Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 505–514
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

GCAN: Graph-aware Co-Attention Networks for Explainable Fake News Detection on Social Media

Yi-Ju Lu
Department of Statistics
National Cheng Kung University
Tainan, Taiwan
l852888@gmail.com

Cheng-Te Li
Institute of Data Science
National Cheng Kung University
Tainan, Taiwan
chengte@mail.ncku.edu.tw

ACL'20

220414 Chia-Chun Ho

Outline

Introduction

Fake news intro

- The convenient and low-cost essence of social networking brings collective intelligence, but at the same time leads to a negative by-product.
  - The propagation of misinformation such as fake news.
- Fake news is a kind of news story possessing intentionally false information on social media.
- The widespread of fake news can mislead the public, and produce unjust political, economic, or psychological profit for some parties.

Introduction
Previous approaches

- Data mining and machine learning techniques were utilized to detect fake news.
- Typical approaches rely on the content of new articles to extract textual features, such as n-gram and bag of words, and apply supervised learning.
- NLP researchers also learn advanced linguistic features, such as factive/assertive verbs and subjectivity and writing styles and consistency.
- Multi-modal context information (user profiles, retweet propagation) is investigated.

Introduction

Challenges (1/4)

- Existing content-based approaches require documents to be long text (news articles).
  - So that the representation of words and sentences can be better learned.
  - However, tweets on social media are usually short text.
    - Data sparsity problem

Introduction

Challenges (2/4)

- Some SOTA models require a rich collection of user comments for every news story.
  - Usually provide strong evidences in identifying fake news.
  - However, most users on social media tend to simply share the source story without leaving any comments.

Introduction
Challenges (3/4)

- Some studies consider that the pathways of information cascade (i.e., retweets) in social media.
  - Learn the representations of the tree-based propagation structures.
  - However, it's costly to obtain the diffusion structure of retweets at most times due to privacy concerns.
  - Many users choose to hide or delete the records of social interaction.

Introduction

Challenges (4/4)

- If the service providers or the government agencies desire to inspect who are the suspicious users who support the fake news.
  - Existing model cannot provide explanations.
  - Although dEFEND('19) can generate reasonable explanation.
    - It requires both long text of source articles and text of user comments.

Introduction

Goal of proposed model

- Predict whether a source tweet story is fake.
  - Given only its short text content and
  - retweet sequence of users, along with user profiles.
- Under three settings:
  - Short-text source tweet
  - No text of user comments
  - No network structures of social network and diffusion network

Introduction

Goal of proposed model

- Moreover, require the fake news detection model to be capable of explainability.
  - i.e., highlighting the evidence when determining a story is fake.
- The model is expected to
  - Point out the suspicious retweets who support the spreading of fake news.
  - Highlight the words they especially pay attention to from the source tweet.

Introduction
GCAN

- Propose a novel model, Graph-aware Co-Attention Network (GCAN).
- First extract user features from profiles and social interactions, and learn word embedding from the source short text.
- Then use CNN & RNN to learn the representation of retweet propagation based on user features.
- Construct a graph to model the potential interactions between users, and GCN is used to learn the graph-aware representation of user interactions.

Introduction
GCAN

- Develop a dual co-attention mechanism to learn
  - The correlation between the source tweet and retweet propagation,
  - The co-influence between the source tweet and user interaction.
- The binary prediction is generated based on the learned embeddings.

Introduction

Contributions

- Study a novel and more realistic scenario of fake news detection on social media.
- Develop new model, GCAN, to better learn the representations of user interactions, retweet propagations, and their correlation with source short text.
- Dual co-attention mechanism can produce reasonable explanations.
- Extensive experiments on real datasets demonstrate the promising performance of GCAN comparing to SOTA models.

Related Works

Fake News Detection

- Content-based
  - TF-IDF, topic feature, language/ writing styles, consistency, and social emotions.
- User-based
  - User profiles features utilized RNN & CNN / heterogeneous graph embedding.
- Structure-based
  - Propagation structure, tree-struct RNN to learn embedding / heterogeneous information network
- Hybrid-based
  - fuse multi-modal context information regarding the source tweets. EANN / dEFEND

Problem Statement

Notations

- $\Psi = \{s_1, s_2 \ldots s_{|\Psi|}\}$: set of tweet stories.
  - $s_i = \{q_1^i, q_2^i, \ldots q_{l_i}^i\} \in \Psi$: source tweet indicating l_i words in tweet s_i.
- $U = \{u_1, u_2 \ldots u_{|U|}\}$: set of users.
  - Each $u_j \in U$ is associated with a user vector $\mathbf{x}_j \in \mathbb{R}^d$.
- $R_i = \{\ldots, (u_j, \mathbf{x}_j, t_j), \ldots\}$: propagation path of s_i.
  - $(u_j, \mathbf{x}_j, t_j)$: u_j (with feature vector $\mathbf{x}_j$) who retweet s_i

