

Multi-Hop RC, HotpotQA & GNNs

Select, Answer and Explain: Interpretable Multi-hop Reading
Comprehension over Multiple Documents – Tu et al., AACL 2020

Presented By:
Lovish Madaan

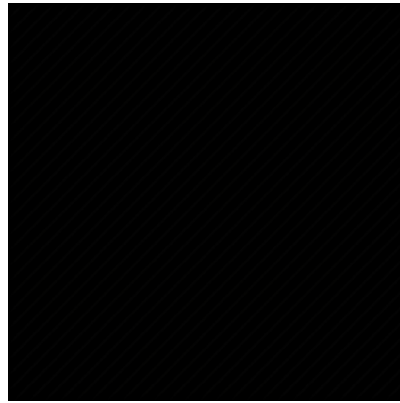
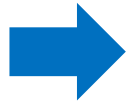
References

- HotpotQA - Peng Qi (Stanford)
- GNNs - Jure Leskovec (Stanford), AAAI 2019 Tutorial by William Hamilton (McGill)
- Some elements and images borrowed from Tu et al. (AAAI 2020), Yang et al. (EMNLP 2018), and Jay Alammarr

Topics

- Introduction and HotpotQA
- Select, Answer and Explain
- GNNs
- Answer and Explain
- Results and Ablation Study
- Reviews

The Promise of Question Answering



In which city was
Facebook first
launched?

Cambridge, Massachusetts.

This is because Mark Zuckerberg and his business partners launched it from his Harvard dormitory [1], and Harvard is located in Cambridge, Massachusetts [2].

[1] https://en.wikipedia.org/wiki/Mark_Zuckerberg

[2] https://en.wikipedia.org/wiki/Harvard_University



The ~~Reality~~ of Question Answering

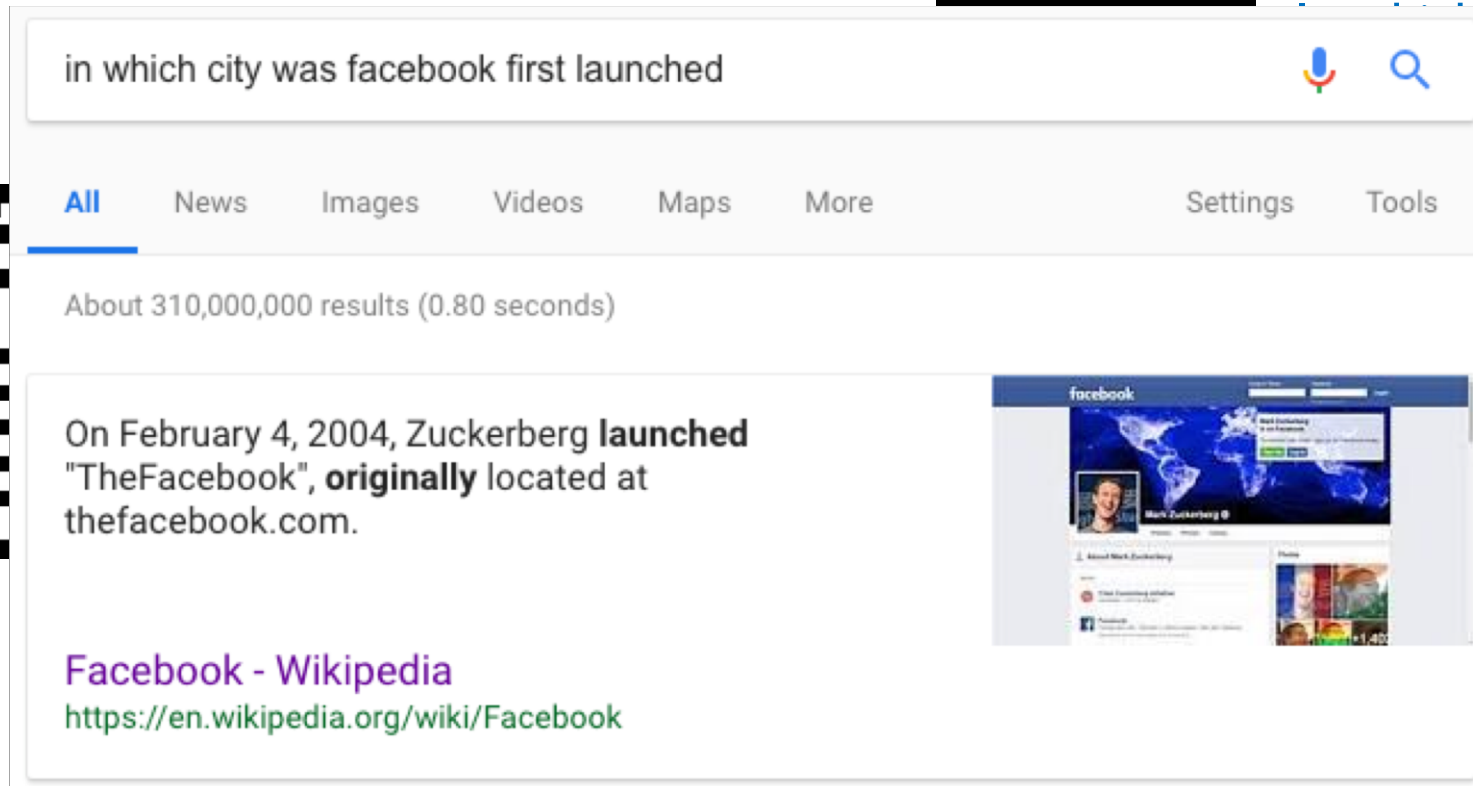
in which city was facebook first launched

All News Images Videos Maps More Settings Tools

About 310,000,000 results (0.80 seconds)

On February 4, 2004, Zuckerberg **launched** "TheFacebook", **originally** located at thefacebook.com.

Facebook - Wikipedia
<https://en.wikipedia.org/wiki/Facebook>



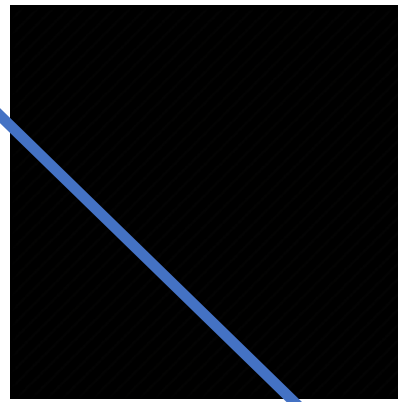
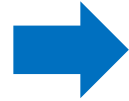
city was
first
?

ts.
Zuckerberg and
founded it from
Harvard], and
Cambridge,

- [1] https://en.wikipedia.org/wiki/Mark_Zuckerberg
- [2] https://en.wikipedia.org/wiki/Harvard_University
Sorry, folks from Google!

The Promise of Question Answering

Multi-hop reasoning



In which city was
Facebook first
launched?

Cambridge, Massachusetts.

This is because Mark Zuckerberg and his business partners launched it from his Harvard dormitory [1], and Harvard is located in Cambridge, Massachusetts [2].

[1] https://en.wikipedia.org/wiki/Mark_Zuckerberg

[2] https://en.wikipedia.org/wiki/Harvard_University



The Promise of Question Answering

Multi-hop reasoning



Text-based, diverse

In which city was Facebook first launched?

Cambridge, Massachusetts.

This is because Mark Zuckerberg and his business partners launched it from his Harvard dormitory [1], and Harvard is located in Cambridge, Massachusetts [2].

[1] https://en.wikipedia.org/wiki/Mark_Zuckerberg

[2] https://en.wikipedia.org/wiki/Harvard_University

The Promise of Question Answering

Multi-hop reasoning



Text-based, diverse

Explainability

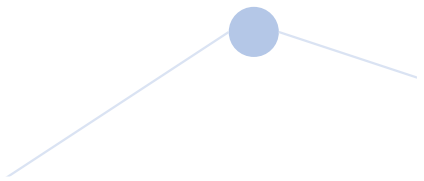
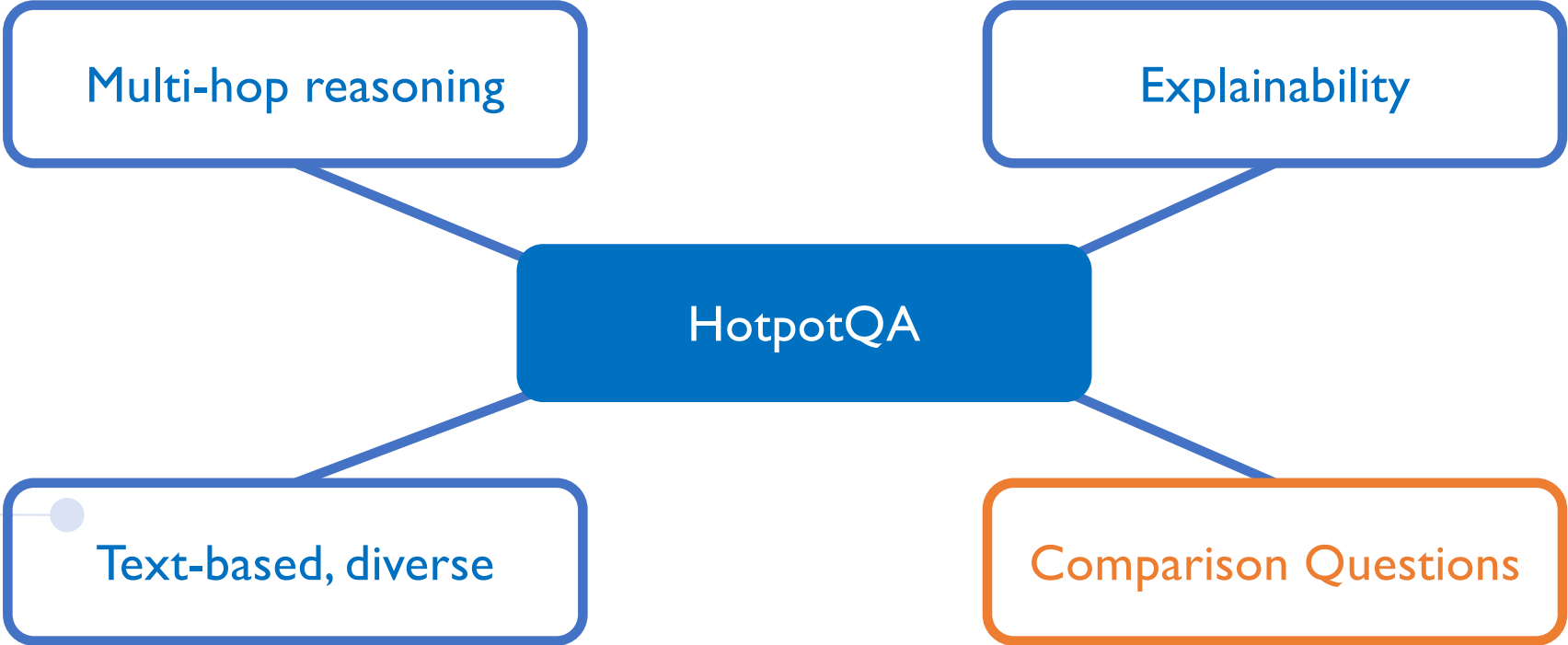
launched?

Cambridge, Massachusetts.

This is because Mark Zuckerberg and his business partners launched it from his Harvard dormitory [1], and Harvard is located in Cambridge, Massachusetts [2].

