NIPS 2018 paper:

GLoMo: Unsupervised Learning of Transferable Relational Graphs

Author: Zhilin Yang, Jake Zhao, Bhuwan Dhingra,
Kaiming He, William W.Cohen,
RuslanSalakhutdinov, Yann LeCun

Carnegie Mellon University
New York University
Facebook AI Research

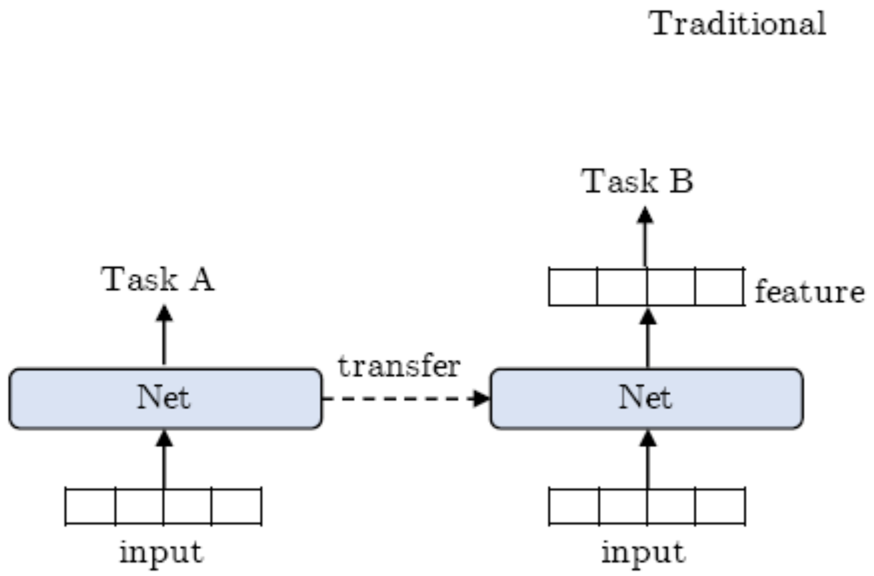
Qing Lyu 2/6/2019

Contribution

- Present a novel transfer learning scheme based on latent relational graph learning
- This framework is capable of improving performance and learning generic graphs applicable to various types of features

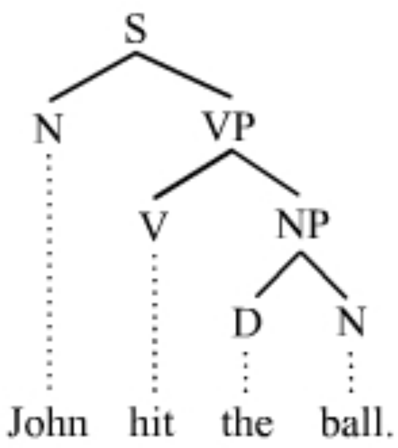
Current CNNs & RNNs

- Primarily operate on grid-like or sequential structures due to their built-in “innate priors”
- Rely on high expressiveness to model complex structural phenomena
- Do not explicitly leverage structural, graphical representations