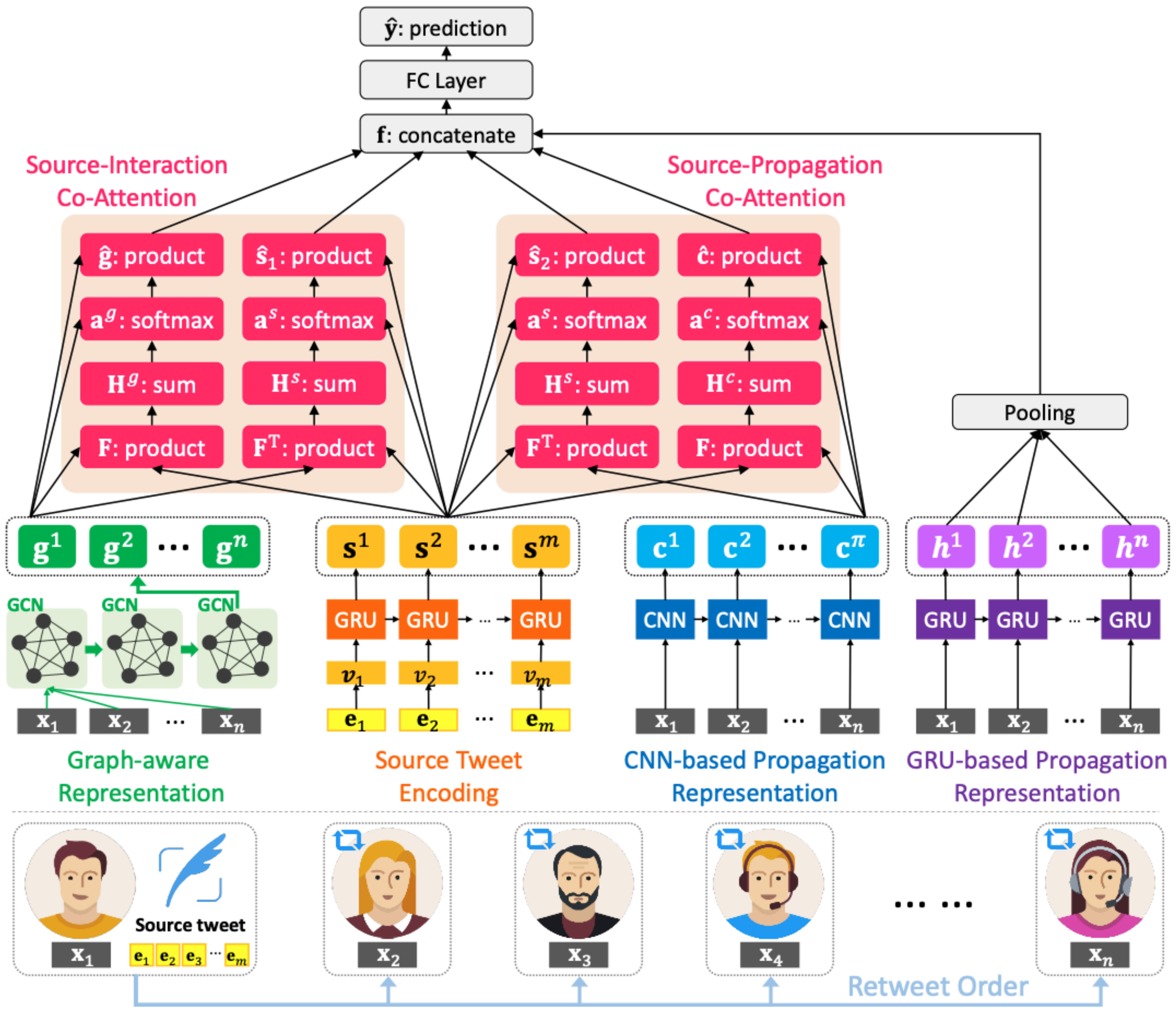
Problem Statement
Problem

- Given a source tweet s_i, along with the corresponding propagation path R_i.
- Goal is to predict the truthfulness y_i of tweet s_i.
- In addition, require model to highlight few users $u_j \in U_i$ who retweet s_i and few words $q_k^i \in s_i$ that can interpret why s_i is identified as a true of fake one.

Methodology

Graph-aware Co-Attention Networks (GCAN)

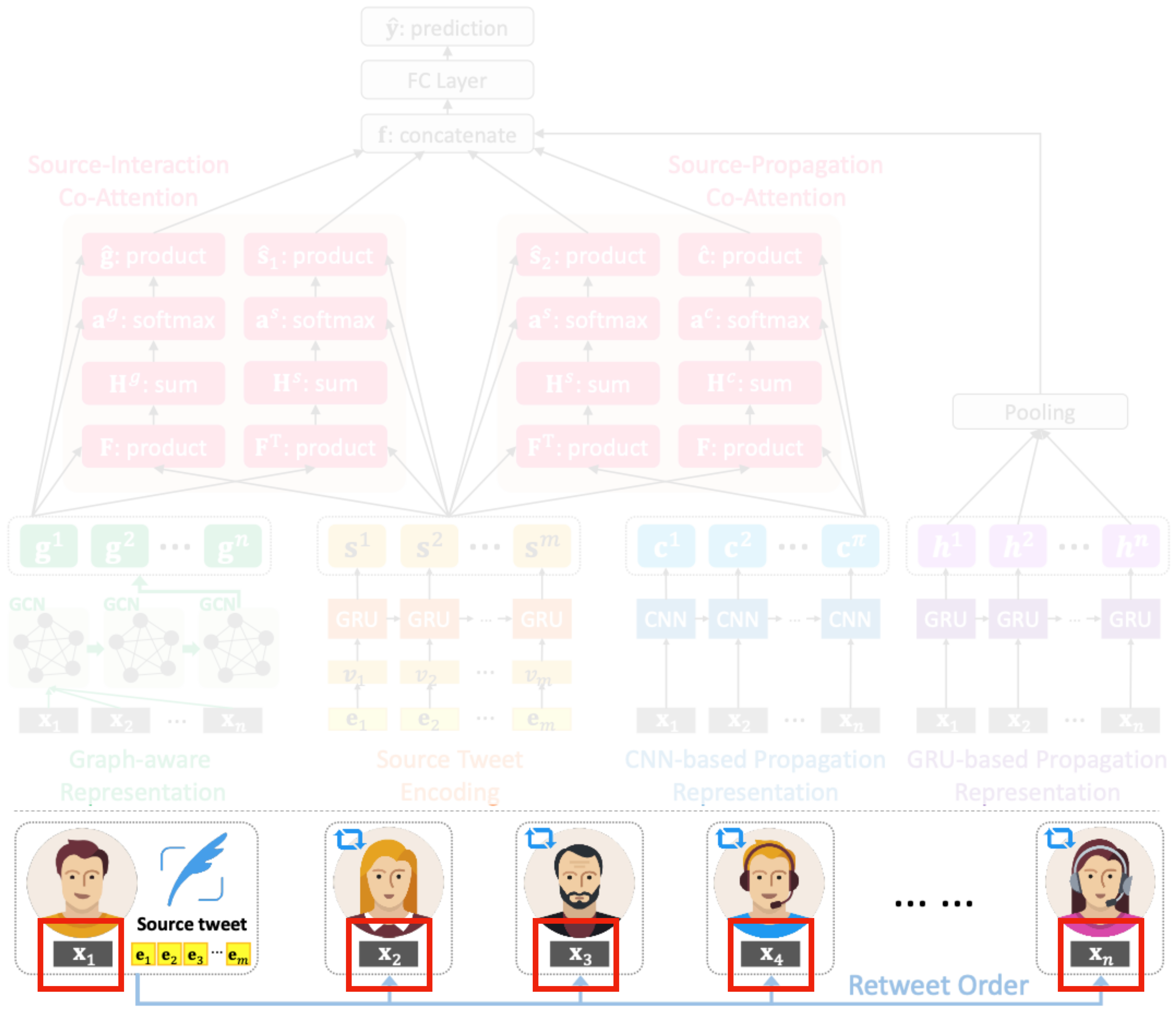
- User characteristic extraction
- News tweet encoding
- User propagation representation
- Co-attention mechanism
- Making prediction



Methodology

Graph-aware Co-Attention Networks (GCAN)

- User characteristic extraction
- News tweet encoding
- User propagation representation
- Co-attention mechanism
- Making prediction