[1] https://en.wikipedia.org/wiki/Mark_Zuckerberg

[2] https://en.wikipedia.org/wiki/Harvard_University



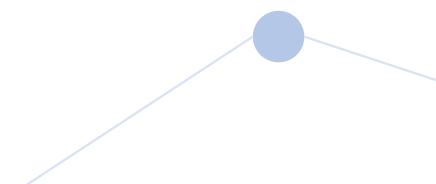
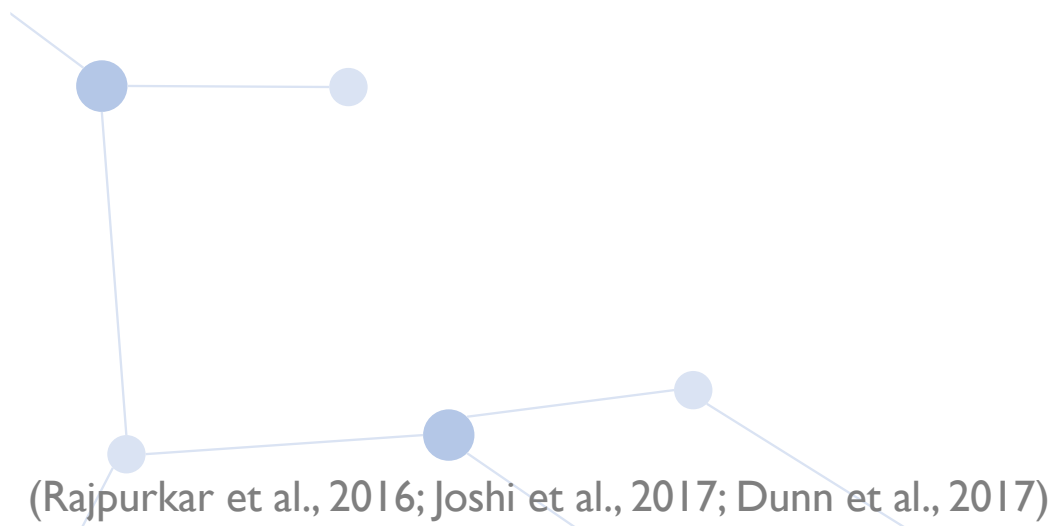
Multi-hop Reasoning across Multiple Documents

- Previous work (SQuAD, TriviaQA, etc)

- HotpotQA

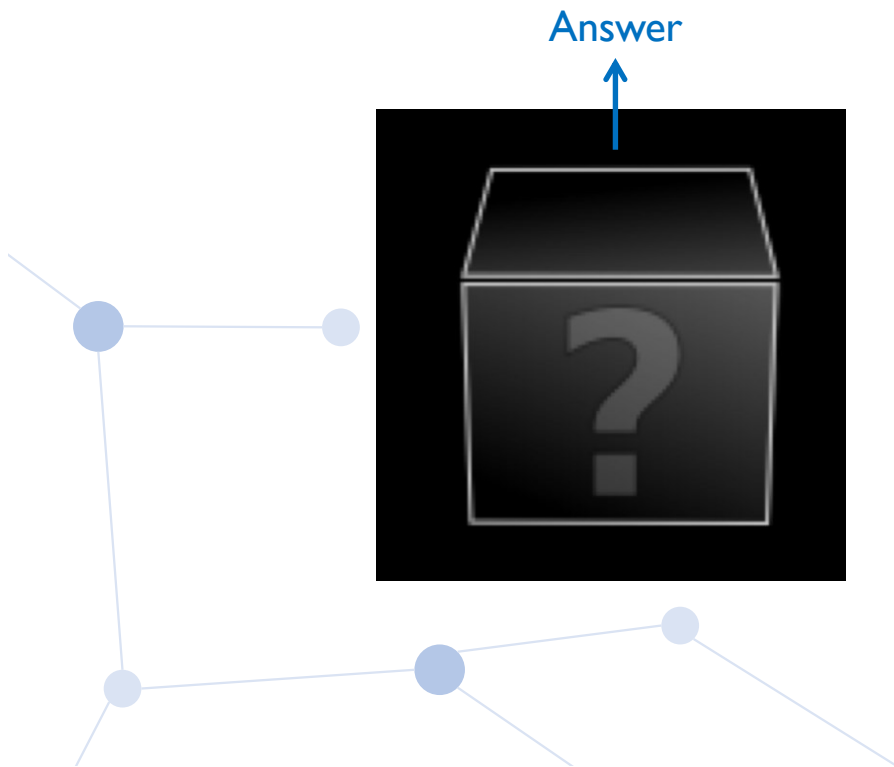
When was *Chris Martin* born?

When was the lead singer of *Coldplay* born?

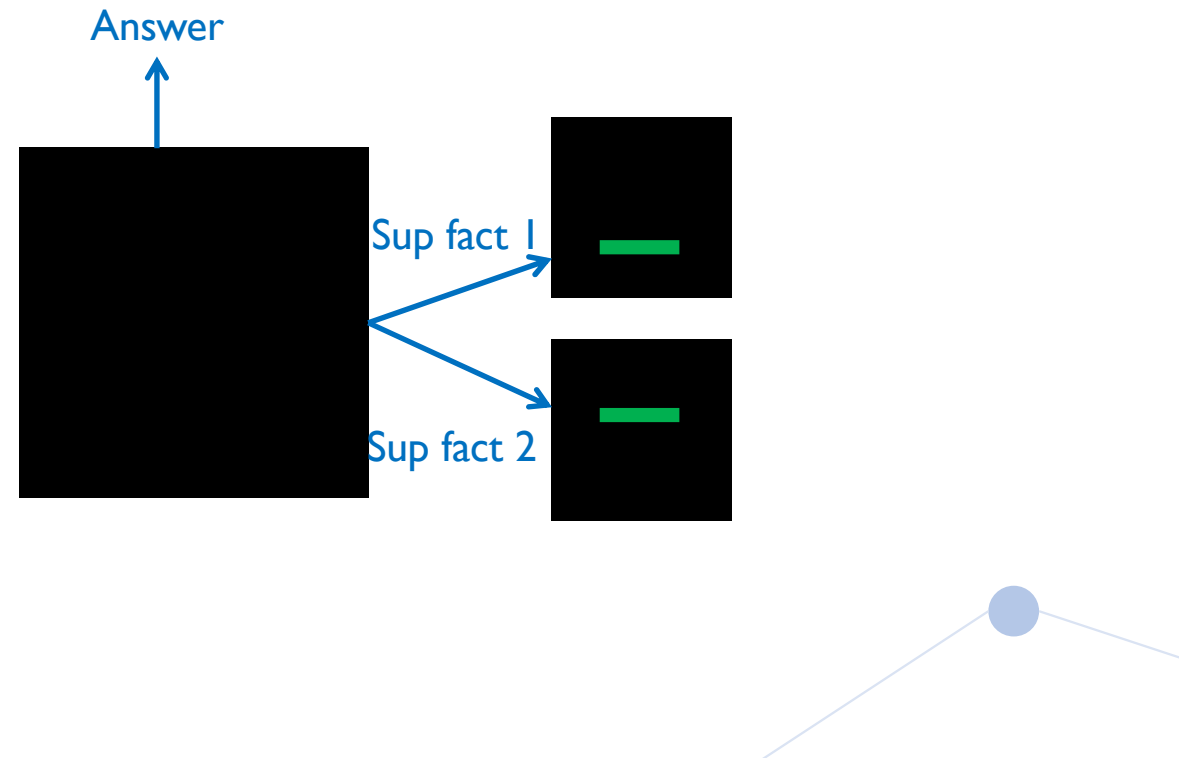


Explainability

- Previous work



- HotpotQA



Evaluation Settings

- Distractor Setting
 - 2 gold paragraphs + 8 extracted from information retrieval
- Fullwiki Setting
 - Entire Wikipedia as context

- Types of Instances
 - Bridge Entity Questions
 - Comparison Questions

Reasoning Type	%	Example(s)
Inferring the <i>bridge entity</i> to complete the 2nd-hop question (Type I)	42	<p>Paragraph A: The 2015 Diamond Head Classic was a college basketball tournament ... <i>Buddy Hield</i> was named the tournament's MVP.</p> <p>Paragraph B: <i>Chavano Rainier "Buddy" Hield</i> is a Bahamian professional basketball player for the Sacramento Kings of the NBA...</p> <p>Q: Which team does the player named 2015 Diamond Head Classic's MVP play for?</p>
Comparing two entities (Comparison)	27	<p>Paragraph A: LostAlone were a British rock band ... consisted of <i>Steven Battelle, Alan Williamson, and Mark Gibson</i>...</p> <p>Paragraph B: Guster is an American alternative rock band ... Founding members <i>Adam Gardner, Ryan Miller, and Brian Rosenworcel</i> began...</p> <p>Q: Did LostAlone and Guster have the same number of members? (yes)</p>

Topics

- Introduction and HotpotQA
- **Select, Answer and Explain**
- GNNs
- Answer and Explain
- Results and Ablation Study
- Reviews

Multi-hop RC – Previous Works

- Adapt techniques from single-hop QA
- Use Graph Neural Networks (GNNs)
 - Cao et al., 2018 – Build entity graph and realize multi-hop reasoning

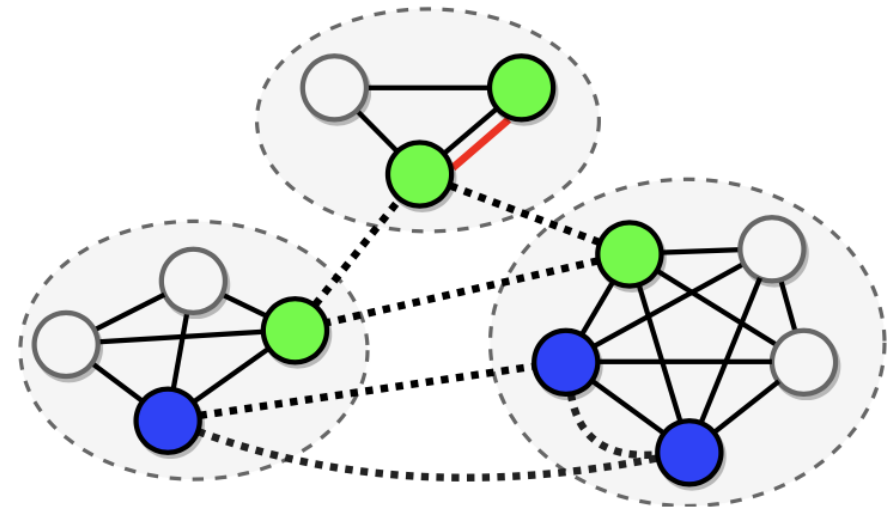


Figure 2: Supporting documents (dashed ellipses) organized as a graph where nodes are mentions of either candidate entities or query entities. Nodes with the same color indicates they refer to the same entity (exact match, coreference or both). Nodes are connected by three simple relations: one indicating co-occurrence in the same document (solid edges), another connecting mentions that exactly match (dashed edges), and a third one indicating a coreference (bold-red line).

Shortcomings – Previous Works

- Concatenate multiple documents / Process documents separately
 - No document filters
- Current application of GNNs
 - Entities as nodes – either pre specified / use NER
 - Further processing if answer is not an entity

Select, Answer and Explain (SAE)

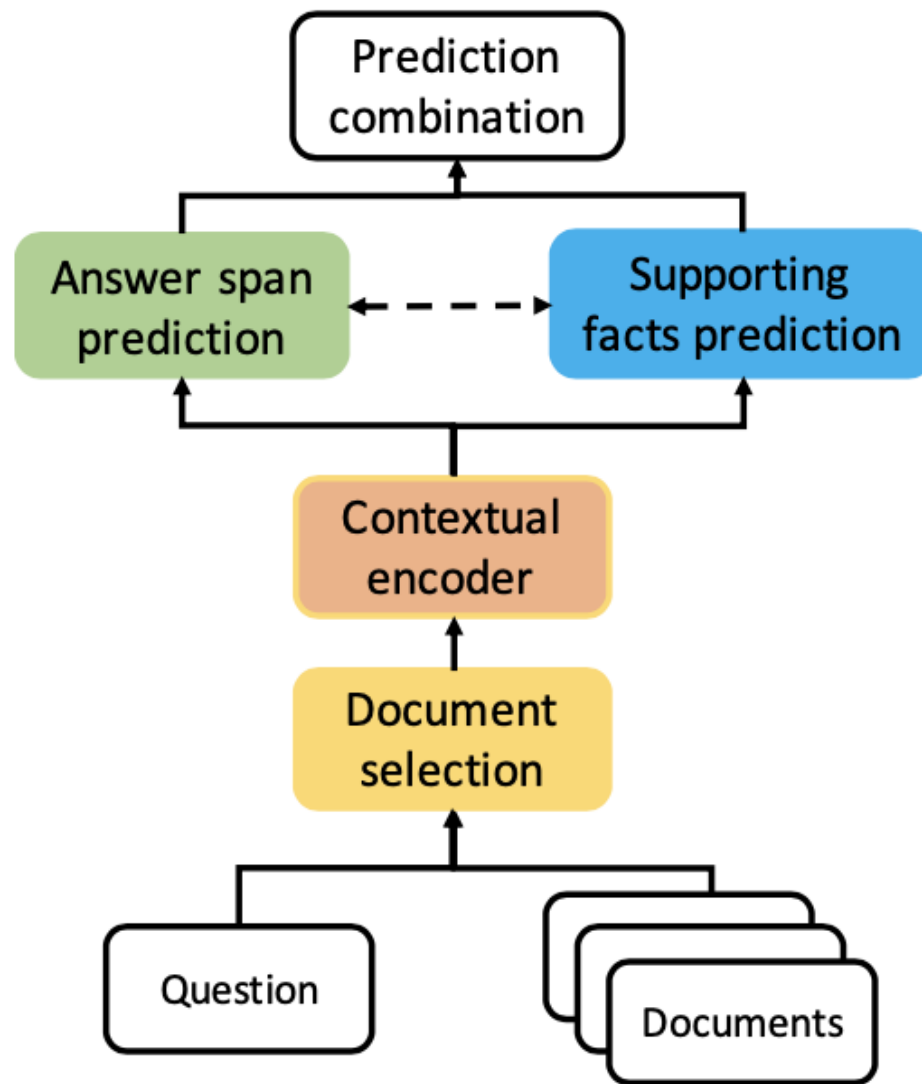


Figure 2: Diagram of the proposed SAE system. The dashed arrow line indicates the mixed attention based interaction between the two tasks

