AutoCite: Multi-Modal Representation Fusion for Contextual Citation Generation

Qingqin Wang[1,2], Yun Xiong[1,2] (✉) , Yao Zhang[1,2], Jiawei Zhang[3], Yangyong Zhu[1,2]
[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai, China
[2]Shanghai Institute for Advanced Communication and Data Science, Fudan University, Shanghai, China
[3]IFM Lab, Department of Computer Science, Florida State University, FL, USA
[1,2] {qqwang18, yunx, yaozhang18, yyzhu}@fudan.edu.cn, [3]jiawei@ifmlab.org

ABSTRACT

Citing comprehensive and correct related work is crucial in academic writing. It can not only support the author's claims but also help readers trace other related research papers. Nowadays, with the rapid increase in the number of scientific literatures, it has become increasingly challenging to search for high-quality citations and write the manuscript. In this paper, we present an automatic writing assistant model, AutoCite, which not only infers potentially related work but also automatically generates the citation context at the same time. Specifically, AutoCite involves a novel multi-modal encoder and a multi-task decoder architecture. Based on the multi-modal inputs, the encoder in AutoCite learns paper representations with both citation network structure and textual contexts. The multi-task decoder in AutoCite couples and jointly learns citation prediction and context generation in a unified manner. To effectively join the encoder and decoder, we introduce a novel representation fusion component, i.e., gated neural fusion, which feeds the multi-modal representation inputs from the encoder and creates outputs for the downstream multi-task decoder adaptively. Extensive experiments on five real-world citation network datasets validate the effectiveness of our model.

KEYWORDS

Multi-Modal Learning; Representation Fusion; Multi-Task Learning; Data Mining

ACM Reference Format:
Qingqin Wang[1,2], Yun Xiong[1,2] (✉) , Yao Zhang[1,2], Jiawei Zhang[3], Yangyong Zhu[1,2]. 2021. AutoCite: Multi-Modal Representation Fusion for Contextual Citation Generation. In *Proceedings of the Fourteenth ACM International Conference on Web Search and Data Mining (WSDM '21), March 8–12, 2021, Virtual Event, Israel*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3437963.3441739

WSDM '21, March 8–12, 2021, Virtual Event, Israel

ACM ISBN 978-1-4503-8297-7/21/03.
https://doi.org/10.1145/3437963.3441739

1 INTRODUCTION

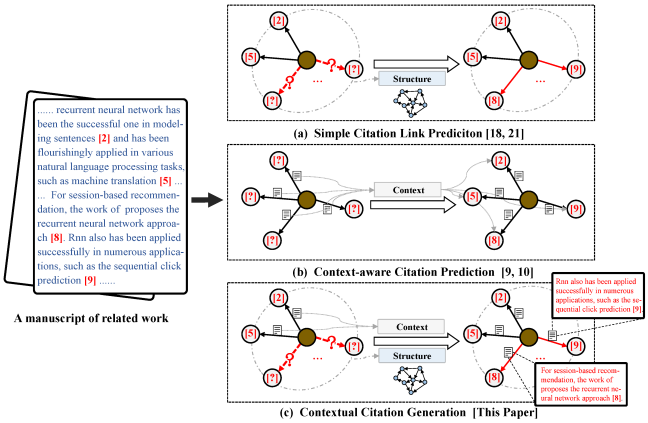


Figure 1: The differences between contextual citation generation and conventional methods. (a) Simple link prediction infers potential citations in citation networks. (b) Context-aware methods take a completed context as a query to predict related papers in marked placeholders '[?]'. (c) Our contextual citation generation captures both network structure and semantic information to predict potential citations (e.g., related work [8], [9]) and generate new contexts simultaneously.

Have you ever had difficulties in finding suitable citations when writing an academic paper? It is crucial to cite correct work in academic writing, which can not only strongly support the claims of the author but also help readers trace other related papers with similar topics. However, with the rapid increase in the number of scientific literatures, it has become increasingly challenging for researchers to search for and cite comprehensive related work nowadays.

Some work proposes to utilize citation networks to help authors find the missing related work [12, 17]. As shown in Figure 1(a), citation recommendation can be regarded as a link prediction task in citation networks. However, most link prediction methods [8, 21] can only capture the network structure but ignore the semantic information of papers. To explore the semantic relevance among papers, extensive context-aware citation prediction methods attempt

to recommend related research for authors [9, 10, 24]. Figure 1(b) shows a paragraph of related work containing unknown citation placeholders marked as '[?]'. To replace these placeholders with appropriate research papers, conventional context-aware methods [9, 10, 24] utilize a specific context that consists of several sentences as the query to predict citations or a suggested list.