Methodology

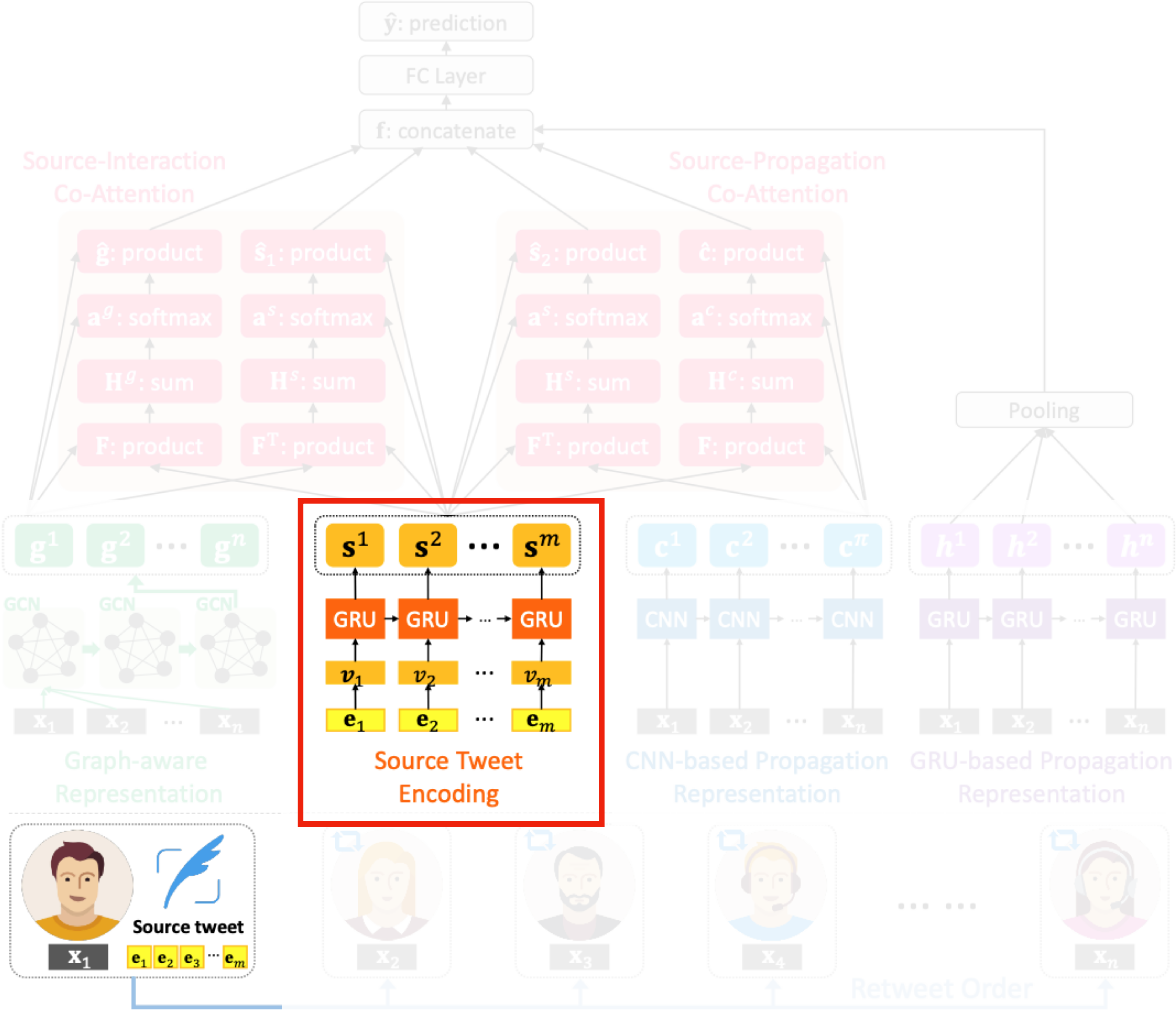
User characteristic extraction

- # of words in a u_j's self-description
- # of words in u_j's screen name
- # of users who follows u_j
- # of users that u_j is following
- # of created tweets for u_j
- Time elapsed after u_j's first tweet
- Whether u_j allows geo-spatial positioning
- Time difference between the source tweet's post time and u_j's retweet.
- Length of retweet path between u_j and the source tweet.

Methodology

Graph-aware Co-Attention Networks (GCAN)

- User characteristic extraction
- News tweet encoding
- User propagation representation
- Co-attention mechanism
- Making prediction



Methodology

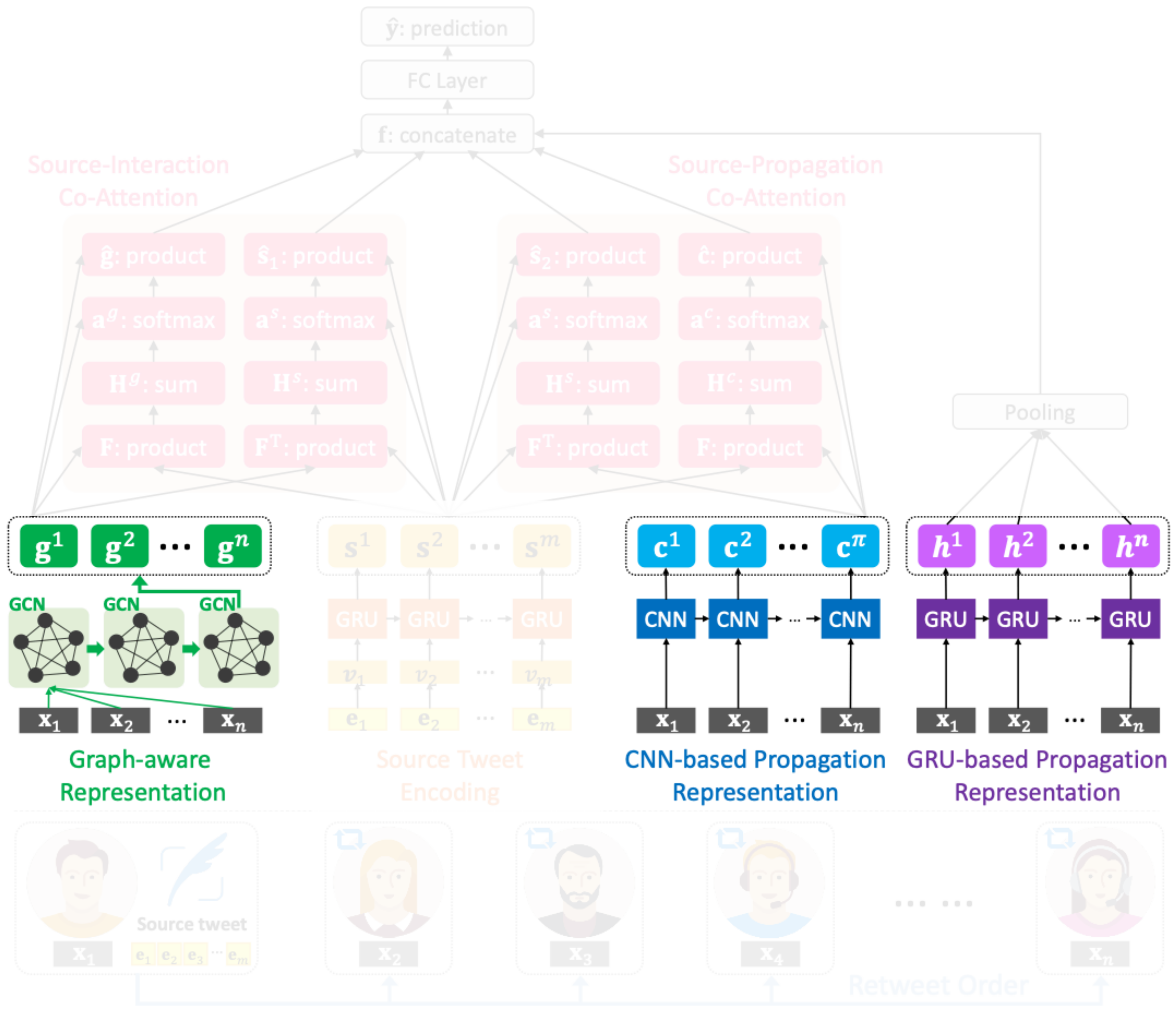
News tweet encoding

- The given source tweet is represented by a word-level encoder.
- The input is one-hot encoding vector of each word in tweet s_i.
- Let $\mathbf{E} = [e_1, e_2, \ldots, e_m] \in \mathbb{R}^m$ be the input vector of source tweet.
- Create a fully-connected layer to generate word embeddings.
    - $\mathbf{V} = \tanh(\mathbf{W}_w\mathbf{E} + \mathbf{b}_w) = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m] \in \mathbb{R}^{d \times m}$
- Then utilize GRU to learn words sequence representation from $\mathbf{V}$.
    - $\mathbf{s}_t = \mathrm{GRU}(\mathbf{v}_\mathrm{t}), \mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m] \in \mathbb{R}^{d \times m}$

Methodology

Graph-aware Co-Attention Networks (GCAN)

- User characteristic extraction
- News tweet encoding
- User propagation representation
- Co-attention mechanism
- Making prediction



Methodology

User propagation representation

- Propagation of s_i is triggered by a sequence of users as time proceeds.
- Aim at exploiting the extracted user feature vectors $\mathbf{x}_j$ to learn user propagation representation.
  - $PF(s_i) = \langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n \rangle$
- Underlying idea is that the user characteristics in real news propagations are difference from those of fake ones.
- Make use of GRU & CNN to learn propagation representation.

Methodology

GRU-based representation

- Given $PF(s_i)$, utilize GRU to learn the propagation representation.
  - $\mathbf{h}_t = \mathrm{GRU}(\mathbf{x}_i), t \in \{1,\ldots,n\}$
- Generate the final GRU-based user propagation embedding by average pooling.
  - $\mathbf{h} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{h}_t \in \mathbb{R}^d$

Methodology

CNN-based representation

- Take advantage of 1D CNN to learn the sequential correlation of user features in $PF(s_i)$.