Preprocessing & Inputs

- Question and set of documents
- Answer text
- Set of labelled support sentences from each document
- Label corresponding to each document - D_i (0/1)
- Answer type – (“Span” / “Yes” / “No”)

Select Module

- [CLS] + Q + [SEP] + D + [SEP]
- One Approach – Use BCE with [CLS] embeddings as features
- Neglects inter-document interactions

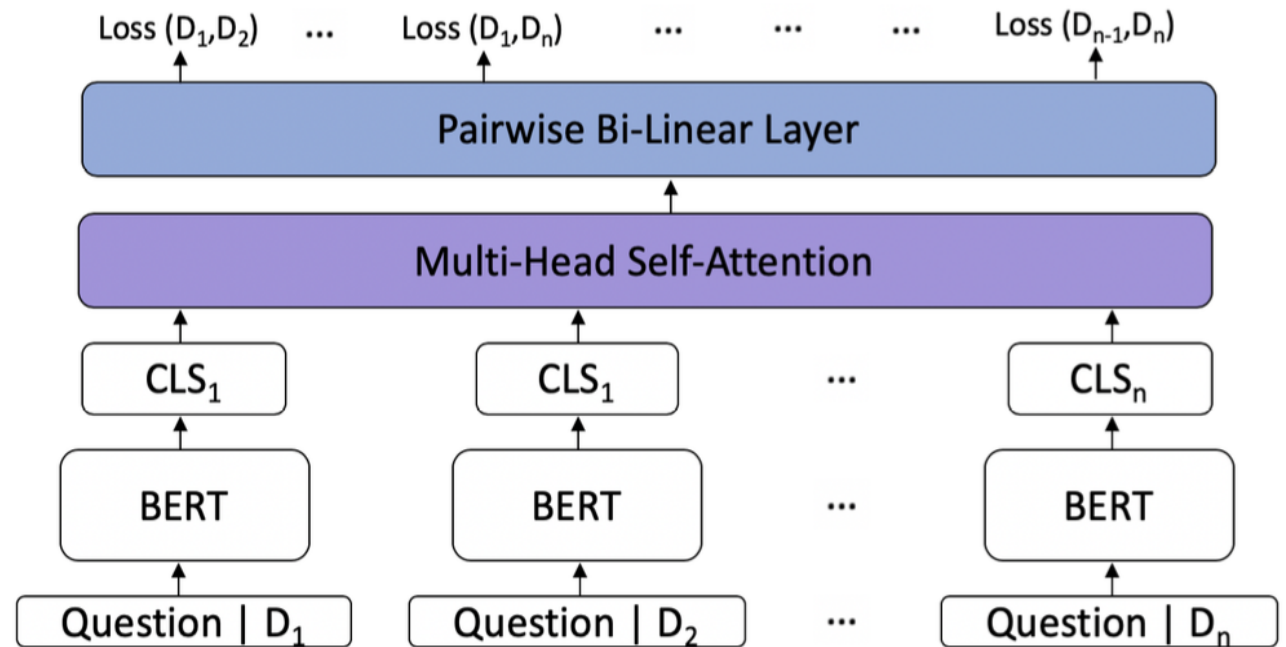
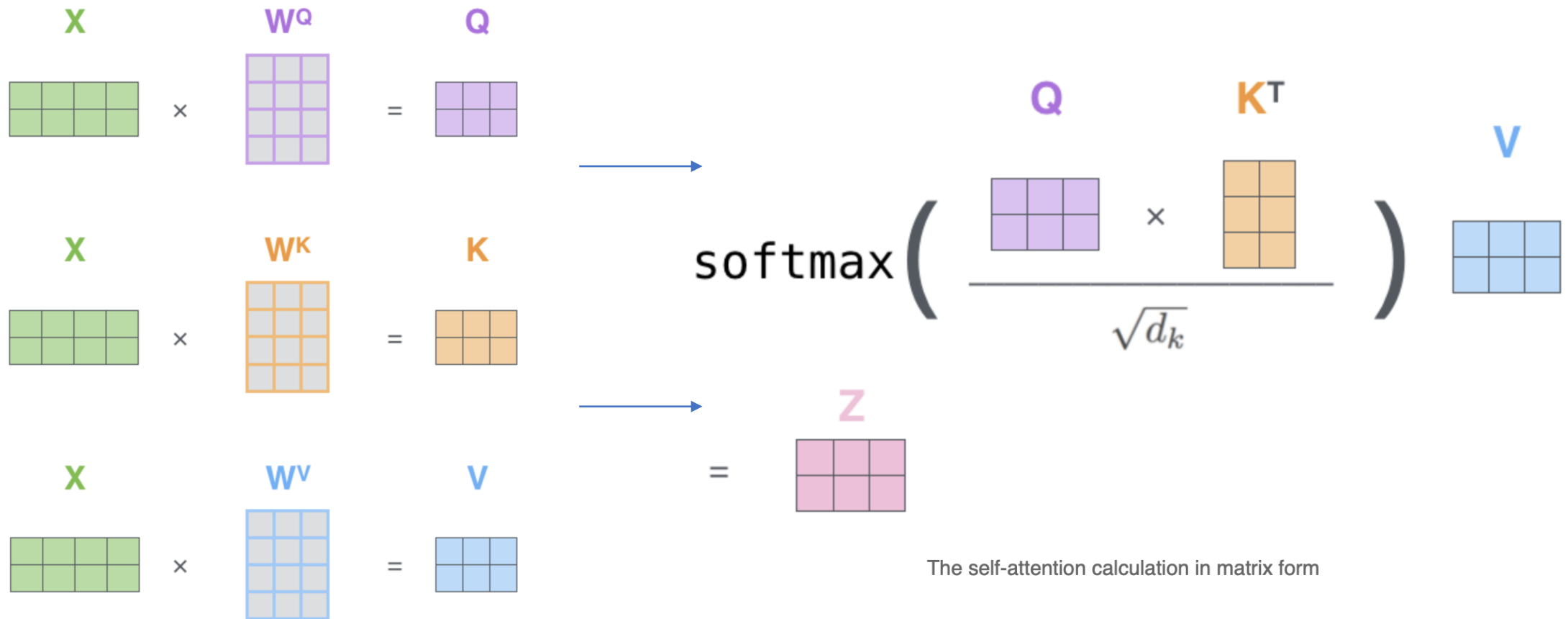


Figure 3: Diagram of document selection module. N indicates the total number of documents.

MHSA – Single Attention Head



X – matrix of [CLS] embeddings of question/document pairs

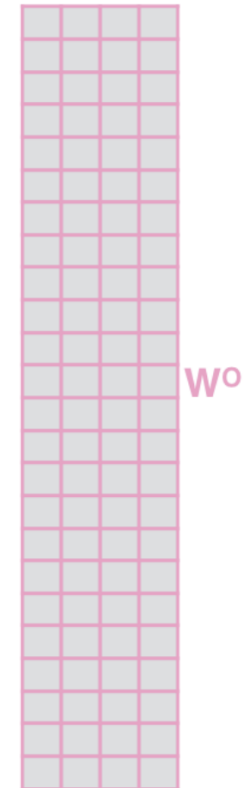
MHSA – Multiple Attention Heads

1) Concatenate all the attention heads

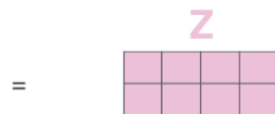


2) Multiply with a weight matrix W^O that was trained jointly with the model

x



3) The result would be the Z matrix that captures information from all the attention heads. We can send this forward to the FFNN



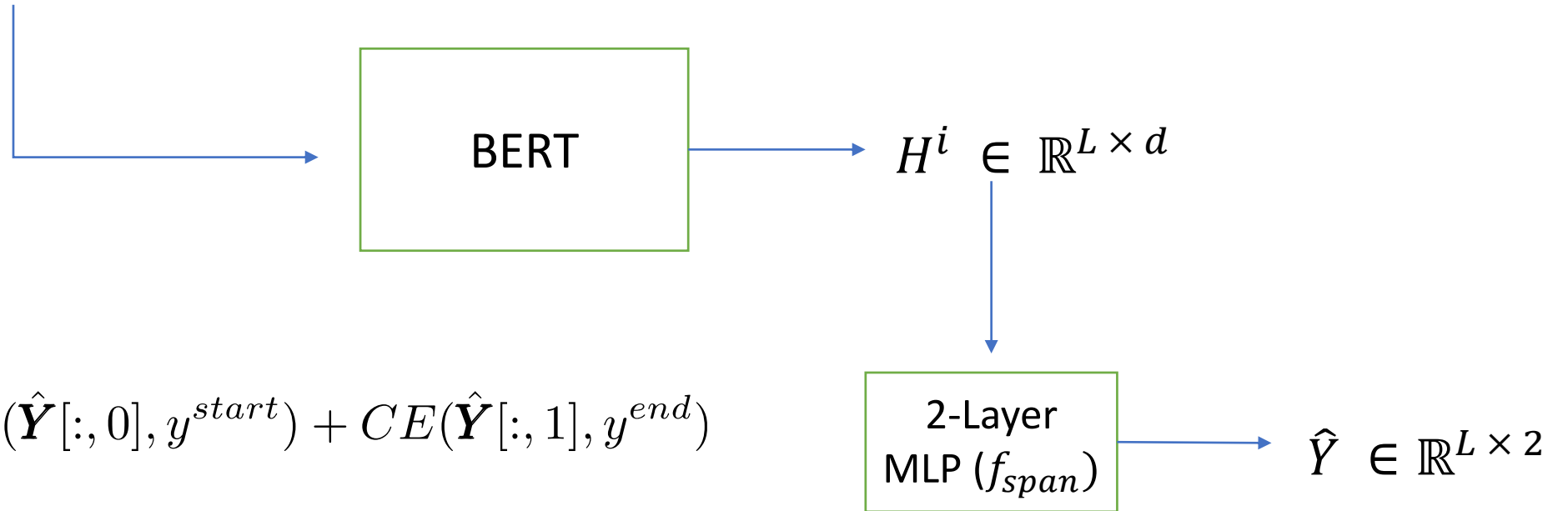
Output is the matrix of modified [CLS] embeddings having contextual information

Pairwise Bi-Linear Layer

- $S(D_i)$ - Score for each document (0/1/2)
- $l_{i,j} = \begin{cases} 1 & \text{if } S(D_i) > S(D_j) \\ 0 & \text{if } S(D_i) \leq S(D_j) \end{cases}$
- $L = -\sum_{i=0}^n \sum_{j=0, j \neq i}^i l_{i,j} \log(P(D_i, D_j)) + (1 - l_{i,j}) \log(1 - P(D_i, D_j))$
- $R_i = \sum_j^n \mathbb{I}(P(D_i, D_j) > 0.5)$ - Relevance score for each document
- Take top-k documents according to this relevance score

Answer Prediction

- Gold Documents extracted from Select Module
- [CLS] + Q + [SEP] + Context + [SEP]