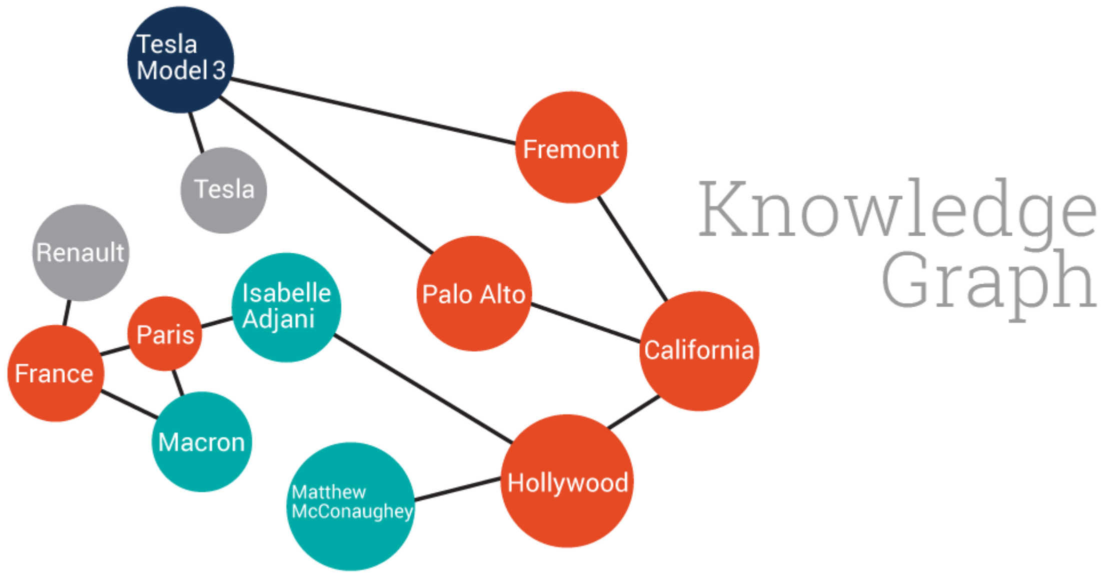


Relational graph structures

- parse trees to represent syntactic dependency between words
- information retrieval systems exploit knowledge graphs to reflect entity relations

