous layers and affinity matrix $f_t^l = v(\sum_j G_{jt}^l f_j^{l-1}, f_t^{l-1})$

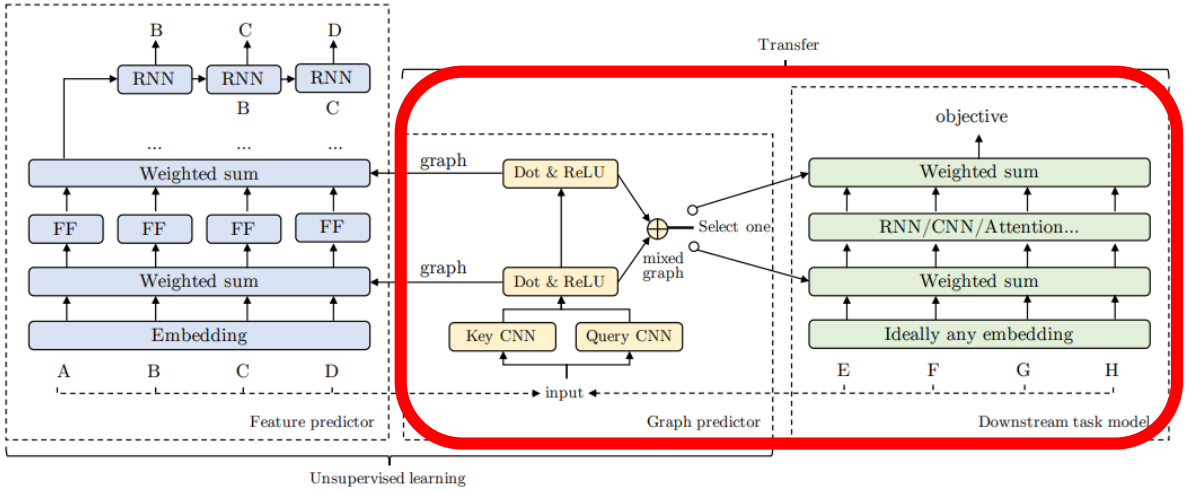
Pretraining part



The objective is context prediction with context length D at time t

$$\max \sum_{t} logP(x_{t+1}, \cdots, x_{t+D} | x_t, f_t^L)$$

Downstream part



-After training graph predictor from large-scale datasets, it is transferred to downstream tasks.

-Various embedding(GloVe, ELMo) can be used feature predictor of downstream tasks.

-Let the features of downstream tasks $H = (h_1, \cdots, h_T)$

-To consider propagating the connections among multiple layers $\Lambda^l = \prod_{i=1}^{l} G^i$

Results

Table 1: Main results on natural language datasets. Self-attention modules are included in all baseline models. All baseline methods are feature-based transfer learning methods, including ELMo and GloVe. Our methods combine graph-based transfer with feature-based transfer. Our graphs operate on various sets of features, including GloVe embeddings, ELMo embeddings, and RNN states. "mism." refers to the "mismatched" setting.

Transfer method	SQuAD GloVe		SQuAD ELMo		IMDB GloVe	MNLI GloVe	
	EM	*F1*	*EM*	*F1*	*Accuracy*	*matched*	*mism.*
transfer feature only (baseline)	69.33	78.73	74.75	82.95	88.51	77.14	77.40
GLoMo on embeddings	70.84	79.90	**76.00**	**84.13**	**89.16**	**78.32**	**78.00**
GLoMo on RNN states	**71.30**	**80.24**	76.20	83.99	-	-	-

-'Self attention' module is incorporated into all of baselines models.

-When pre-trained 'feature transfer' is adopted, GloMo is able to yield further improvement.

-GLoMo can be adopted to various embedding and RNN states

Thank you