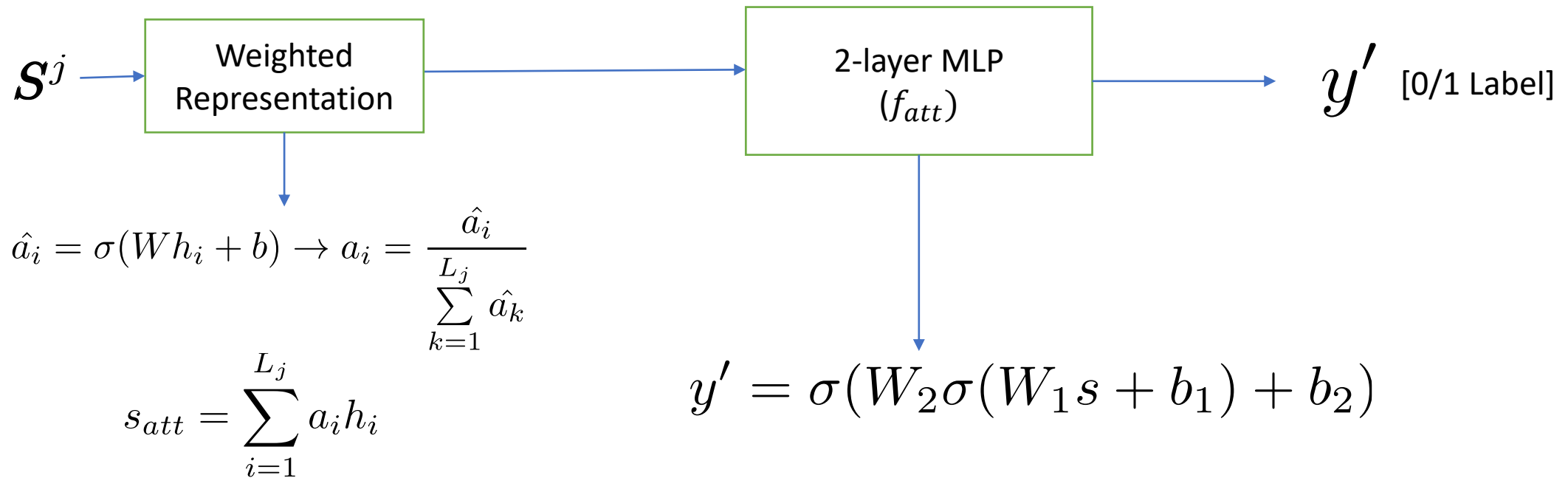


$$L^{span} = \frac{1}{2} (CE(\hat{Y}[:, 0], y^{start}) + CE(\hat{Y}[:, 1], y^{end}))$$

Contextual Sentence Embeddings

- Sentence Representation: $\mathbf{S}^j = \mathbf{H}[j^s : j^e, :] \in \mathbb{R}^{L^j \times d}$
- Self Attention Weights: $f_{att}(S_j)$

$$L_{sent} = \sum_t (y'_t - y_t)^2$$



Contextual Sentence Embeddings - 2

- Motivation for adding start and end span probabilities
 - Answer span -> Supporting Sentence

$$\alpha^j = \sigma(f_{att}(\mathbf{S}^j) + \hat{\mathbf{Y}}[j^s : j^e, 0] + \hat{\mathbf{Y}}[j^s : j^e, 1])$$

- Final sentence embeddings:

$$\mathbf{s}^j = \sum_{k=0}^{L^j} \alpha_k^j \mathbf{S}^j[k, :] \in \mathbb{R}^{1 \times d}$$

Sentence Graph

- Construct a graph with the following properties:
 - Nodes represent the sentences
 - Each node has label 0/1 (supporting sentence)
 - 3 types of edges
 - Between nodes present in the same document (Type 1)
 - Between nodes of different documents if they have named entities / noun phrases (can be different) present in the question (Type 2)
 - Between nodes of different documents if they have the same named entity / noun phrase (Type 3)

Sentence Graph

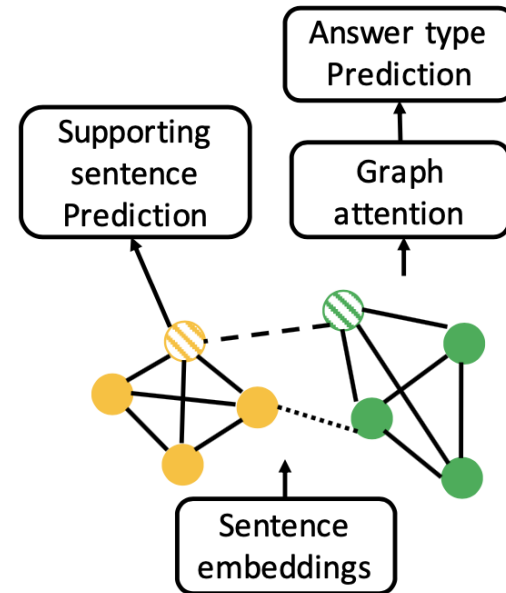
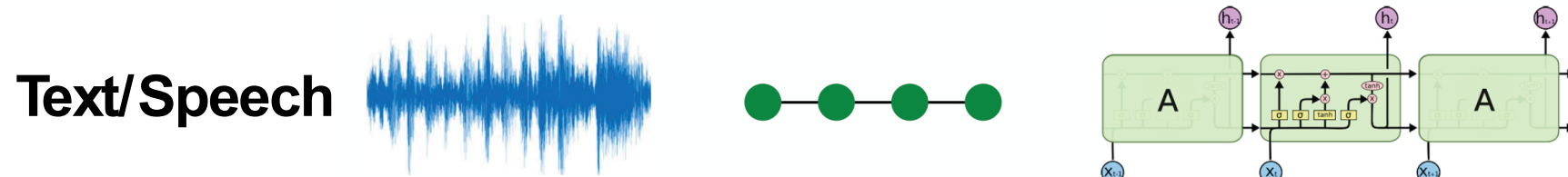
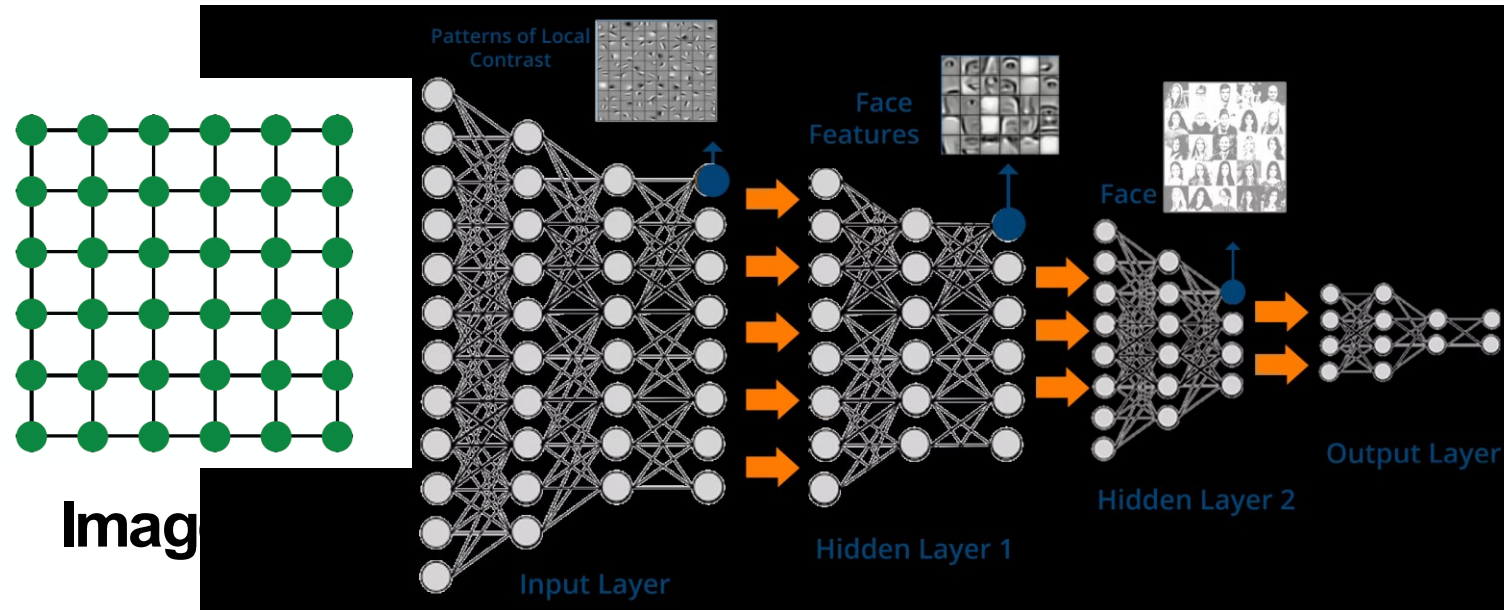


Figure 4: Diagram of the supporting sentence prediction module. Different node colors indicate that nodes are from different documents. Nodes with diagonal stripes indicate that the corresponding input sentences are supporting sentences. We use solid, dashed and dotted lines to differentiate different edge types.

Topics

- Introduction and HotpotQA
- Select, Answer and Explain
- **GNNs**
- Answer and Explain
- Results and Ablation Study
- Reviews

Modern ML Toolbox



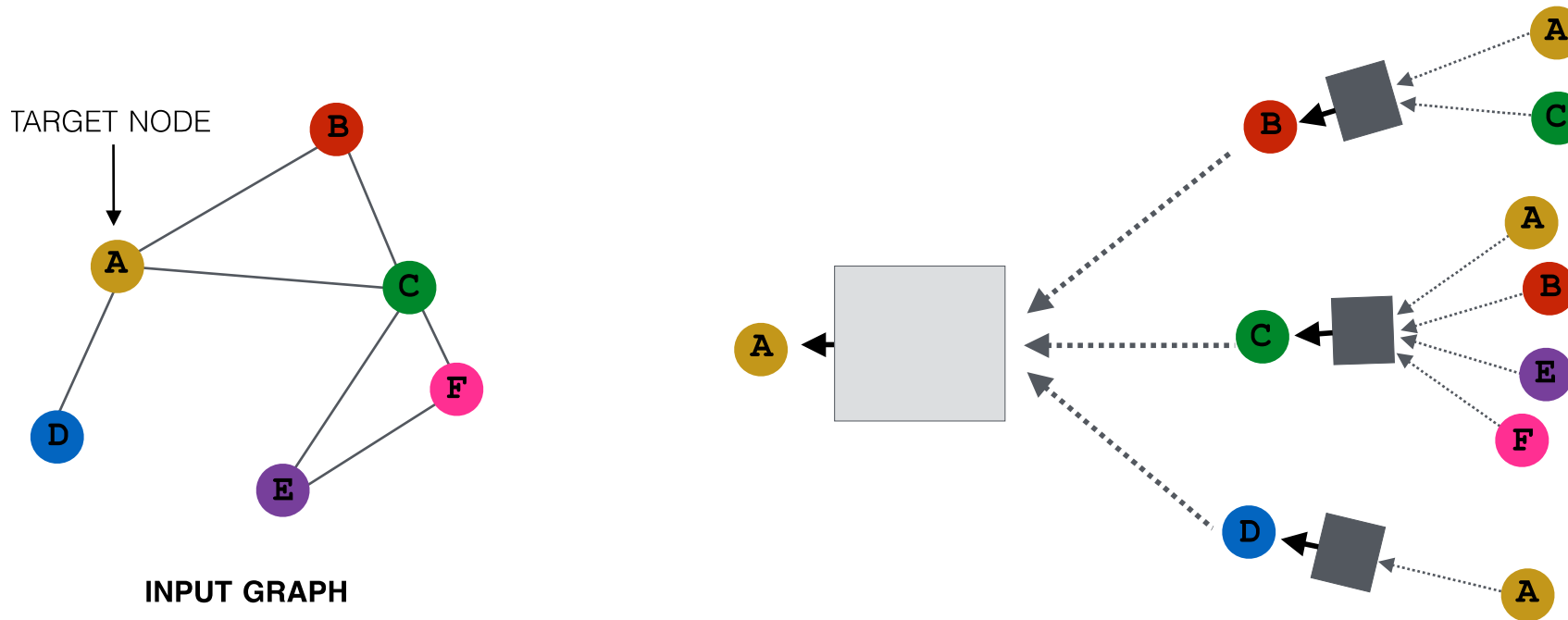
Modern deep learning toolbox is designed for simple sequences & grids

Setup

- Assume we have a graph G :
 - V is the vertex set.
 - A is the adjacency matrix (assume binary).
 - $X \in \mathbb{R}^{m \times |V|}$ is a matrix of node features.
 - Categorical attributes, text, image data
 - E.g., profile information in a social network.
 - Node degrees, clustering coefficients, etc.
 - Indicator vectors (i.e., one-hot encoding of each node)