Although such methods provide citations that are related to the complete manuscript, replacing only placeholders in a manuscript cannot generate more related work that the author may has missed. Especially without considering the associated sentences generation may not help much in assisting manuscript writing. Some interesting applications attempt to write the academic paper automatically, e.g., generating abstracts based on titles [27] or writing key elements of a new paper based on predicted related entities for input [26]. Due to the huge amount of literatures, even for senior researchers with lots of experiences, it is also difficult to know all relevant research papers and cite them within appropriate contexts [11]. Therefore, besides inferring research papers for the placeholders, generating appropriate sentences (i.e., the citation context) describing the related work tends to be more important. As illustrated in Figure 1(c), in this paper, we introduce a novel research problem to predict potential citations (e.g., related work [8] in Figure 1(c)) that the author may ignore and generate corresponding citation contexts at the same time.

Citation networks consist of nodes representing papers and directed edges representing citation relationships, where each edge is associated with a textual attribute as its citation context. It is observed that papers citing (or cited by) the same citations may be relevant. Such citation relevance is multi-modal, which is revealed in both the similar local topology of network structure and semantic information of textual contexts. As mentioned in [30], descriptive text on edges can encode rich semantic information and structural relationship between networks. It motivates us to exploit both network structure and context semantics in a multi-modal way to generate high-quality citations in this paper. Different from the existing citation prediction tasks [9, 24], we formulate a novel contextual citation generation problem that predicts potential citations and automatically generates citation contexts at the same time. Essentially, it is a kind of link inference task with textual attributes, which aims to infer links and the attributed contexts simultaneously.

However, achieving the goal of contextual citation generation is challenging due to the following reasons:

• **Multi-modal**. In citation networks, the local topology reveals structural characteristics of nodes, and textual contexts on edges encode rich semantic information of nodes. It is essential but challenging to capture complex interaction between information from different modalities and learn multi-modal representations.

• **Multi-task**. Our objective is a multi-task learning problem, which calls for not only citation link prediction but also context generation. Such a multi-task problem needs information from all uni-modality but with specific task-oriented adaption requirements. Thus, how to design an effective fusion strategy to integrate multi-modal representations for downstream tasks adaptively is challenging.

• **Diverse roles**. Since citation links in networks are directed from the outer-citer (papers cite others) to the inner-citer (papers cited by others), it corresponds to the different roles of nodes. For a specific node, the contextual citation pointing to it and from it both reveal its characteristics in various aspects, which need to be captured crucially.

To solve the challenges mentioned above, we propose an automatic writing assistant model AutoCite to address the contextual citation generation task. Specifically, AutoCite involves a novel multi-modal encoder and multi-task decoder architecture, where the encoder learns multi-modal representations of papers with both citation network structure and textual contexts, and the decoder coherently achieves citation link prediction and context generation in a unified manner. To join the encoder and decoder effectively, we introduce a novel gated neural fusion component. It feeds the multi-modal representation inputs and creates outputs for the multiple downstream tasks, which realizes the feature cross of modalities and meet specific task-oriented adaptation requirements adaptively. Furthermore, to deal with diverse roles challenge, the encoder in AutoCite improves attention networks to capture critical characteristics with different roles. Extensive experiments on five real-world datasets validate the effectiveness of our model. Our contributions are summarized as follows:

- We propose a novel multi-modal multi-task learning model for contextual citation generation problem.
- We introduce a gated neural fusion mechanism to integrate multi-modal representations and control features transfer for downstream tasks adaptively.
- Our model captures the characteristic differences of nodes acting in diverse roles. It includes dual-role graph attention for network structure and co-attention for context semantics.
- Extensive experiments are conducted on five real-world datasets, and the results demonstrate the superiority of our model over other state-of-the-art methods.

2 PRELIMINARIES

In this section, we provide some background and formally define the problem of contextual citation generation.

Definition 1. (Contextual Citation Network): We formally define a contextual citation network as a directed graph $G = (\mathcal{V}, \mathcal{E}, f)$, where $\mathcal{V} = \{v_1, v_2, ..., v_n\}$ consists of a set of paper node instances, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ is a set of edges representing the citation relationships between nodes in $\mathcal{V}$. Edge $e_{ij} \in \mathcal{E}$ represents a citation link from node v_i to node v_j. Each edge is associated with a textual citation context $f(e_{ij}) = \mathcal{D}_{ij}$, where $\mathcal{D}_{ij}$ consists of a sequence of words.