- Consider λ consecutive users at one time to model their sequential correlation, $\langle \mathbf{x}_t, \ldots, \mathbf{x}_{t+\lambda-1} \rangle$.

  - $\mathbf{C} = \mathrm{ReLU}(\mathbf{W}_f \cdot \mathbf{X}_{t:t+\lambda-1} + b_f)$

Methodology

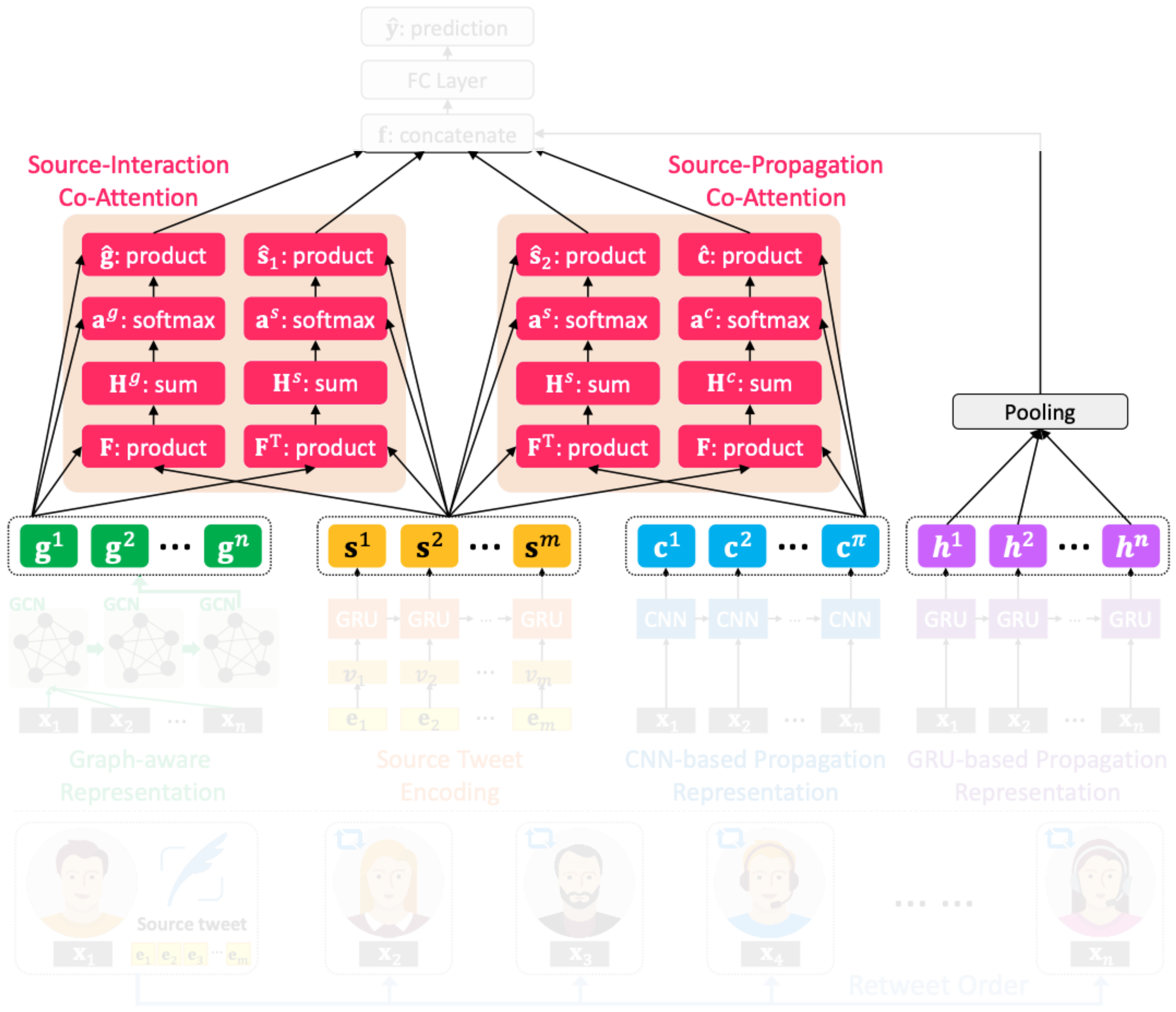
Graph-aware propagation representation

- Aim at creating a graph to model the potential interaction among users who retweet.
- Since true interactions between users are unknown, consider graph is a fully-connected graph.
- To incorporate user features in the graph, each edge is associated with a weight ω.
  - Weight is derived based on cosine similarity between $\mathbf{x}_a, \mathbf{x}_b$. $\omega_{\alpha\beta} = \dfrac{\mathbf{x}_\alpha \cdot \mathbf{x}_\beta}{||\mathbf{x}_\alpha|| \, ||\mathbf{x}_\beta||}$
- Choose to stack two GCN layer to generates embedding vectors of nodes according to their neighborhoods.

Methodology

Graph-aware Co-Attention Networks (GCAN)

- User characteristic extraction
- News tweet encoding
- User propagation representation
- **Co-attention mechanism**
- Making prediction



Methodology
Dual co-attention mechanism

- Develop a dual co-attention mechanism to model the mutual influence
  - between the source tweet and user propagation,
  - between the source tweet and graph-aware interaction embeddings.
- Equipped with co-attention learning, proposed model is capable of the explainability by looking into the attention weights
  - between retweet users in the propagation,
  - words in the source tweet.

Methodology

Source-Interaction co-attention

- First compute a proximity matrix $\mathbf{F} = \tanh(\mathbf{S}^\top \mathbf{W}_{sg} \mathbf{G})$.
- By treating $\mathbf{F}$ as a feature, can learn to predict source and interaction attention maps.
- Generate attention weights of source words and interaction users through the softmax function.
- Generate attention vectors of source tweet and interaction users weighted sum using the derived attention weights.

$$\mathbf{H}^s = \tanh(\mathbf{W}_s \mathbf{S} + (\mathbf{W}_g \mathbf{G})\mathbf{F}^\top) \quad \mathbf{a}^s = \mathrm{softmax}(\mathbf{w}_{hs}^\top \mathbf{H}^s) \quad \hat{\mathbf{s}}_1 = \sum_{i=1}^{m} \mathbf{a}_i^s \mathbf{s}^i,$$

- $$\mathbf{H}^g = \tanh(\mathbf{W}_g \mathbf{G} + (\mathbf{W}_s \mathbf{S})\mathbf{F}) \quad \mathbf{a}^g = \mathrm{softmax}(\mathbf{w}_{hg}^\top \mathbf{H}^g) \quad \hat{\mathbf{g}} = \sum_{j=1}^{n} \mathbf{a}_j^g \mathbf{g}^j$$

Methodology

Source-Propagation co-attention

- The process to generate the co-attention feature vectors $\hat{\mathbf{s}}_2$ for source tweet and $\hat{\mathbf{c}}$ for user propagation.
  - Same as source-interaction co-attention.