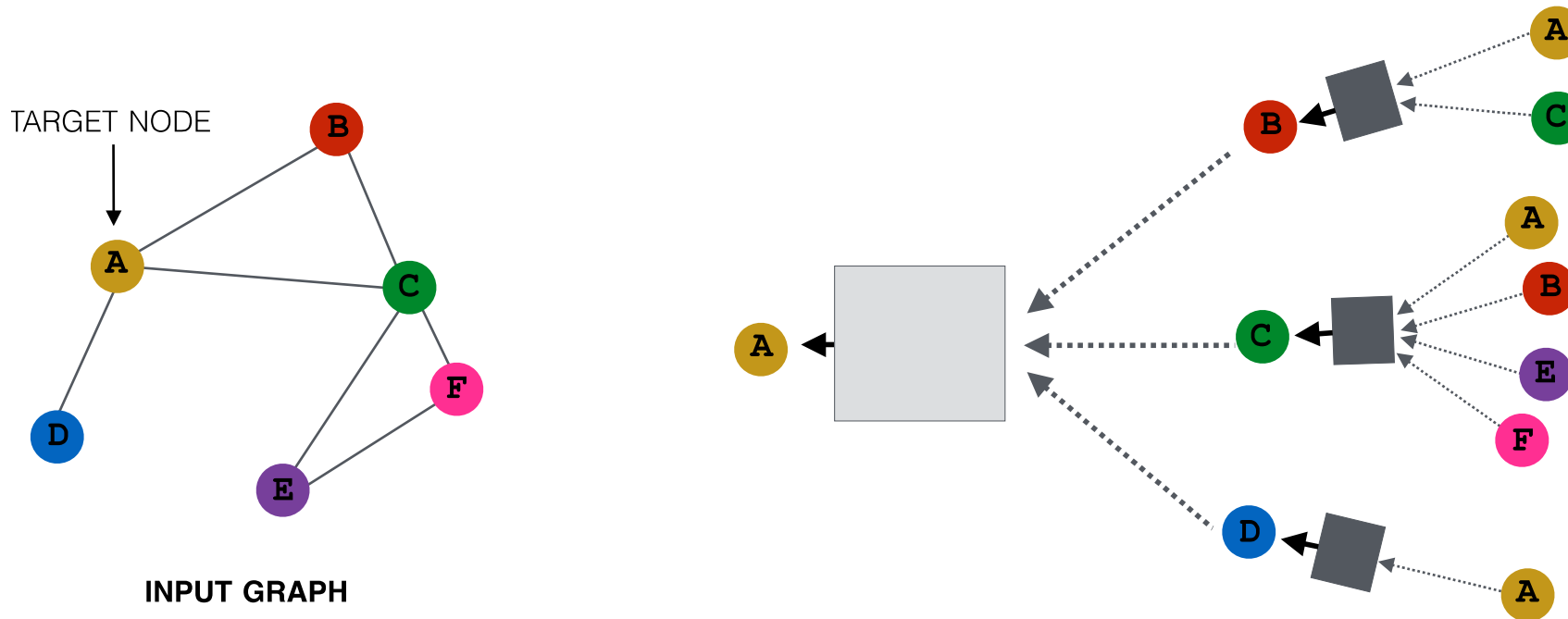
Neighborhood Aggregation

- **Key idea:** Generate node embeddings based on local neighborhoods.



Neighborhood Aggregation

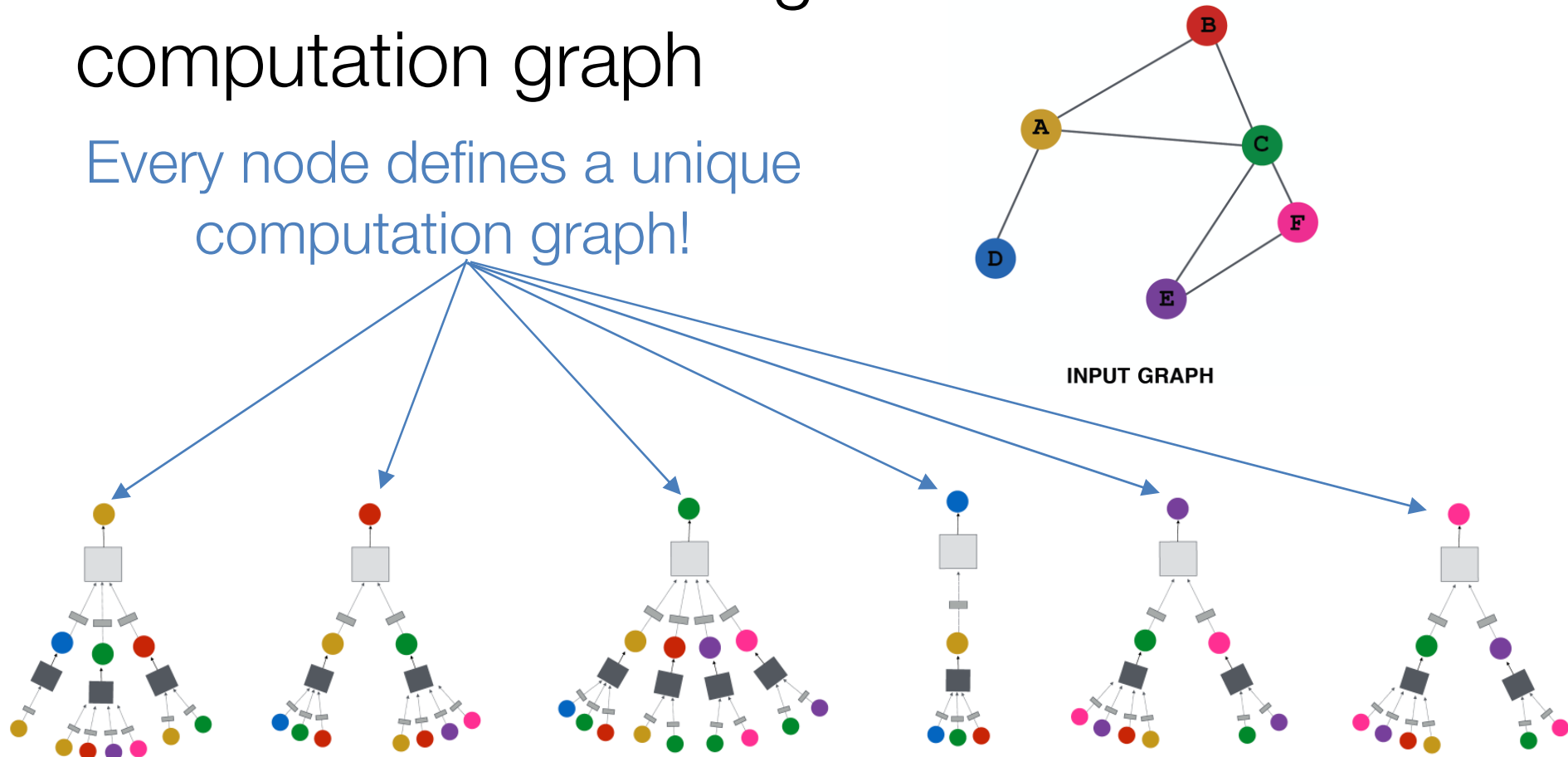
- **Intuition:** Nodes aggregate information from their neighbors using neural networks



Neighborhood Aggregation

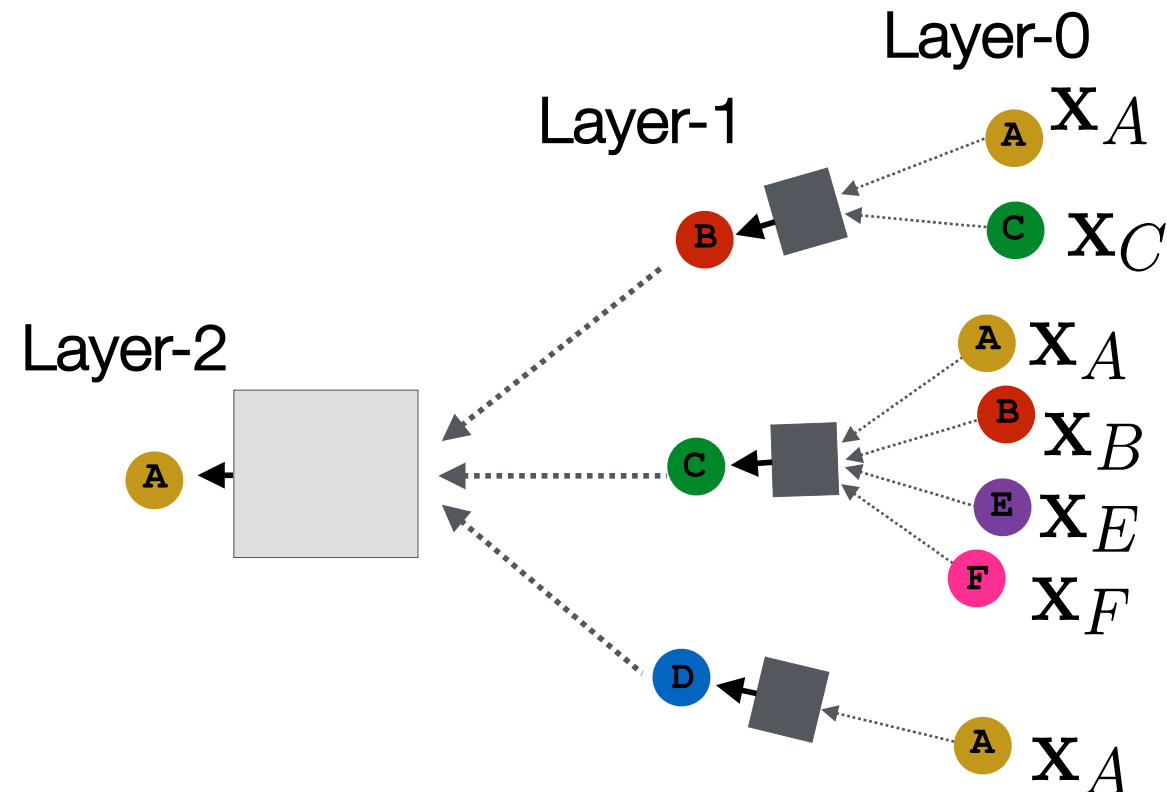
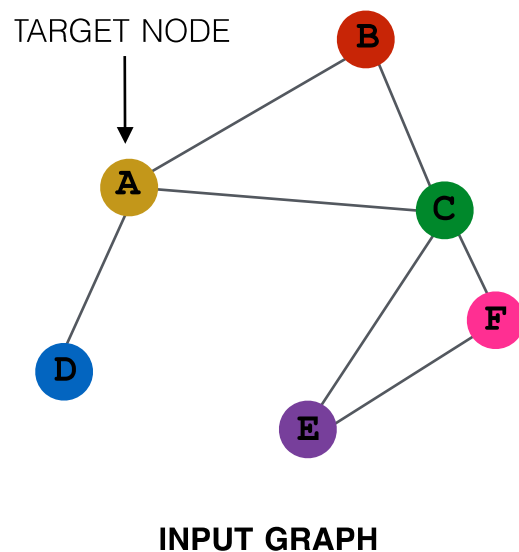
- **Intuition:** Network neighborhood defines a computation graph

Every node defines a unique computation graph!



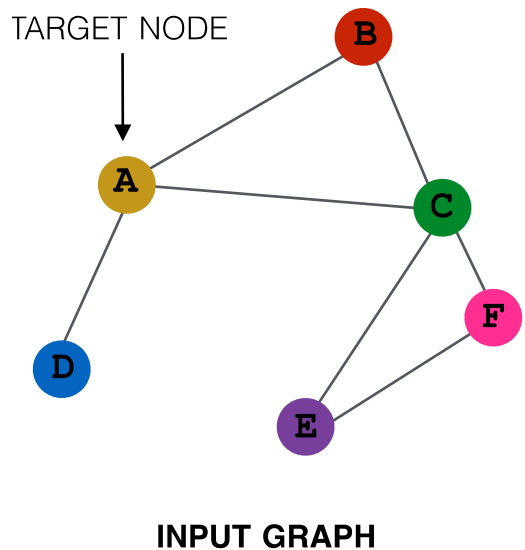
Neighborhood Aggregation

- Nodes have embeddings at each layer.
- Model can be arbitrary depth.
- “layer-0” embedding of node u is its input feature, i.e. x_u .

