Information diffusion analysis in online social networks based on deep representation learning

Chen, X.

Chapter 8

Conclusions and Future work

In this thesis, we focused on two specific tasks in the field of information diffusion analysis in online social networks, namely, information cascades modeling and rumor detection. In this chapter, we first briefly summarize the main contributions of this thesis, and then conclude by discussing some important future directions.

8.1 Summary of Contributions

In Chapter 1, we stated four different research questions that we now recall so to summarize for each question the contribution of this thesis.

Research Question 1 (RQ 1) *Can we develop an effective deep learning-based model to capture structural and temporal features from the observed cascade graph for macro-level information cascade prediction?*

To the best of our knowledge, the proposed *Recurrent Cascades Convolutional Network* (CasCN) [10] (Chapter 4) is the first neural diffusion model answering this question. Besides that, CasCN also provides an important first step to sample the observed cascade graph as a sequence of subgraphs rather than random walks and diffusion paths. Based on the experimental results and feature visulizations, we find that (1) the way to sample an observed cascade as a sequence of subgraphs is much informative than other node sequence-based sampling methods; (2) CasCN can effectively learn the structural-temporal features from a sequence of subgraphs; (3) CasCN significantly outperform earlier deep learning-based methods for the task of macro-level information cascade prediction.

Research Question 2 (RQ 2) *How can we improve upon earlier deep learning-based models that learn the latent representation for the observed cascade graph from a multi-scale perspective to predict the future size of this cascade?*

We have introduced the *Multi-scale Cascades* (MUCas) model [42] (Chapter 5) to address this question. MUCas abstracts the multi-scale information for cascade graphs as a collection of direction-scale, high-order-scale, position-scale, and dynamic-scale

features, and it uses a multi-scale graph capsule network and an influence attention to learn and fuse the multi-scale information to form a unique cascade representation. This marks an important first step towards solving RQ 2. Different from previous works, MUCas innovatively propose a time interval-based sampling method, which significantly reduces the number of subgraphs, and increases differences between each subgraph as much as possible. Moreover, we find that: (1) each of the scale information is vital for information cascades; (2) aggregate features in a more fine-grained way by introducing routing mechanism can be a more effective and simple way to replace multi-head attention from prior works.

Research Question 3 (RQ 3) *Can we detect rumors at an early stage by learning various diffusion patterns from rumor spreading threads?*

The *Macroscopic and Microscopic-aware Rumor Detection* (MMRD) model [44] (Chapter 6) addresses this research question by two newly designed encoding components MacroE and MicroE, which were used for modeling rumor diffusion from macroscopic and microscopic perspectives, respectively. To the best of our knowledge, this is the first work to solve the rumor detection task by only exploring different levels of diffusion patterns. Also, MMRD innovatively introduces cross-learning and hybrid aggregation mechanisms to improve the model ability in feature learning and feature fusion. Furthermore, MMRD successfully leverages knowledge distillation to increase detection performance, which is a meaningful attempt that can be further considered in future research. We conduct our experiments on two real-world datasets, the experimental results demonstrate MMRD outperforms other state-of-the-art methods and can be applied in an extremely early rumor detection, i.e., with only 10 retweet user observations.

Research Question 4 (RQ 4) *Can we improve the model performance by developing effective rumor detection models at the participant level?*

We have proposed two complementary models to solve the **RQ 4** in Chapter 7: the *Participant-level Rumor Detection* (PLRD) model [45] and the *User-aspect Multi-view Learning with Attention for Rumor Detection* (UMLARD) model [46]. Both of them aim to extract features at the user level rather than the event level. PLRD serves as the first deep learning-based participant-level rumor detection model rooted in social influence and propagation theory, which provides demonstrations of the importance of users in rumor spreading from both theoretical and model performance perspectives. PLRD comprehensively exploits various fine-grained user features from the diffusion threads, i.e., the users' social homophily, influence, susceptibility, temporal features, and then uses these features to determine whether the information is true or false. PLRD also introduces a variational autoencoder (VAE) to handle the uncertainty which exists in the feature learning phase. Moreover, Compared with prior deep learning-based rumor detection models, PLRD can make good predictions only based on user-level features and also provides explainability from both feature-level and user-level. UMLARD solves one burning limitation left by PLRD and

other existing rumor detection methods, i.e., input features entangled with learned high-level features, by using three view-specific embedding methods with distinct inputs. Compared with PLRD, UMLARD innovatively proposes a capsule-based attention layer to replace the original attention mechanism in PLRD, which is more effective from both performance and time cost. UMLARD classifies information into more fine-grained labels, which is rarely considered in existing works because it increases the difficulty in detection.