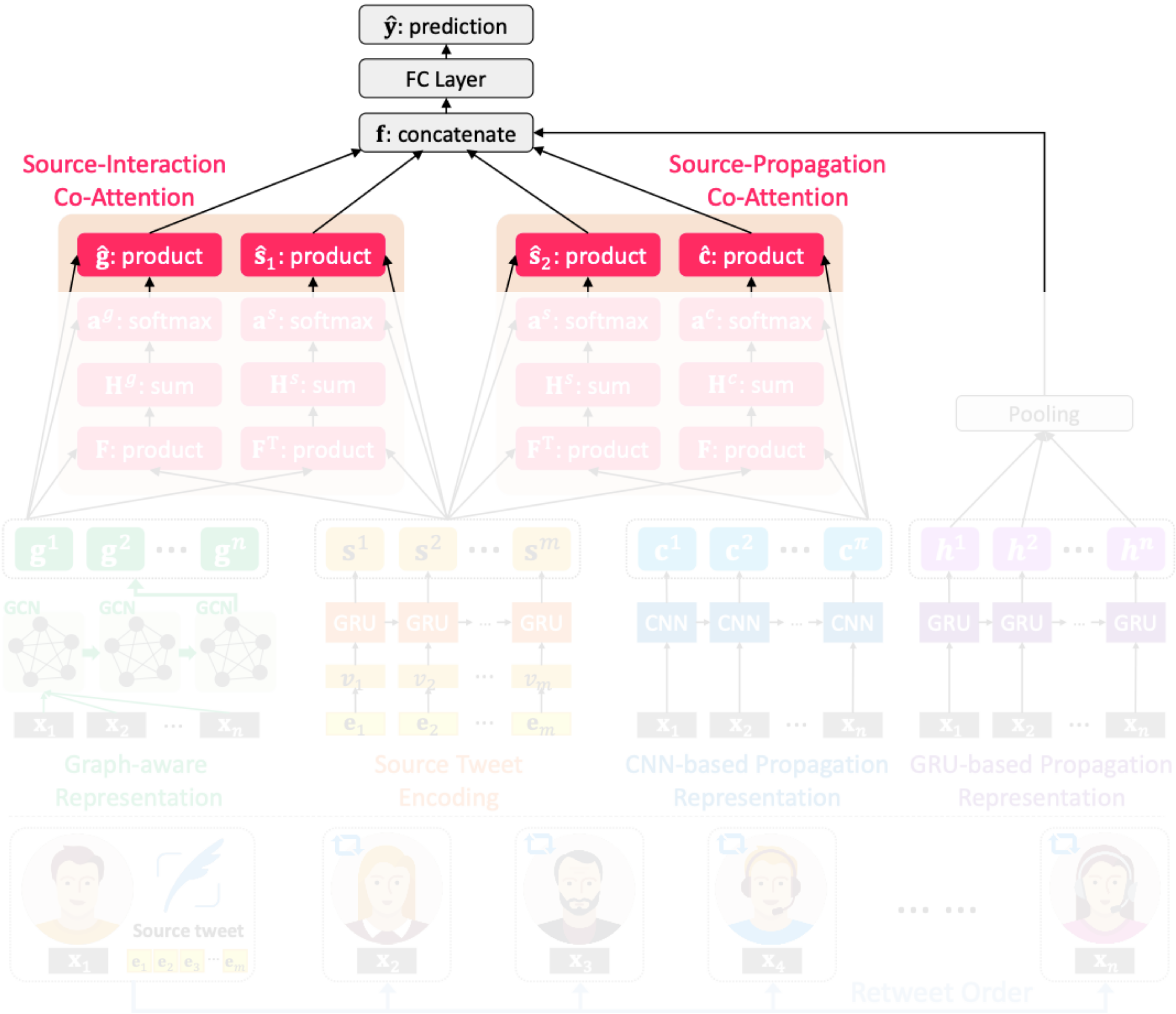
$$\mathbf{H}^s = \tanh(\mathbf{W}_s\mathbf{S} + (\mathbf{W}_c\mathbf{C})\mathbf{F}^\top) \quad \mathbf{a}^s = \mathrm{softmax}(\mathbf{w}_{hs}^\top\mathbf{H}^s) \quad \hat{\mathbf{s}}_2 = \sum_{i=1}^{m} \mathbf{a}_i^s\mathbf{s}^i,$$

- $$\mathbf{H}^c = \tanh(\mathbf{W}_c\mathbf{C} + (\mathbf{W}_s\mathbf{S})\mathbf{F}) \quad \mathbf{a}^c = \mathrm{softmax}(\mathbf{w}_{hc}^\top\mathbf{H}^c) \quad \hat{\mathbf{c}} = \sum_{j=1}^{n} \mathbf{a}_j^c\mathbf{c}^j$$

Methodology

Graph-aware Co-Attention Networks (GCAN)

- User characteristic extraction
- News tweet encoding
- User propagation representation
- Co-attention mechanism
- Making prediction



Methodology

Make prediction

- Fed into a multi-layer feedforward neural network that finally predicts the label.
- Loss function is devised to minimize the cross-entropy value.
  - $\mathbf{f} = [\hat{\mathbf{s}}_1, \hat{\mathbf{g}}, \hat{\mathbf{s}}_2, \hat{\mathbf{c}}, \mathbf{h}]$
  - $\hat{\mathbf{y}} = \text{softmax}(\text{ReLU}(\mathbf{f}\mathbf{W}_f + \mathbf{b}_f))$
  - $\mathscr{L}(\Theta) = -y\log(\hat{y}_1) - (1-y)\log(1-\hat{y}_0)$

Experiments

Datasets

- Twitter15, Twitter16 (MediaEval)
- Each dataset contains a collection of source tweets, along with their corresponding sequences of retweet users.
- Choose only “true” & “fake” labels as the ground truth.
- Since original data does not contain user profiles, use user IDs to crawl user information via Twitter API.
- Train: test = 70: 30