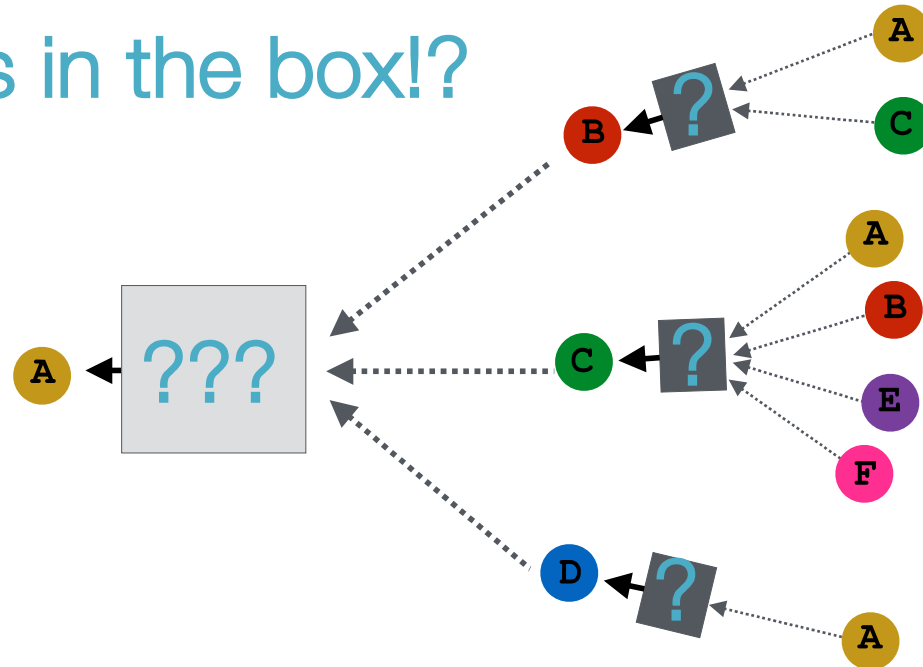


Neighborhood Aggregation

- Key distinctions are in how different approaches aggregate information across the layers.

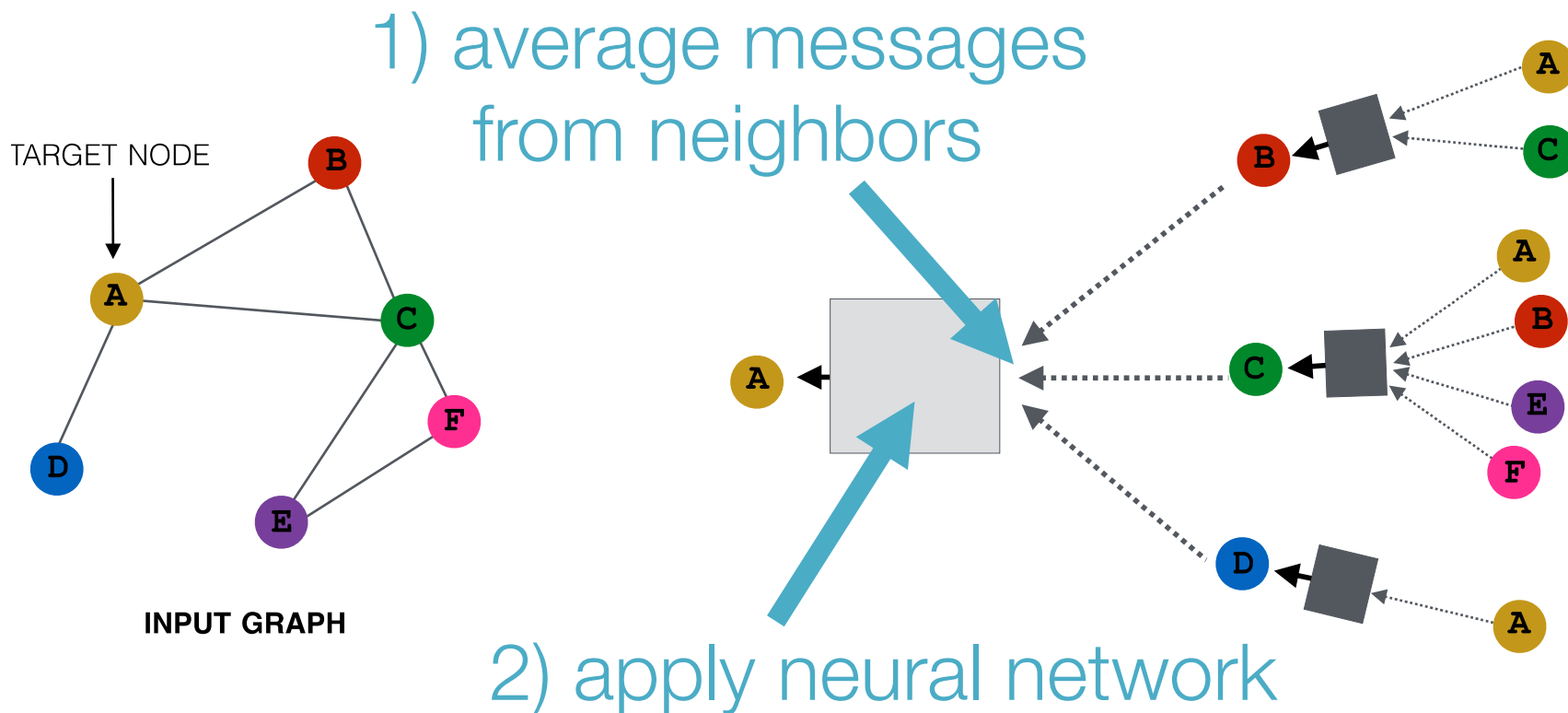


what's in the box!?



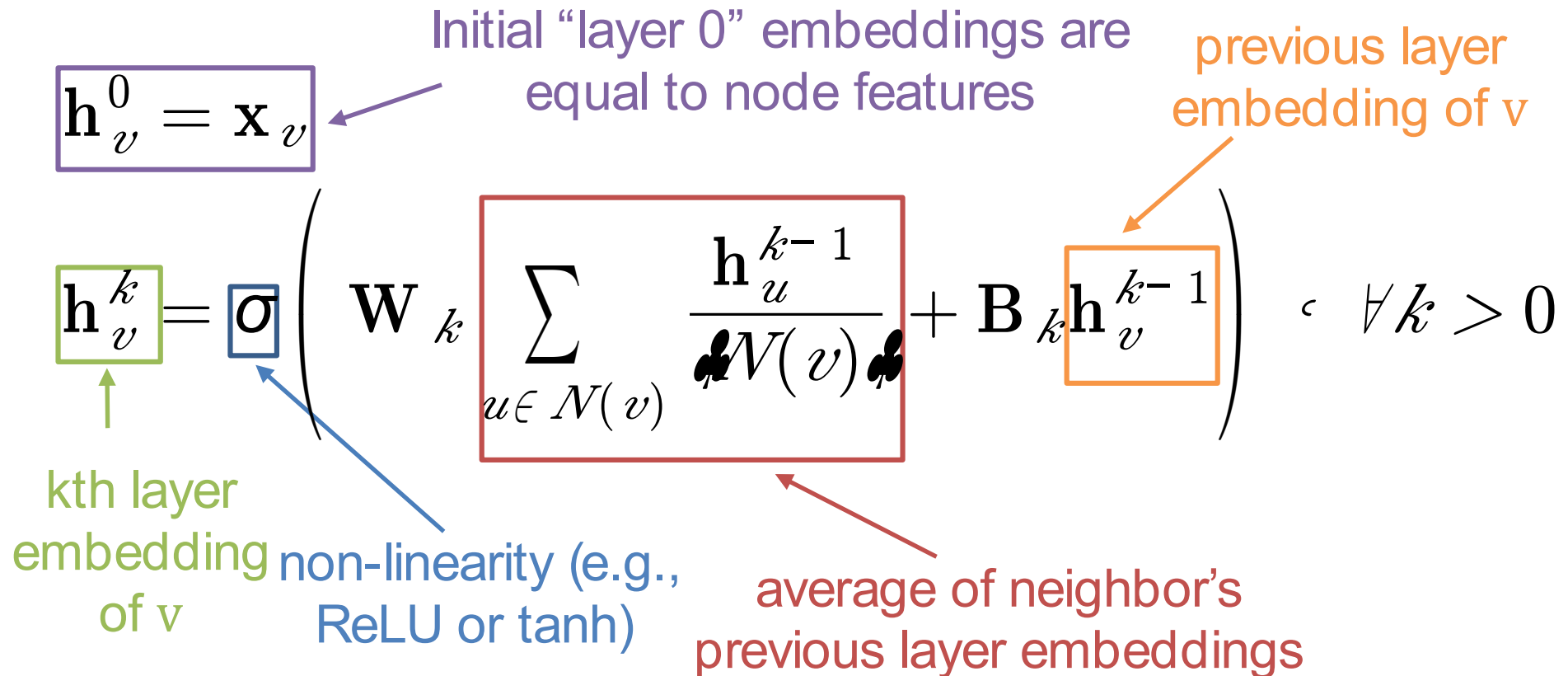
Neighborhood Aggregation

- Basic approach: Average neighbor information and apply a neural network.



The Math

- Basic approach: Average neighbor messages and apply a neural network.



Graph Convolutional Networks

Basic Neighborhood Aggregation

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right)$$

VS.

GCN Neighborhood Aggregation

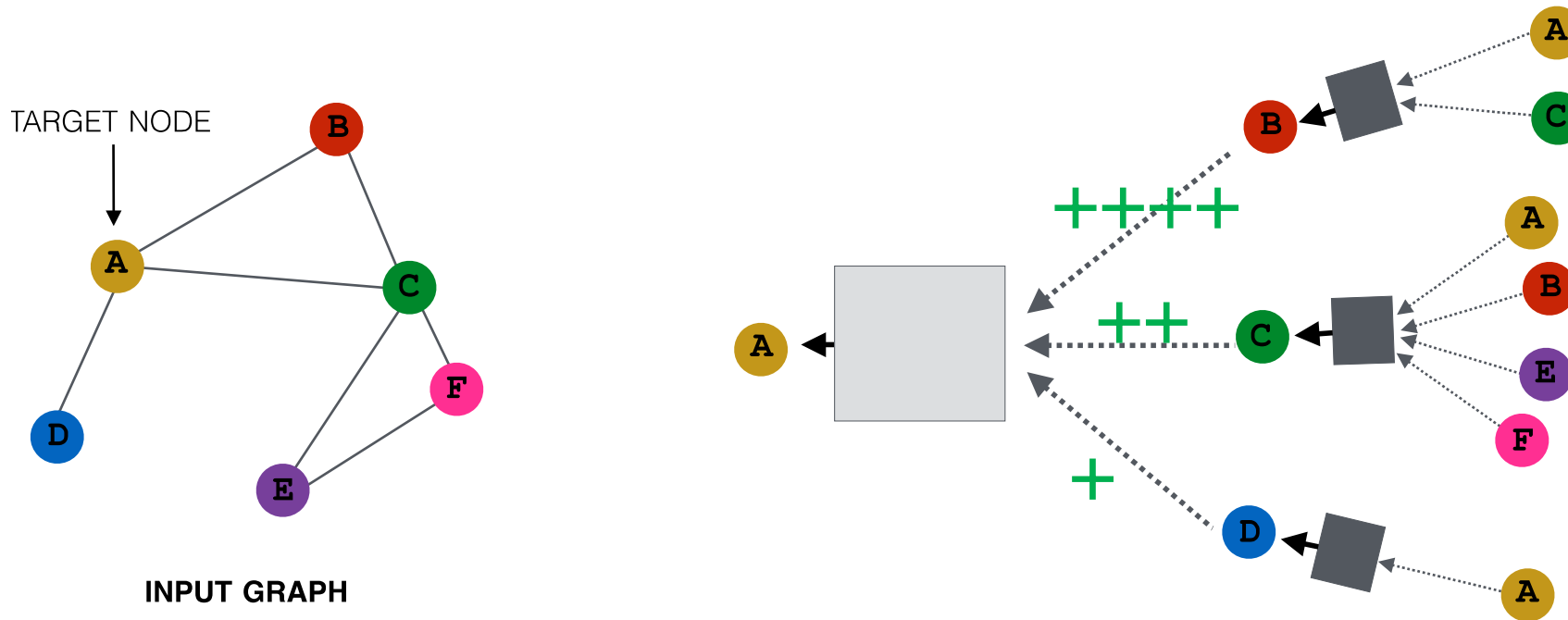
$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}} \right)$$

same matrix for self and
neighbor embeddings

per-neighbor normalization

Neighborhood Attention

- What if some neighbors are more important than others?