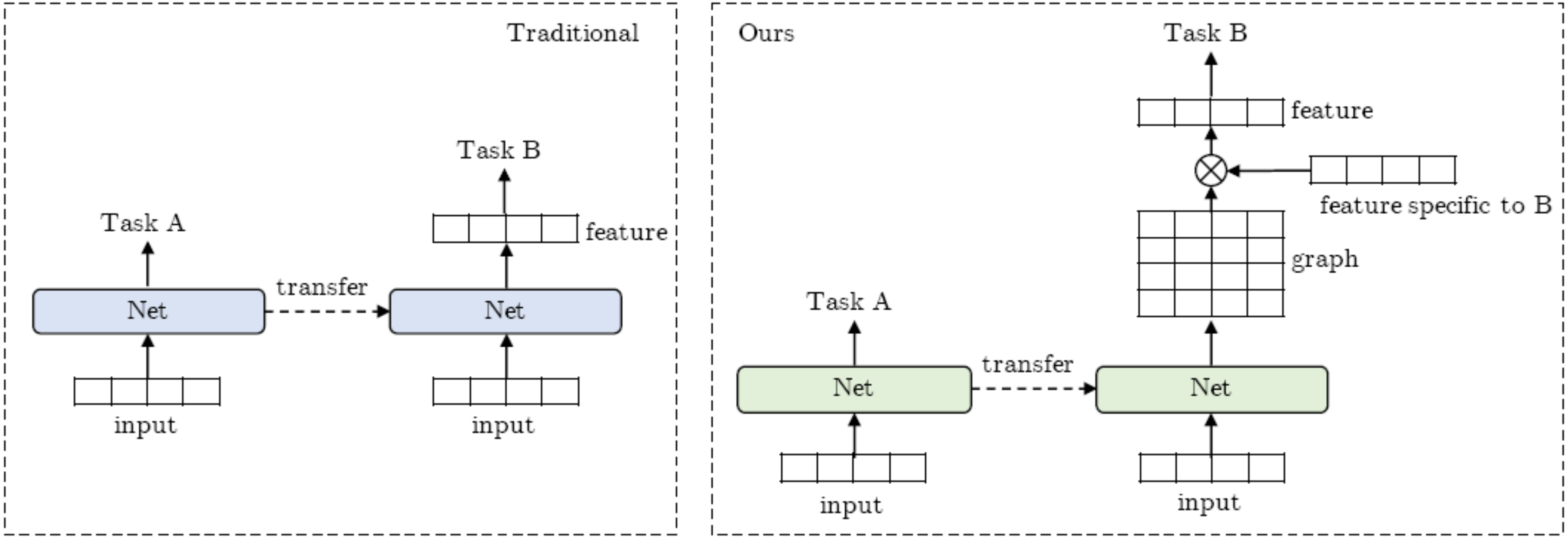


Constituency-based parse tree

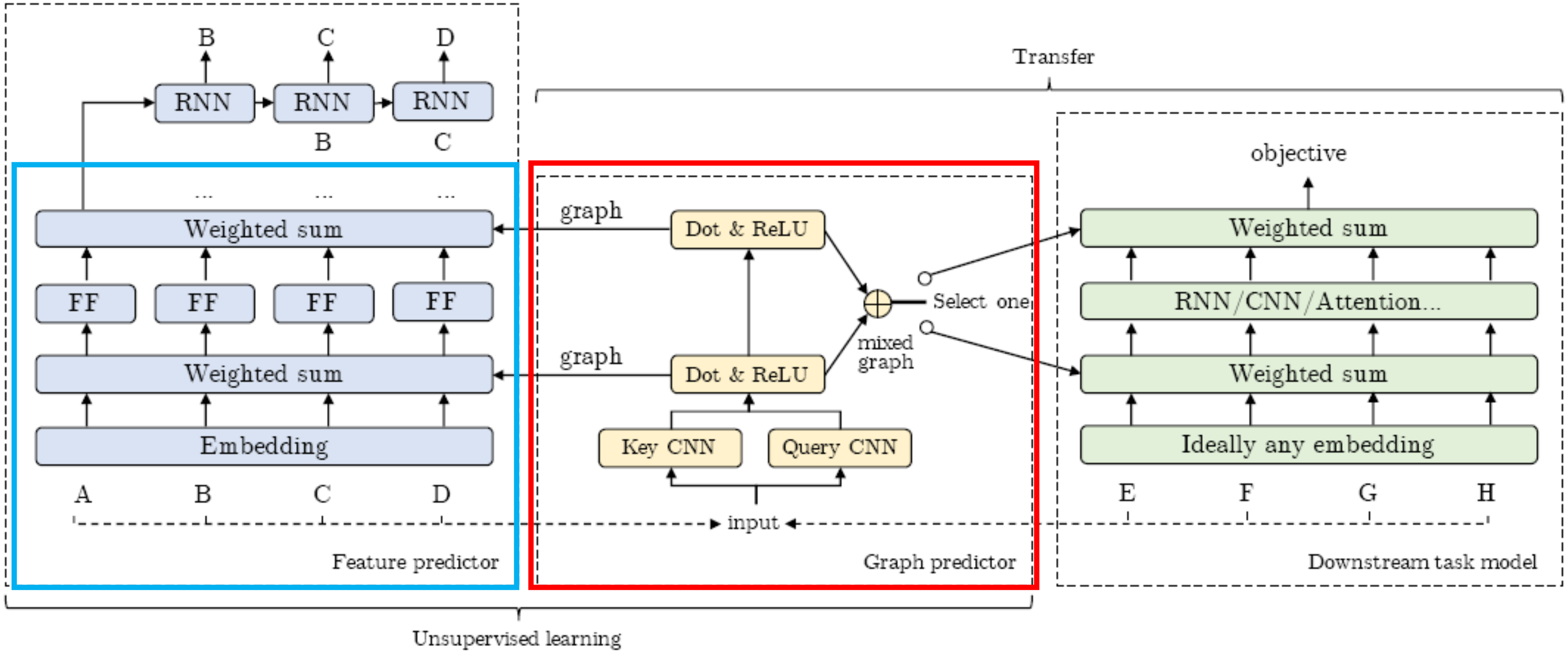


- These exemplified structures are universally present in almost any natural language data regardless of the target tasks
- Suggesting the possibility of transfer across tasks
- For vision domain, modeling the relations between pixels is proven useful

“we are interested in learning transferable latent relational graphs, where the nodes of a latent graph are the input units, e.g., all the words in a sentence.”