For node $v_i \in \mathcal{V}$, it may cite others or be cited by others. We denote the node that cites others as **outer-citer**, and the node cited by others as **inner-citer** conversely.

Problem Definition: The contextual citation generation problem studied in this paper includes predicting citation links and generating corresponding contexts simultaneously. Formally, given a contextual citation network $G = (\mathcal{V}, \mathcal{E}, f)$, we represent unknown links in G as a set $\hat{\mathcal{E}} = \{\mathcal{V} \times \mathcal{V}\}/\mathcal{E}$. Our task is to capture network structure and context semantic information of nodes to predict the probability that links may exist between node-pairs in $\hat{\mathcal{E}}$. More importantly, our model also generates the corresponding context on the citation link when it does exist. For each node-pair (v_i, v_j) in $\hat{\mathcal{E}}$, AutoCite predicts the probability $Pr(v_i, v_j)$ that outer-citer v_i may cite inner-citer v_j, and automatically generates the possible context $\mathcal{D}_{ij}$ when v_i cites v_j.

3 METHODOLOGY

In this section, we will introduce the multi-modal representation learning model in detail. Figure 2 illustrates the overview of AutoCite, which consists of a multi-modal encoder, gated neural fusion, and a multi-task decoder.

3.1 Multi-Modal Encoder

To deeply explore the relevance of nodes in contextual citation networks, we propose a novel multi-modal encoder, including two components, i.e., a graph structure encoder and a textual context encoder.

3.1.1 Graph Structure Encoder. Since the node roles in citation networks are diverse, each node is cooperatively characterized by the information of both other nodes it cites, and other nodes cite it. We propose dual-role graph attention to accurately capture directed graph structure information to learn node representations.

Formally, each node v_i is represented as a structural feature vector $\mathbf{v}_i \in \mathbb{R}^d$. Here, d denotes the dimension of node embeddings. We define the nodes cited by v_i as its outer-neighbors $\mathcal{N}_{v_i}^{\vdash} = \{v_j | (v_i, v_j) \in \mathcal{E}\}$, and nodes cite v_i as its inner-neighbors $\mathcal{N}_{v_i}^{\dashv} = \{v_j | (v_j, v_i) \in \mathcal{E}\}$. Next, we introduce the dual-role graph attention.

Outer-graph attention. For the target outer-citer v_i and its outer-neighbor $v_j \in \mathcal{N}_{v_i}^{\vdash}$, the outer-cite attention score of v_j to v_i is defined as:

$$\alpha_g^{\vdash}(v_i, v_j) = \text{LeakyReLU}\left((\mathbf{a}_g^{\vdash})^T [\mathbf{W}_g^{\vdash}\mathbf{v}_i || \mathbf{W}_g^{\dashv}\mathbf{v}_j]\right), \quad (1)$$

where $\cdot^T$ represents transpose and $||$ is the concatenation operation. $(\mathbf{a}_g^{\vdash})^T \in \mathbb{R}^{2d}$ is a single-layer attention network and subscript $\cdot_g$ denotes "graph" for short. Superscript $\cdot^{\vdash}$ and $\cdot^{\dashv}$ represent the outer-citer and inner-citer roles, respectively. $\mathbf{W}_g^{\vdash}$ (or $\mathbf{W}_g^{\dashv}$) $\in \mathbb{R}^{d \times d}$ is a shared weight matrix for all nodes.

Inner-graph attention. To capture the impact of nodes citing v_i, we also define inner-graph attention. For the target inner-citer v_i and its inner-neighbors $v_j \in \mathcal{N}_{v_i}^{\dashv}$, the inner-cite attention score of v_j to v_i is defined as:

$$\alpha_g^{\dashv}(v_i, v_j) = \text{LeakyReLU}\left((\mathbf{a}_g^{\dashv})^T [\mathbf{W}_g^{\dashv}\mathbf{v}_i || \mathbf{W}_g^{\vdash}\mathbf{v}_j]\right), \quad (2)$$

where notations are similarly defined as above. Then, the attention coefficients across all outer- and inner-neighbors are normalized based on a unified softmax function $\alpha_g^{\cdot}(v_i, v_j) = \text{Softmax}_j(\alpha_g^{\cdot}(v_i, v_j))_{v_j \in \mathcal{N}_{v_i}^{\cdot}}$. We aggregate the outer-neighbor contributions from v_j to outer-citer v_i as a vector:

$$\mathbf{e}_g^{\vdash}(v_i, v_j) = \alpha_g^{\vdash}(v_i, v_j)\mathbf{W}_g^{\vdash}\mathbf{v}_j. \quad (3)$$