Graph Attention Networks

- Augment basic graph neural network model with attention.

$$\mathbf{h}_v^k = \sigma \left(\sum_{u \in \mathcal{N}(v) \cup \{v\}} \alpha_{v,u} \mathbf{W}^k \mathbf{h}_u^{k-1} \right)$$

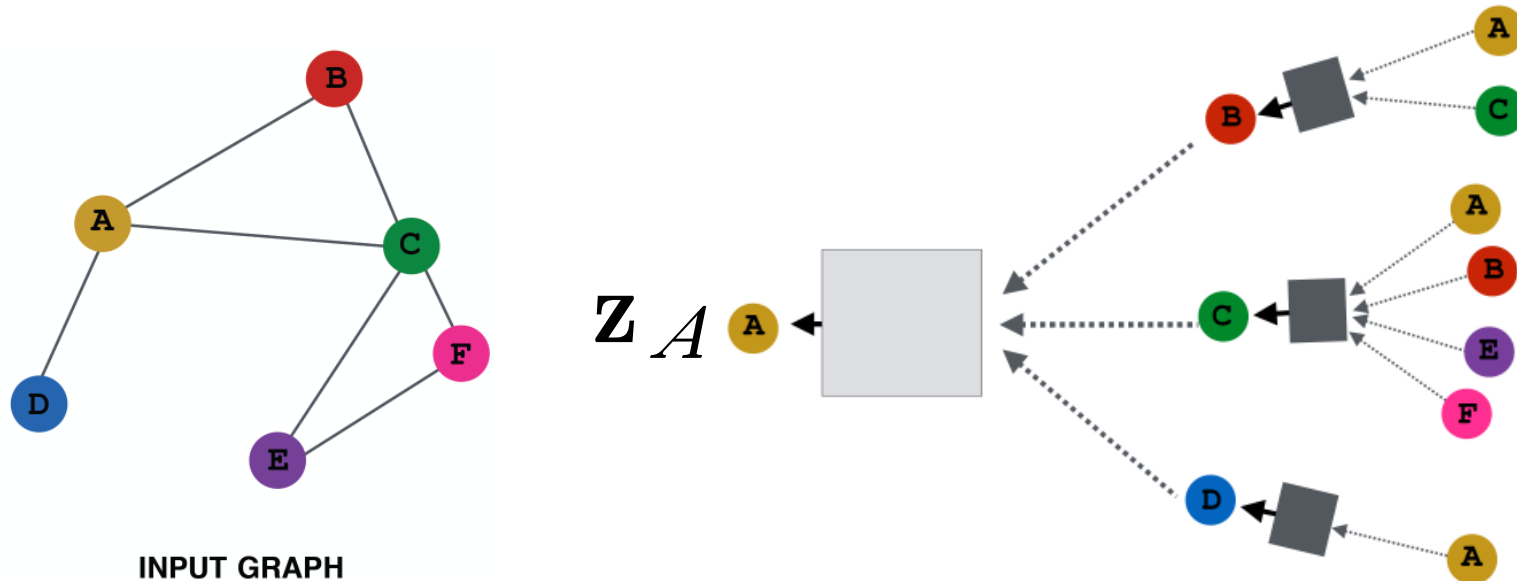
Non-linearity

Sum over all neighbors (and the node itself)

Learned attention weights!

Training the Model

- How do we train the model to generate “high-quality” embeddings?



Need to define a loss function on the embeddings, $\mathcal{L}(z_u)$!

Training the Model

trainable matrices
(i.e., what we learn)

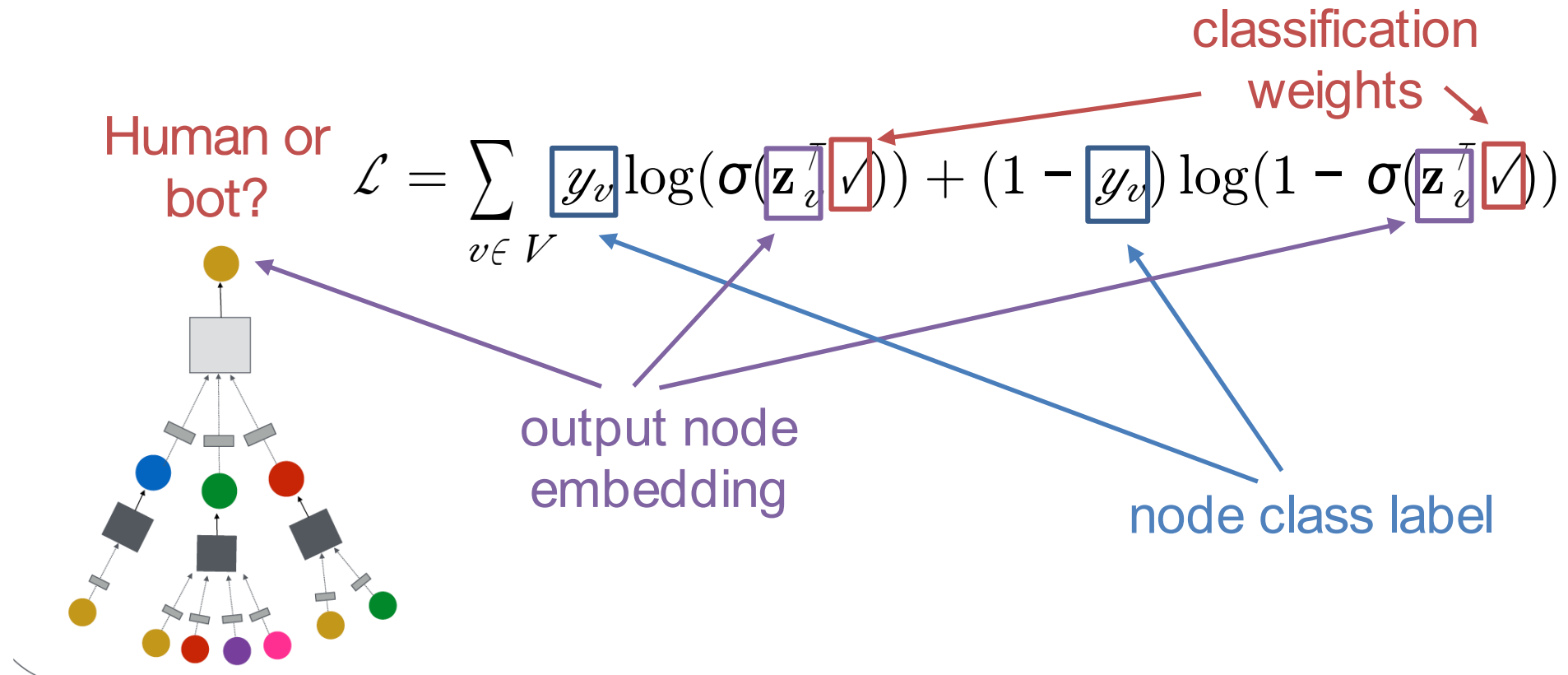
$$\mathbf{h}_v^0 = \mathbf{x}_v$$
$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in \mathcal{N}(v)} \frac{\mathbf{h}_u^{k-1}}{|\mathcal{N}(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right) \quad \forall k \in \{1, \dots, K\}$$

$\mathbf{z}_v = \mathbf{h}_v^K$

- After K -layers of neighborhood aggregation, we get output embeddings for each node.
- We can feed these embeddings into any loss function and run stochastic gradient descent to train the aggregation parameters.

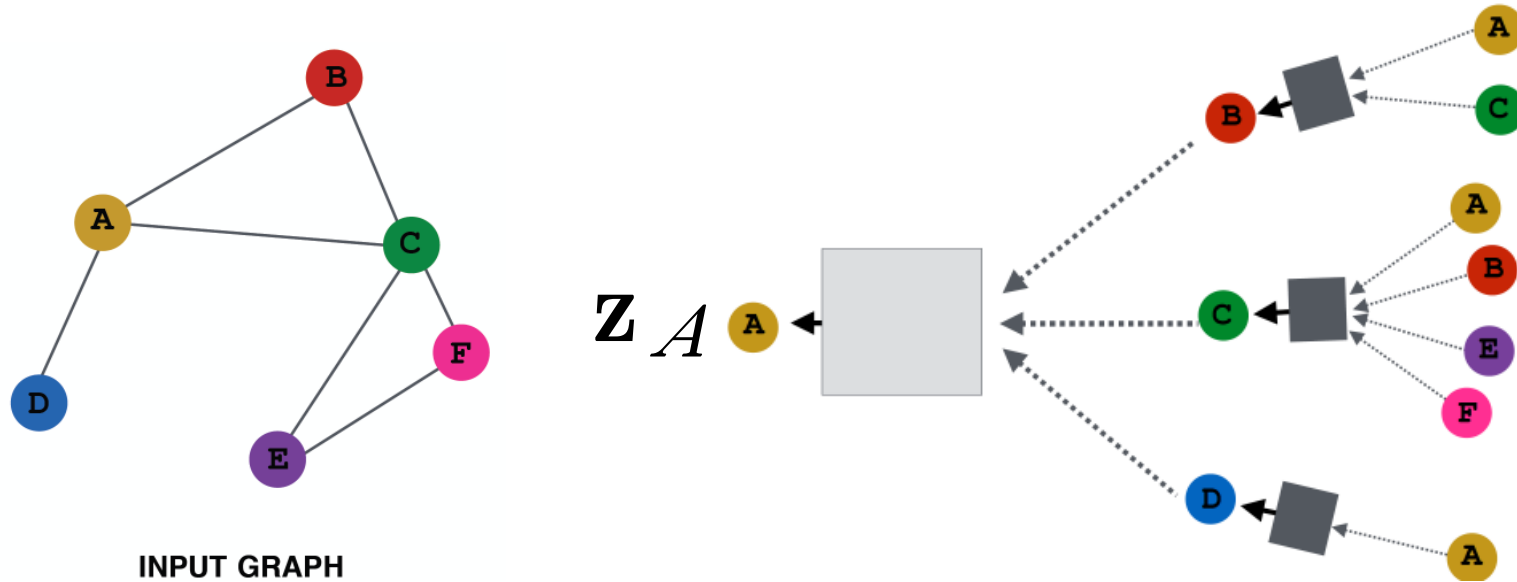
Training the Model

- Alternative: Directly train the model for a supervised task (e.g., node classification):



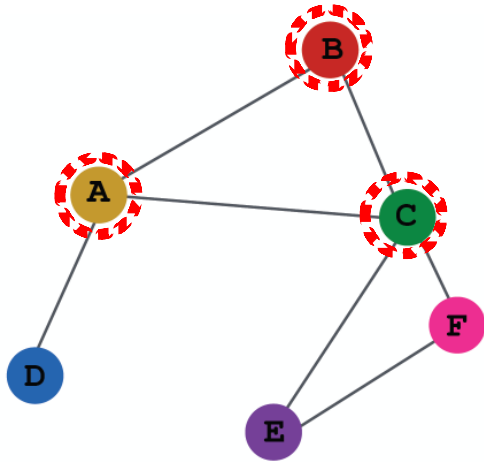
Overview of Model

1) Define a neighborhood aggregation function.