8.2 Future Work

The research presented in this thesis achieved some interesting results, opens some potential research directions that we leave for future work.

Data collection and processing: Current open benchmark datasets (such as Weibo [9], Twitter [33], etc.) face several challenges. First, due to the strict privacy protection policies in online social platforms, the benchmark datasets can not open access to all resources, such as user profiles, which leads to difficulty in reproducing the same results as the state-of-the-art methods provided. Secondly, the data collection APIs of online social platforms do not provide true retweet paths. In fact, the raw data looks like all retweets point to the original post, which is generally not the case in reality. To deal with this problem, the current method infers the true retweet path based on the follower relationship between all users enrolled in the diffusion process and the timestamps for all retweets, however, this method is time-consuming and biased. Last, the collection of the complete diffusion graph and the social graph is difficult for researchers due to the access limitation of data collection APIs. There is the needs to develop an open-source data collection platform rather than only provide public available datasets, such as FakeNewsNet[1]. To construct the diffusion graph, we can further improve the quality of inferred edges by learning edge uncertainty. And in order to acquire the complete graphs, a corporation with online social platforms is indispensable.

Developing self-supervised and unsupervised model: Most existing efforts on information cascade modeling and rumor detection have been mainly focused on developing supervised and semi-supervised models, both of which require a dataset with sufficient labeled data. However, deciding on a label for a specific message requires a lot of manual labor, especially when it comes to rumor detection. For example, the label of each tweet in Twitter15/16 was confirmed from the fact-checking systems (e.g., Snopes [2], Factcheck [3], etc.). Hence, designing a model that can reduce dependence on labeled data while nevertheless doing well in information diffusion

[1]https://github.com/KaiDMML/FakeNewsNet
[2]https://www.snopes.com/
[3]https://www.factcheck.org/

analysis tasks is an urgent task. Ideally, one could design self-supervised and unsupervised frameworks to encode the information diffusion including structural and temporal features and facilitate the downstream tasks, such as information cascade prediction and rumor detection. Moreover, self-supervised learning always asks for a data augmentation operating, different from the augmentation on images, existing augmentation on graphs, such as edge-delete/add, node-delete/add, etc., which will harm the real diffusion networks and cause information biases. Therefore, how to design a diffusion-aware augmentation method or develop an augmentation-free self-supervised model is another interesting research topic.

Incorporating multi-modal and external information: To improve the model performance for both information cascade modeling and rumor detection, incorporating the knowledge from different aspects is demonstrated to be indispensable in Chapter 7. Apart from the previously used graph, sequence, and text, how to comprehensively extract features from other related sources, such as images, videos, websites, and so on, in a unified framework and efficiently fuse multiple features is an interesting research topic. Incorporating the external information from knowledge graphs in information diffusion prediction or rumor detection can introduce interpretability and inference ability into models, how to combine multi-modal data and external information is absolutely an interesting research direction.

Learning robust embedding for tail-nodes in the graph: The long-tail distribution phenomena can be discovered not only in datasets, such as the label imbalance, but also in graphs, where the majority of nodes are tail-nodes (with small degree) and only a small fraction have a big degree (head-node). Most of the existing works of graph neural networks treat all nodes equally, and do not pay more attention to the difference between tail-nodes and head-nodes, which have limited ability in learning distinguishable and robust embedding for the most vulnerable tail-nodes. As future work, one could borrow the idea from the fields of meta-learning and transfer learning to design a more unified graph neural network for tail-node embedding.

Developing interpretable deep learning-based models: Despite the significant achievements made by employing deep learning methods in information analysis tasks, compared with the traditional hand-crafted feature-based methods, the deep learning-based models do not provide enough interpretability due to their “black-box" models. However, besides looking for improvements of model performance, researchers also want to know the reason behind a message going viral or the intentions behind rumors. Thus, developing interpretable deep learning-based models without significantly sacrificing model performance, is another interesting direction to explore. For example, we plan in future work to keep designing new attention-based and disentangled models, as well as introducing causality into model learning.

.