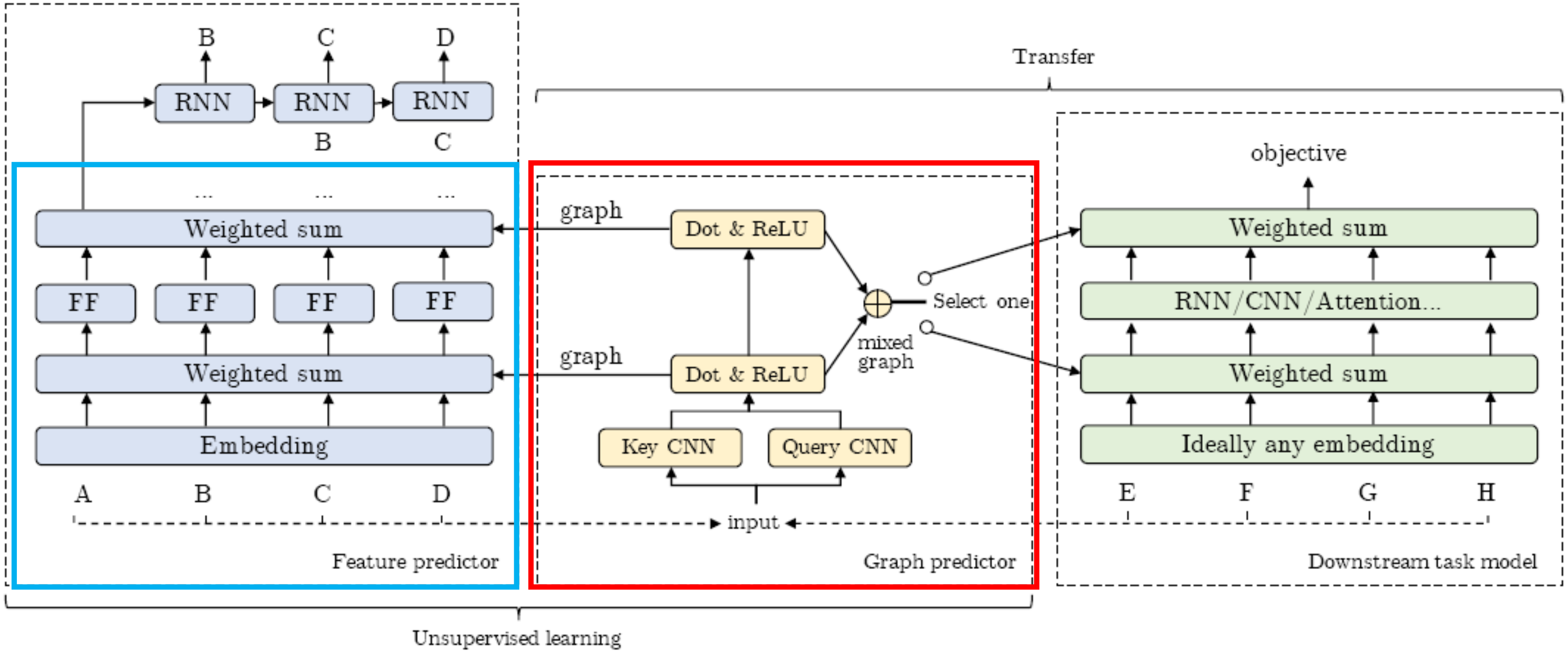

Traditional
Task A
Task B
feature
transfer
Net
Net
input
input
Ours
Task B
feature
feature specific to B
graph
Task A
transfer
Net
Net
input
input

GLoMO(Graphs from LOw-level unit MOdeling)