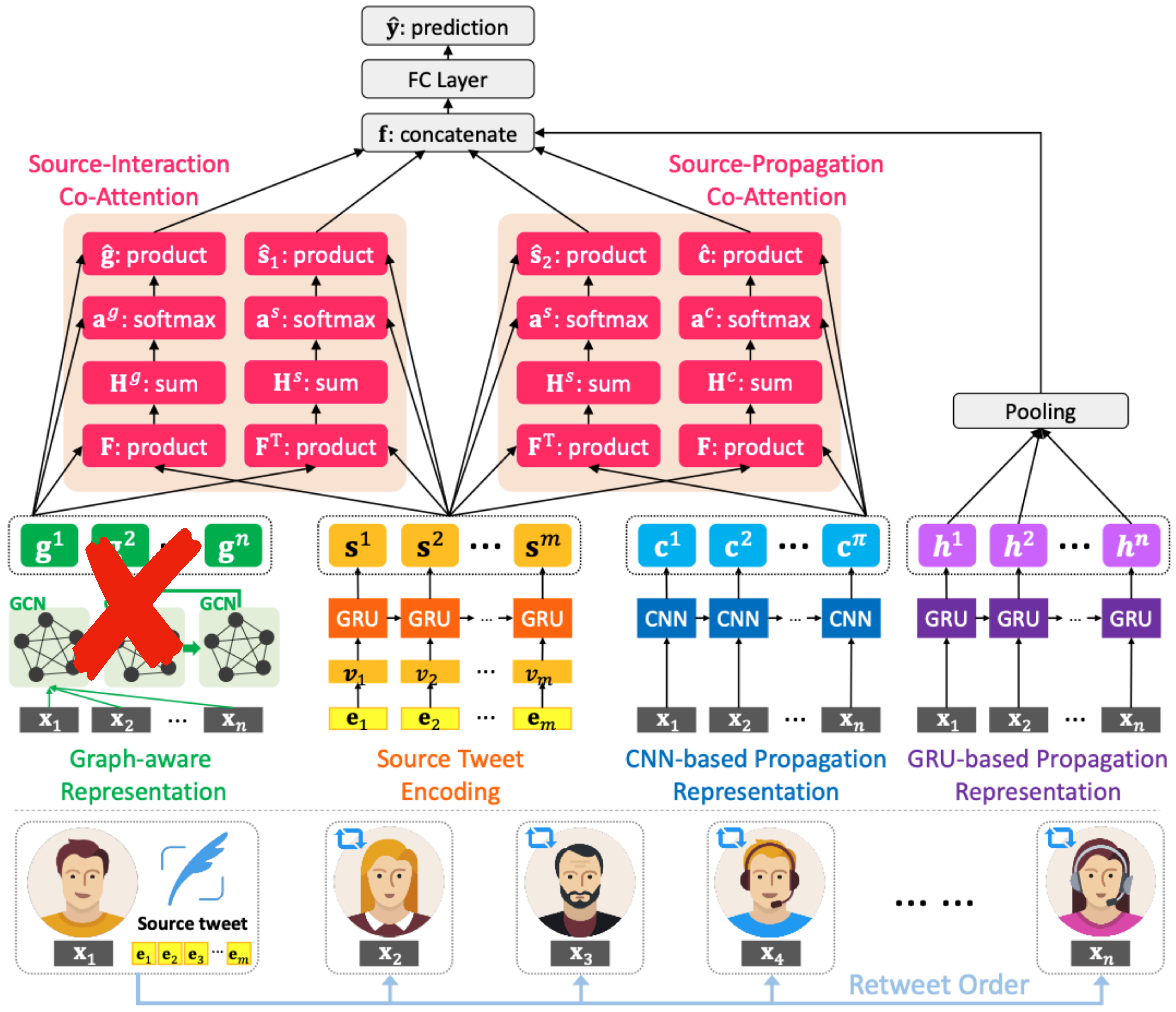
Experiments
Baselines

- DTC: decision tree-based model using user profiles & source tweet.
- SVM-TS: linear SVM using source tweet & sequence of retweet users' profiles.
- mGRU: modified GRU to learn temporal patterns from retweet user profile & source's features.
- RFC: random forest model using retweet user profile & source tweet.
- CSI: SOTA model using articles & group behavior of users who propagate fake news by LSTM.
- tCNN: propose modality-similarity method by caption news image compare with news text content.
- CRNN: SOTA joint CNN & RNN to learn local & global variations of retweet user profiles and resource tweet.
- dEFEND: SOTA co-attention-based model to learn correlation between source tweet's sentences & user profiles.

Experiments

Model configuration

- To examine the effectiveness of graph-aware representation.
  - Create another version GCAN-G, denoting model w/o graph convolution part.



Experiments

Main Result

	Twitter15				Twitter16			
Method	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
DTC	0.4948	0.4806	0.4963	0.4949	0.5616	0.5369	0.5753	0.5612
SVM-TS	0.5190	0.5186	0.5195	0.5195	0.6915	0.6910	0.6928	0.6932
mGRU	0.5104	0.5148	0.5145	0.5547	0.5563	0.5618	0.5603	0.6612
RFC	0.4642	0.5302	0.5718	0.5385	0.6275	0.6587	0.7315	0.6620
tCNN	0.5140	0.5206	0.5199	0.5881	0.6200	0.6262	0.6248	0.7374
CRNN	0.5249	0.5305	0.5296	0.5919	0.6367	0.6433	0.6419	0.7576
CSI	0.7174	0.6867	0.6991	0.6987	0.6304	0.6309	0.6321	0.6612
dEFEND	0.6541	0.6611	0.6584	0.7383	0.6311	0.6384	0.6365	0.7016
GCAN-G	0.7938	0.7990	0.7959	0.8636	0.6754	0.6802	0.6785	0.7939
GCAN	**0.8250**	**0.8295**	**0.8257**	**0.8767**	**0.7593**	**0.7632**	**0.7594**	**0.9084**
Improvement	**15.0%**	**20.8%**	**18.1%**	**18.7%**	**19.3%**	**15.9%**	**3.8%**	**19.9%**

- GCAN is outperform than other methods on both datasets.
  - Even w/o graph-aware part, GCAN-G also improve the best competing method.
- GCAN > GCAN-G, imply some insights.
  - Exhibit usefulness of graph-aware representation.
  - Dual co-attention is powerful, as it clearly outperforms non-co-attention SOTA model CSI.

Experiments

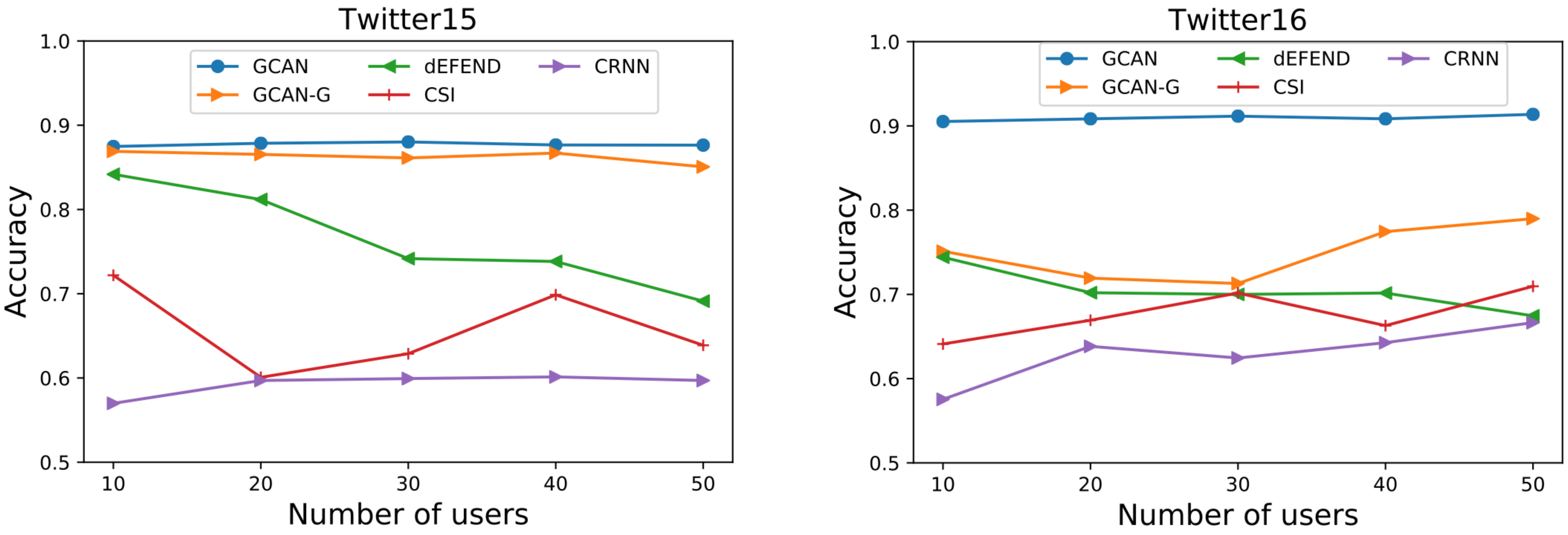
Main Result

	Twitter15				Twitter16			
Method	F1	Rec	Pre	Acc	F1	Rec	Pre	Acc
DTC	0.4948	0.4806	0.4963	0.4949	0.5616	0.5369	0.5753	0.5612
SVM-TS	0.5190	0.5186	0.5195	0.5195	0.6915	0.6910	0.6928	0.6932
mGRU	0.5104	0.5148	0.5145	0.5547	0.5563	0.5618	0.5603	0.6612
RFC	0.4642	0.5302	0.5718	0.5385	0.6275	0.6587	0.7315	0.6620
tCNN	0.5140	0.5206	0.5199	0.5881	0.6200	0.6262	0.6248	0.7374
CRNN	0.5249	0.5305	0.5296	0.5919	0.6367	0.6433	0.6419	0.7576
CSI	0.7174	0.6867	0.6991	0.6987	0.6304	0.6309	0.6321	0.6612
dEFEND	0.6541	0.6611	0.6584	0.7383	0.6311	0.6384	0.6365	0.7016
GCAN-G	0.7938	0.7990	0.7959	0.8636	0.6754	0.6802	0.6785	0.7939
GCAN	**0.8250**	**0.8295**	**0.8257**	**0.8767**	**0.7593**	**0.7632**	**0.7594**	**0.9084**
Improvement	**15.0%**	**20.8%**	**18.1%**	**18.7%**	**19.3%**	**15.9%**	**3.8%**	**19.9%**

- GCAN-G and dEFEND are co-attention-based, additional sequential features learned from the retweet user sequence in GCAN-G can significantly boost the performance.