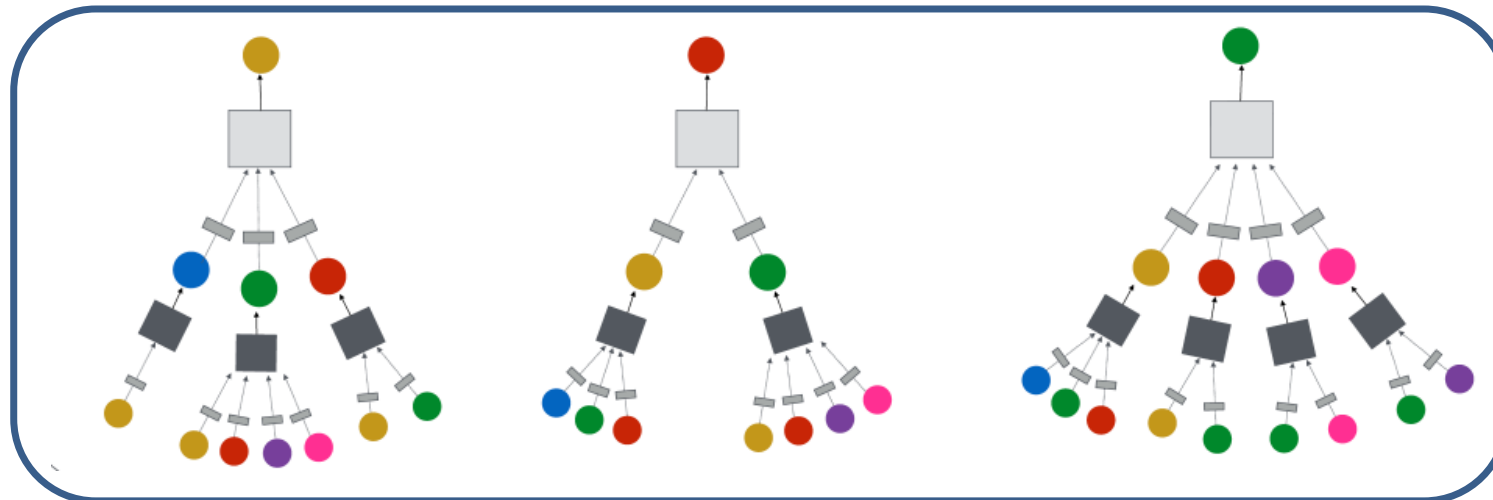
2) Define a loss function on the embeddings, $\mathcal{L}(z_u)$

Overview of Model

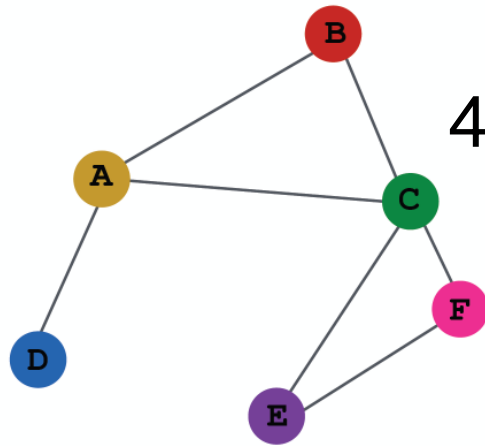


3) Train on a set of nodes, i.e., a batch of compute graphs

INPUT GRAPH



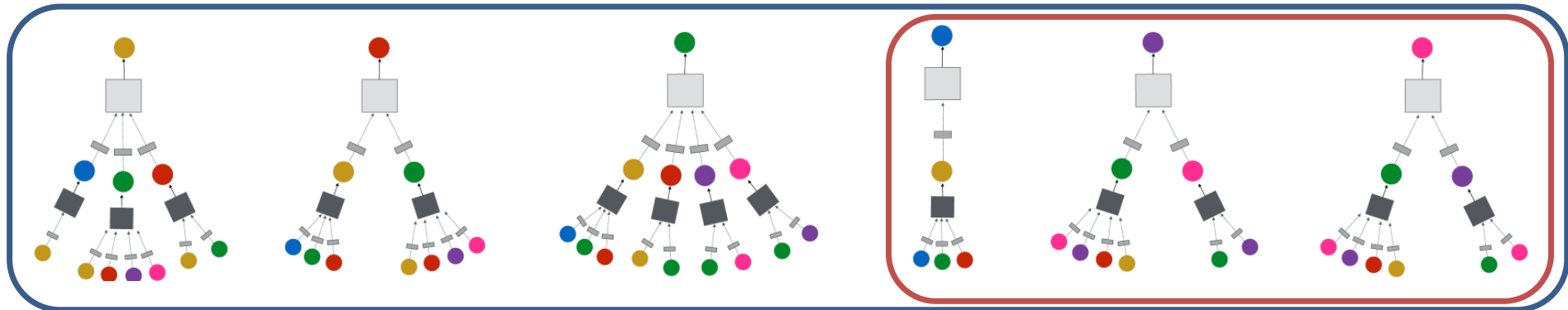
Overview of Model



4) Generate embeddings for nodes as needed

Even for nodes we never trained on!!!!

INPUT GRAPH



Topics

- Introduction and HotpotQA
- Select, Answer and Explain
- GNNs
- **Answer and Explain**
- Results and Ablation Study
- Reviews

Aggregation mechanism in SAE

$$\mathbf{h}_j^{k+1} = \text{act}(\mathbf{u}_j^k) \odot \mathbf{g}_j^k + \mathbf{h}_j^k \odot (1 - \mathbf{g}_j^k) \quad (11)$$

where

$$\mathbf{u}_j^k = f_s(\mathbf{h}_j^k) + \sum_{r \in \mathcal{R}} \frac{1}{|\mathcal{N}_j^r|} \sum_{n \in \mathcal{N}_j^r} f_r(\mathbf{h}_n^k), \quad (12)$$

$$\mathbf{g}_j^k = \text{sigmoid}(f_g([\mathbf{u}_j^k; \mathbf{h}_j^k])). \quad (13)$$

Graph Representation

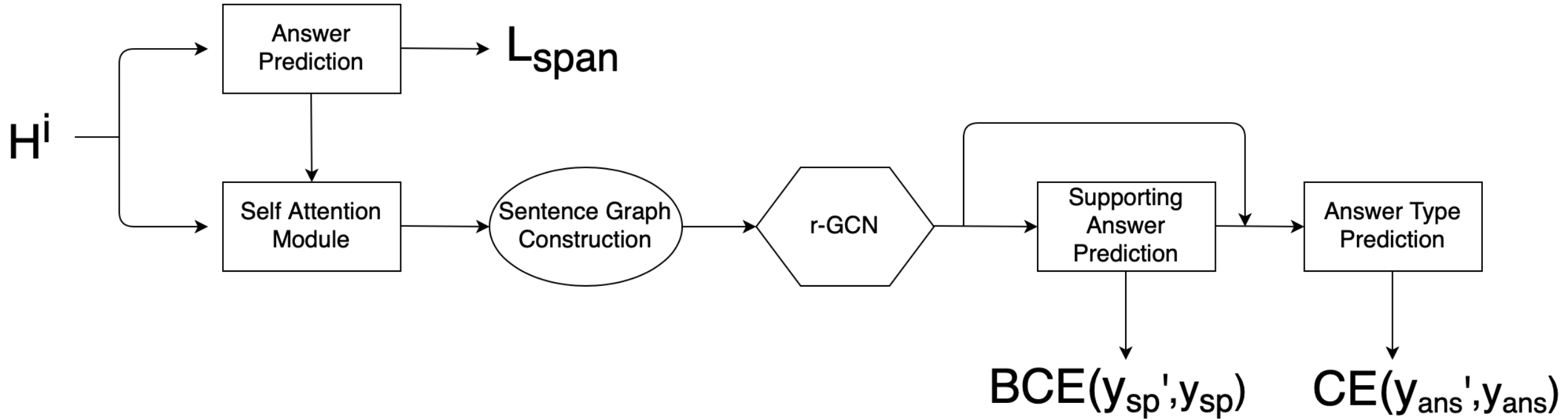
- Weighted sum of the embeddings of the nodes of the graph

$$h = \sum_j a_j h_j$$

- The weights are given by

$$a = \sigma(\hat{\mathbf{y}}^{sp})$$

Answer and Explain Pipeline



Topics

- Introduction and HotpotQA
- Select, Answer and Explain
- GNNs
- Answer and Explain
- Results and Ablation Study
- Reviews

Dataset Details

- Train – 90K
- Validation/Dev – 7.4K
- Test – 7.4K

Results

Table 1: Results comparison between our proposed SAE system with other methods. * indicates unpublished models.

	Model	Ans		Sup		Joint	
		EM	F_1	EM	F_1	EM	F_1
Dev	Baseline(Yang et al. 2018)	44.44	58.28	21.95	66.66	11.56	40.86
	QFE(Nishida et al. 2019)	53.70	68.70	58.80	84.70	35.40	60.60
	DFGN(Xiao et al. 2019)	55.66	69.34	53.10	82.24	33.68	59.86
	SAE(ours)	61.32	74.81	58.06	85.27	39.89	66.45
	SAE-oracle(ours)	63.48	77.16	62.80	89.29	42.77	70.13
	SAE-large(ours)	67.70	80.75	63.30	87.38	46.81	72.75
Test	Baseline(Yang et al. 2018)	45.46	58.99	22.24	66.62	12.04	41.37
	QFE(Nishida et al. 2019)	53.86	68.06	57.75	84.49	34.63	59.61
	DFGN(Xiao et al. 2019)	56.31	69.69	51.50	81.62	33.62	59.82
	SAE(ours)	60.36	73.58	56.93	84.63	38.81	64.96
	SAE-large(ours)	66.92	79.62	61.53	86.86	45.36	71.45
	C2F Reader*	67.98	81.24	60.81	87.63	44.67	72.73

Ablation Study – Document Selection Module

Table 2: Ablation study results on HotpotQA dev set. PR(0,1) stands for giving 0 score to non-gold documents and 1 score to all gold documents when preparing pairwise labels, and PR(0,1,2) stands for giving 2 score to the gold document with answer span.

	EM _S	Recall _S	Acc _{span}	joint EM	joint F_1
BERT only	70.65	89.16	90.08	31.87	59.33
+MHSA	87.07	94.65	92.54	38.54	65.00
+PR(0,1)	89.76	94.75	94.53	39.53	65.44
+PR(0,1,2)	91.40	95.61	95.86	39.89	66.45

Ablation Study – Answer & Explain Module

Table 3: Ablation study results on HotpotQA dev set.

	joint EM	joint F_1
full model	39.89	66.45
-mixed attn	39.59	66.28
-attn sum	38.04	65.33
-GNN	38.46	65.53
-type 1 edge	38.15	65.00
-type 2 edge	39.55	66.13
-type 3 edge	39.32	66.03
-type 2&3 edge	39.16	65.76

Ablation Study – Bridge / Comp. Questions

Table 4: Performance comparison in terms of joint EM and F_1 scores under different reasoning types.

	Bridge (5918 samples)		Comparison (1487 samples)	
	joint EM	joint F_1	joint EM	joint F_1
Baseline	8.80	39.77	20.91	43.24
DFGN	30.09	58.61	47.95	64.79
SAE	37.07	66.12	51.18	67.73

Attention Heatmap Example







Figure 5: Attention heatmap of a sample from dev set. Each cell is a word piece token returned by BERT. Sentences with different colors are from different documents.

Question - "Were Scott Derrickson and Ed Wood of the same nationality?"

HotpotQA Leaderboard

Leaderboard (Distractor Setting)

In the *distractor* setting, a question-answering system reads 10 paragraphs to provide an answer (Ans) to a question. They must also justify these answers with supporting facts (Sup).

	Model	Code	Ans		Sup		Joint	
			EM	F ₁	EM	F ₁	EM	F ₁
1 Dec 1, 2019	HGN-large (single model) <i>Anonymous</i>		69.22	82.19	62.76	88.47	47.11	74.21
2 Oct 18, 2019	C2F Reader (single model) <i>Joint Laboratory of HIT and iFLYTEK Research</i>		67.98	81.24	60.81	87.63	44.67	72.73
3 Nov 19, 2019	SAE-large (single model) <i>JD AI Research</i> Tu, Huang et al., AACL 2020		66.92	79.62	61.53	86.86	45.36	71.45
4 Sep 27, 2019	HGN (single model) <i>Microsoft Dynamics 365 AI Research</i> Fang et al., 2019		66.07	79.36	60.33	87.33	43.57	71.03

Topics

- Introduction and HotpotQA
- Select, Answer and Explain
- GNNs
- Answer and Explain
- Results and Ablation Study
- **Reviews**

Reviews (Pros)

- Detailed Ablation Study [Atishya, Pratyush, Rajas, Saransh]
- Usage of contextualized sentence embeddings [Atishya, Jigyasa]
- MHSA in Document Selection [Pratyush, Shubham, Rajas, Siddhant]
- “Learning to Rank” framework is general [Keshav]
- Top 3 position on the leaderboard [Pratyush, Keshav, Rajas,]
- Simple Idea [Soumya]
- Single Model gives good performance [Keshav]
- Careful modelling of the loss function [Vipul]
- “Explainability” of the model [Various people]

Reviews (Cons)

- Motivation for Type 2 edges not present [Pratyush, Rajas]
- No clear flow [Atishya]
- Entire context fed to BERT [Pratyush, Jigyasa]
- Pairwise ranking costly [Siddhant, Saransh, Jigyasa]
- Do not evaluate on Fullwiki setting, simple method for edges [Keshav]
- Post-facto explanation [Rajas, Soumya]
- Layers for GCN not mentioned [Vipul]
- GNN not explained clearly, performance gain is low [Pratyush]

Reviews (Extensions)

- Extract relevant spans instead of documents [Pratyush]
- Modify the above extension as span-prediction [Keshav]
- Replace pairwise ranking [Shubham, Saransh]
- End-to-end training (RL/integrate REALM) [Siddhant, Jigyasa]
- OpenIE for graph generation [Keshav]
- Enforce constraints in pairwise prediction models [Atishya]
- Handle exposure bias by gradually replacing gold documents with retrieved documents [Rajas]
- Link sentences using clustering methods [Soumya]

Thank You!

Questions?