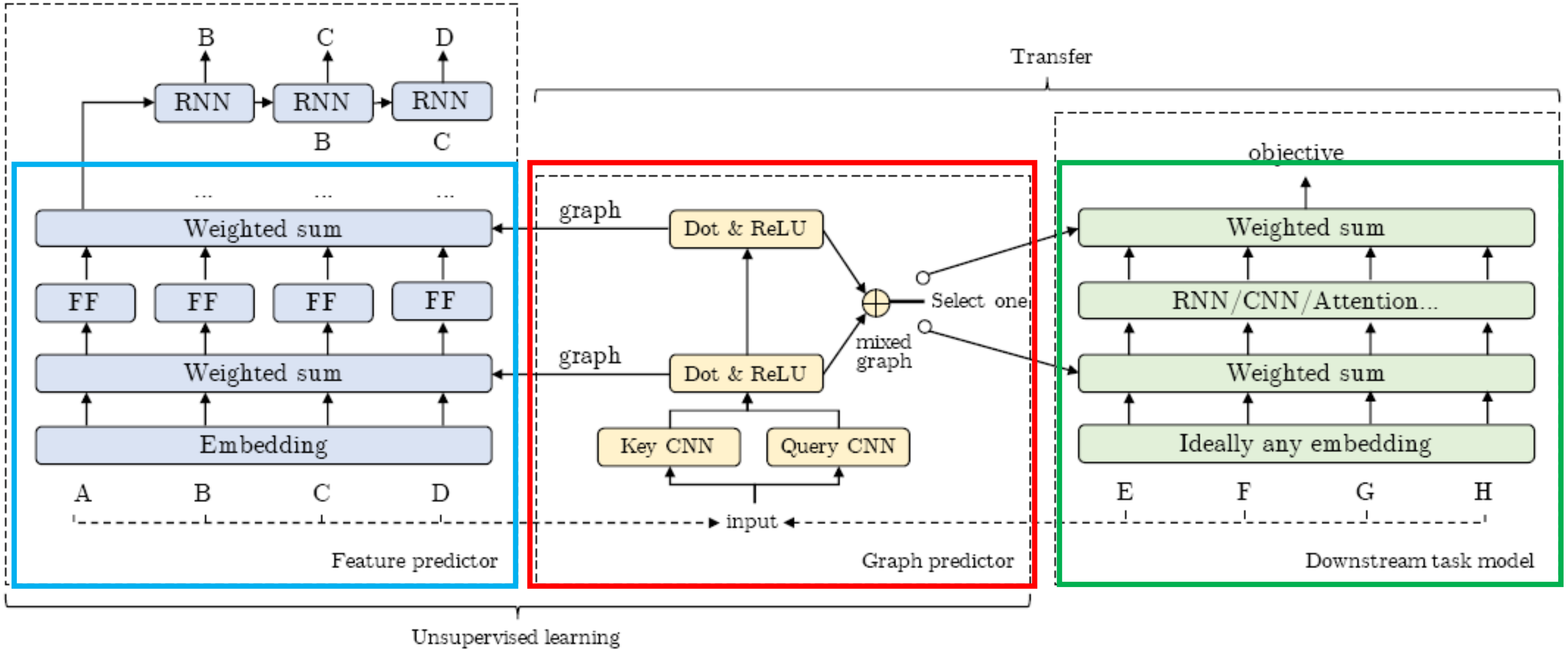
$$\mathbf{f}_t^l = v\Big(\sum_j G^l_{jt}\mathbf{f}_j^{l-1}, \mathbf{f}_t^{l-1}\Big) \qquad G^l_{ij} = \frac{\left(\mathrm{ReLU}(\mathbf{k}_i^{l\top}\mathbf{q}_j^l + b)\right)^2}{\sum_{i'}\left(\mathrm{ReLU}(\mathbf{k}_{i'}^{l\top}\mathbf{q}_j^l + b)\right)^2}$$

GLoMO(Graphs from LOw-level unit MOdeling)



Objective Function $\max \sum_{t} \log P(x_{t+1}, \cdots, x_{t+D} | x_t, \mathbf{f}_t^L)$

GLoMO(Graphs from LOw-level unit MOdeling)



$$\mathbf{f}_t^l = v\left(\sum_j G_{jt}^l \mathbf{f}_j^{l-1}, \mathbf{f}_t^{l-1}\right) \qquad G_{ij}^l = \frac{\left(\mathrm{ReLU}(\mathbf{k}_i^{l\top}\mathbf{q}_j^l + b)\right)^2}{\sum_{i'}\left(\mathrm{ReLU}(\mathbf{k}_{i'}^{l\top}\mathbf{q}_j^l + b)\right)^2} \qquad \mathbf{M} = \sum_{l=1}^{L} m_G^l \mathbf{G}^l + \sum_{l=1}^{L} m_\Lambda^l \boldsymbol{\Lambda}^l, \text{ s.t. } \sum_{l=1}^{L}(m_G^l + m_\Lambda^l) = 1$$

Experiments

- Transfer Setting
  - preprocessed the Wikipedia dump and obtained a corpus of over 700 million tokens
- Question Answering
  - The Stanford question answering dataset (SQuAD)
    - 100,000+ question-answer pairs from 500+ Wikipedia articles
- Natural Language Inference
  - Multi-Genre NLI corpus (MNLI)
    - 433k sentence pairs annotated with textual entailment information
- Sentiment Analysis
  - movie review dataset
    - 25,000 training and 25,000 testing samples

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.