Experiments

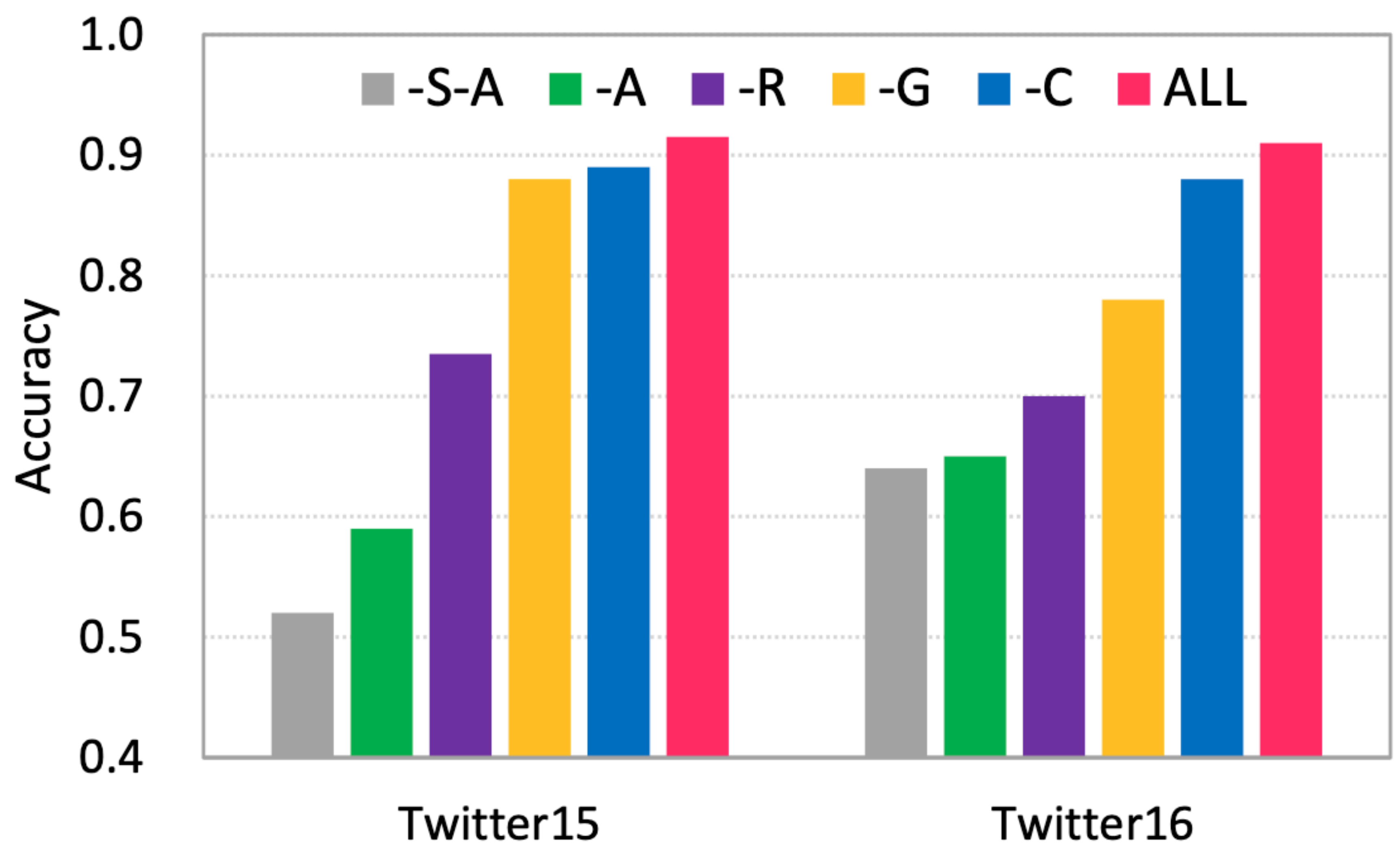
Early detection

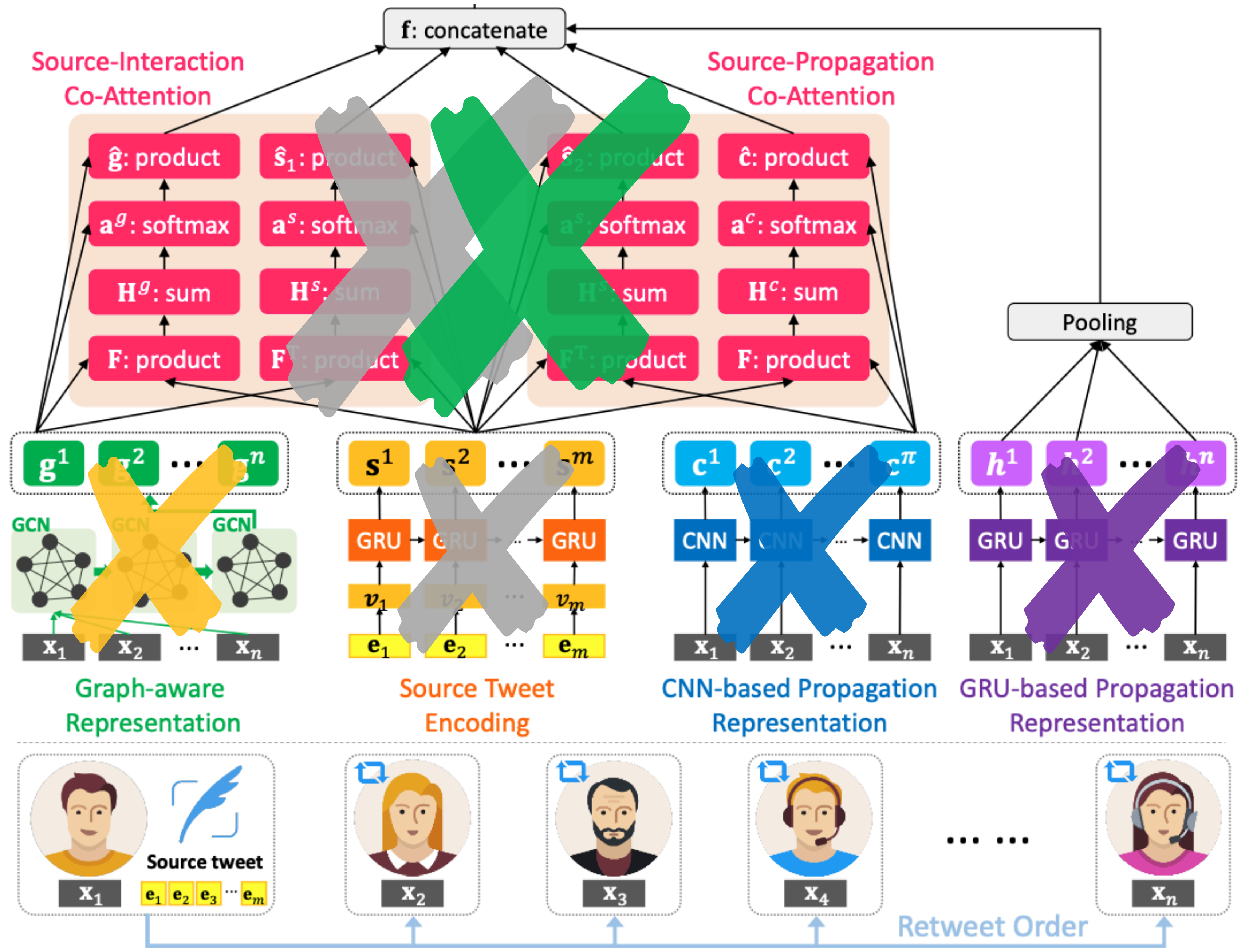


- Further report the performance by varying the number of observed retweet users per source tweet.
- GCAN consistently and significantly outperforms the competitors.
  - Even with only 10 retweeters, GCAN can still achieve 90% accuracy.
- Results tell GCAN is able to generate accurate early detection of the spreading fake news, which is crucial when defending misinformation.