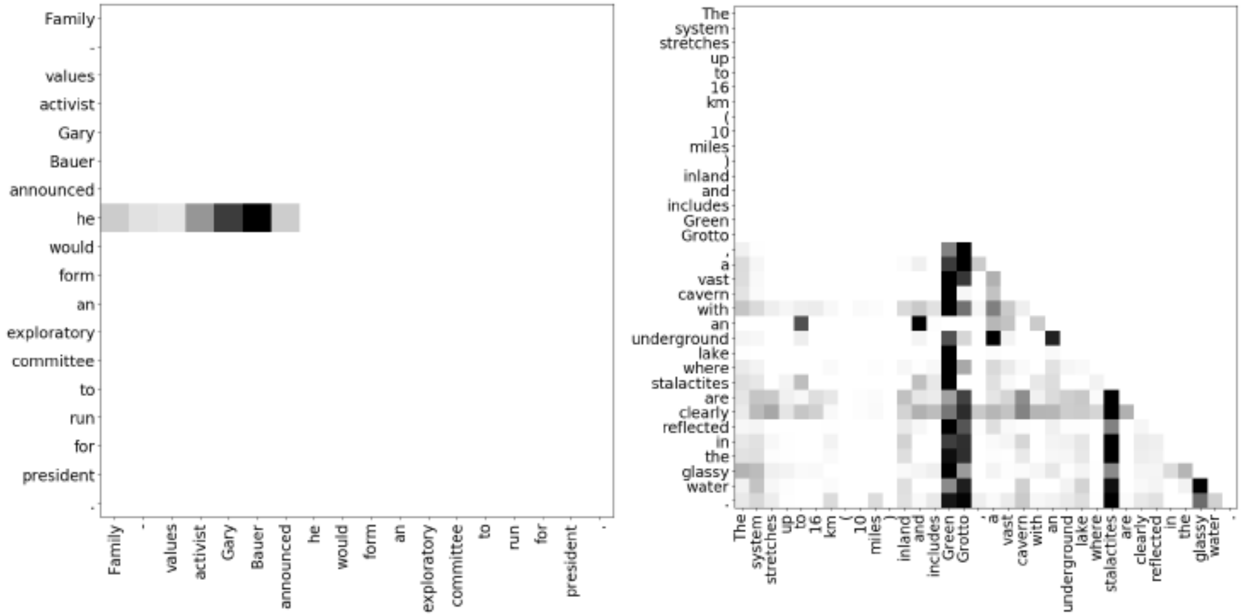
Transfer method	SQuAD GloVe		SQuAD ELMo		IMDB GloVe	MNLI GloVe	
	EM	*F1*	*EM*	*F1*	*Accuracy*	*matched*	*mism.*
transfer feature only (baseline)	69.33	78.73	74.75	82.95	88.51	77.14	77.40
GLoMo on embeddings	70.84	79.90	**76.00**	**84.13**	**89.16**	**78.32**	**78.00**
GLoMo on RNN states	**71.30**	**80.24**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