Experiments

Ablation analysis





- Observe every component indeed plays a significant contribution,
  - Especially for dual co-attention ("-A") part.
  - Then representation learning of user propagation and interactions ("-R" and "G").
  - Since the source tweet provides fundamental clues, the accuracy drops significantly without it ("-S-A").

Experiments

GCAN explainability

- The co-attention weights attended on source tweet words and retweet users (source-propagation co-attention) allow GCAN to be capable of explainability.
- By exhibiting where attention weights distribute, evidential words and users in predicting fake news can be revealed.
- Note that do not consider source-interaction co-attention for explainability.
    - Because user interaction features learned from the constructed graph cannot be intuitively interpretable.

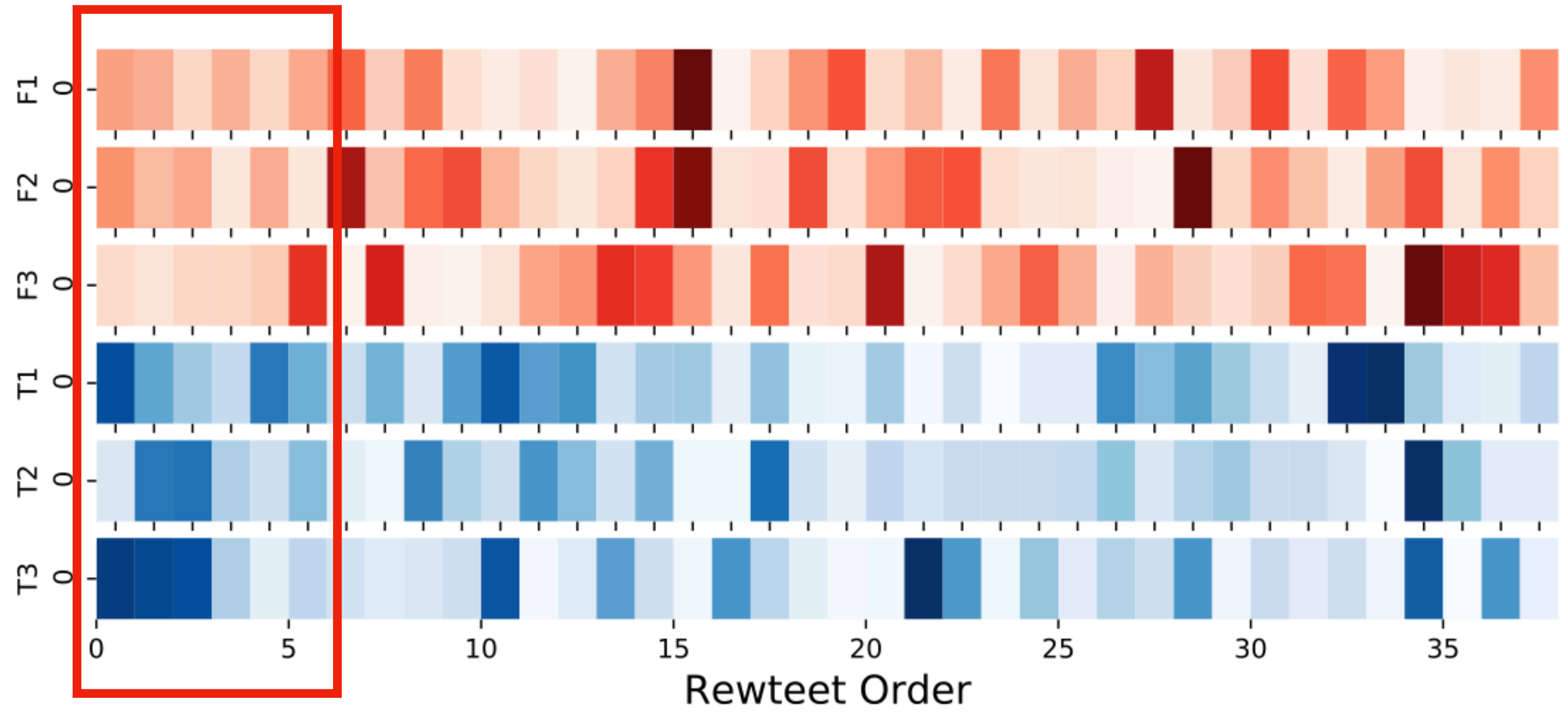
Experiments

Explainability on source words

- Select two source tweets in the test data.
  - Fake (“breaking: ks patient at risk for ebola: in strict isolation at ku med center in kansas city #kwch12”)
  - Real (“confirmed: this is irrelevant. rt @ks-dknews: confirmed: #mike-brown had no criminal record. #Ferguson”)
- Highlight evidential words with higher co-attention weights in font sizes of word clouds.
- Such results may correspond to the common knowledge that fake news tends to use dramatic and obscure words while real news is attended by confirmed and fact checking-related words.

Experiments

Explainability on retweet propagation



- Aim to exploit the retweet order in propagations to unfold the behavior difference between fake and real news.

- Randomly pick three fake (F1-F3) and three true (T1-T3) source stories, and plot their weights from source-propagation co-attention.

- Results show that to determine whether a story is fake, one should first examine the characteristics of users who early retweet the source story.

- Evidences of fake news in terms of user characteristics may be evenly distributed in the propagation.

Experiments

Explainability on retweeter characteristics

- Provide an explanation to unveil the traits of suspicious users and the words they focus on.
- Find that traits of suspicious users in retweet propagation can be:
  - accounts are not verified
  - shorter account creation time
  - shorter user description length
  - shorter graph path length to the user who posts the source tweet.

Source Tweet

Breaking : huge explosion of an #oil pipeline belonging to @saudi_aramco near sudair, #saudiarabia.

Ans: **fake** news

Retweet Propagation

uid	verified	creation time	descpt. length	path to source
		⋮		
14	0	4	7	1
15	0	5	11	1
16	0	6	8	1
		⋮		
32	0	9	17	1
33	0	7	13	2
34	1	9	20	2
		⋮		

highlighted by attention weights on **fake** news (uid 14–16)

highlighted by attention weights on **real** news (uid 32–34)

Experiments

Explainability on retweeter characteristics

- In addition, what they highly attend are words "breaking" and "pipeline."
- Such kind of explanation can benefit interpret the detection of fake news so as to understand their potential stances.

Source Tweet

Breaking : huge explosion of an #oil pipeline belonging to @saudi_aramco near sudair, #saudiarabia.

Ans: **fake** news

Retweet Propagation

uid	verified	creation time	descpt. length	path to source
		⋮		
14	0	4	7	1
15	0	5	11	1
16	0	6	8	1
		⋮		
32	0	9	17	1
33	0	7	13	2
34	1	9	20	2
		⋮		

highlighted by attention weights on **fake** news (rows 14–16)

highlighted by attention weights on **real** news (rows 32–34)

Conclusion

- Proposed a novel FND method, GCAN, which is able to predict whether a short-text tweet is fake, given the sequence of its retweeters.
- Evaluation results show the powerful effectiveness and the reasonable explainability of GCAN.
- Besides, GCAN can also provide early detection of fake news with satisfying performance.
- Besides, while fake news usually targets at some events, authors will also extend GCAN to study how to remove event-specific features to further boost the performance and explainability.

Comments
of GCAN

- Concept of dual co-attention is good.
  - Provide explainable reason.
- Not use tree-based propagation structure indeed reduce the complexity.
- Text-encoding part is out-of-date method.
  - Just use one-hot encoding and utilize GRU to learn word embedding.
- Curious on extract user feature is too simple and normal(?
- All baselines & datasets are too old to cannot known actual performance.