Results

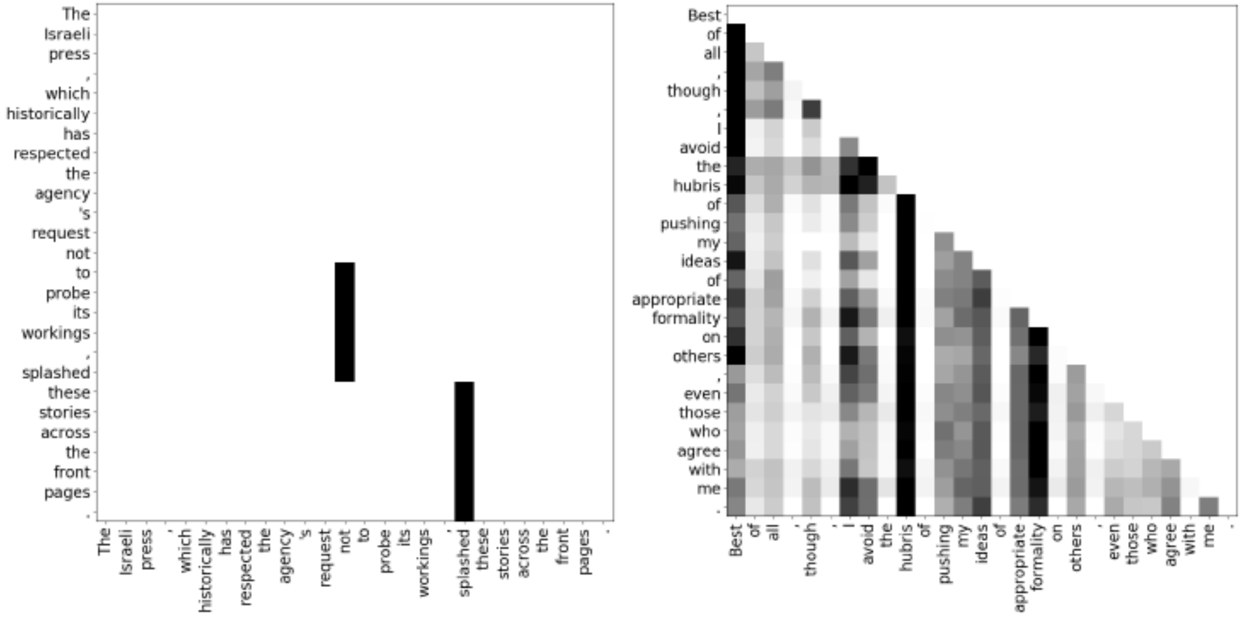
Table 2: Ablation study.

Method	SQuAD GloVe		SQuAD ELMo		IMDB GloVe	MNLI GloVe	
	EM	*F1*	*EM*	*F1*	*Accuracy*	*matched*	*mism.*
GLoMo	**70.84**	**79.90**	**76.00**	**84.13**	**89.16**	**78.32**	78.00
- decouple	70.45	79.56	75.89	83.79	-	-	-
- sparse	70.13	79.34	75.61	83.89	88.96	78.07	77.75
- hierarchical	69.92	79.23	75.70	83.72	88.71	77.87	77.85
- unit-level	69.23	78.66	74.84	83.37	88.49	77.58	**78.05**
- sequence	69.92	79.29	75.50	83.70	88.96	78.11	77.76
uniform graph	69.48	78.82	75.14	83.28	88.57	77.26	77.50

Results



(a) Related to coreference resolution. (b) Attending to objects for modeling long-term dependency.



(c) Attending to negative words and predicates. (d) Attending to nouns, verbs, and adjectives for topic modeling.

Figure 3: Visualization of the graphs on the MNLI dataset. The graph predictor has not been trained on MNLI. The words on the y-axis "attend" to the words on the a-axis; i.e., each row sums to 1.

Results

Figure 4: Visualization. Left: a shark image as the input. Middle: weights of the edges connected with the *central* pixel, organized into 24 heads (3 layers with 8 heads each). Right: weights of the edges connected with the *bottom-right* pixel. Note the use of masking.

Method / Base-model	ResNet-18	ResNet-34
baseline	90.93±0.33	91.42±0.17
GLoMo	**91.55±0.23**	**91.70±0.09**
ablation: uniform graph	91.07±0.24	-

Table 3: CIFAR-10 classification results. We adopt a 42,000/8,000 train/validation split—once the best model is selected according to the validation error, we directly forward it to the test set without doing any validation set place-back retraining. We only used horizontal flipping for data augmentation. The results are averaged from 5 rounds of experiments.