Also, the inner-neighbor contributions from v_j to inner-citer v_i is aggregated as a vector $\mathbf{e}_g^{\dashv}(v_i, v_j)$ in the same way. Then, we combine the contributions of all neighbors to v_i. Inspired by [25], we also employ multi-head attention to learn a more stable graph-modal representation:

$$\widehat{\mathbf{e}}_i = \Big\|_{k=1}^{K} \sigma\Big(\sum_{v_j \in \mathcal{N}_{v_i}^{\vdash}} \mathbf{e}_g^{\vdash}(v_i, v_j)^{(k)} + \sum_{v_j \in \mathcal{N}_{v_i}^{\dashv}} \mathbf{e}_g^{\dashv}(v_i, v_j)^{(k)} \Big), \quad (4)$$

where K denotes multiple independent attention mechanisms, and σ is the sigmoid activation. $\mathbf{e}_g^{\cdot}(v_i, v_j)^{(k)}$ are contributions of neighbors computed by the k-th attention $(\mathbf{a}_g^{\cdot})^{(k)}$. The output vector $\widehat{\mathbf{e}}_i \in \mathbb{R}^d$ denotes the graph-modal representation of node v_i. Our graph structure encoder can capture characteristics differences of nodes playing different roles in sub-networks through attention coefficients.

3.1.2 Textual Context Encoder. Except for graph structure characteristics, the textual contexts on edges encode rich semantic information of nodes [30]. For outer-citer v_i, we denote $\{\mathcal{D}_{i1}^{\vdash}, ..., \mathcal{D}_{il}^{\vdash}\}$ as its textual citation contexts that v_i cites others, where l indicates the maximum number of contexts. Equally, $\{\mathcal{D}_{j1}^{\dashv}, ..., \mathcal{D}_{jl}^{\dashv}\}$ represents the textual contexts of inner-citer v_j from citations of others cite v_j.

Each context is encoded into a semantic embedding based on the last hidden state of Bi-directional Long Short-term Memory: $\mathbf{d}_{ij}^{\vdash} = \text{BiLSTM}^{\vdash}(\mathcal{D}_{ij}^{\vdash})$ and $\mathbf{d}_{ij}^{\dashv} = \text{BiLSTM}^{\dashv}(\mathcal{D}_{ij}^{\dashv})$. $\mathbf{d}_{ij}^{\cdot} \in \mathbb{R}^d$ where d is the same dimension as node embeddings. After that, we get the corresponding semantic embeddings $\mathbf{D}_{v_i}^{\vdash} = [\mathbf{d}_{i1}^{\vdash}, ..., \mathbf{d}_{il}^{\vdash}]$ for outer-citer v_i and $\mathbf{D}_{v_j}^{\dashv} = [\mathbf{d}_{j1}^{\dashv}, ..., \mathbf{d}_{jl}^{\dashv}]$ for inner-citer v_j, respectively. Since each context has different priorities to characterize the semantic relevance between nodes, we apply co-attention to identify the more important context. Formally, the (r, c)-th entry in the co-attention weight matrix $\mathbf{M} \in \mathbb{R}^{l \times l}$ is defined as:

$$\mathbf{M}_{r,c} = \left(\mathcal{F}^{\vdash}(\mathbf{d}_{ir}^{\vdash})\right)^T \mathbf{W}_{co} \left(\mathcal{F}^{\dashv}(\mathbf{d}_{jc}^{\dashv})\right), \quad (5)$$

where $\mathbf{W}_{co} \in \mathbb{R}^{l \times l}$ is a weight matrix and the subscript $\cdot_{co}$ denotes "context" for short. $\mathcal{F}^{\cdot}$ are feed-forward neural networks. By taking the row- and column-wise maximum sum of matrix $\mathbf{M}$, the co-attention is able to select the contexts with maximum correlations:

$$\mathbf{m}_{co}^{\vdash}(v_i, v_j) = \mathbf{M}\mathbf{u}^{\vdash} \quad \text{and} \quad \mathbf{m}_{co}^{\dashv}(v_j, v_i) = (\mathbf{u}^{\dashv})^T\mathbf{M}, \quad (6)$$

where $\mathbf{u}^{\vdash}$ (or $\mathbf{u}^{\dashv}$)$\in \mathbb{R}^l$ is a constant vector with all values of 1. $\mathbf{m}_{co}^{\vdash}(v_i, v_j)$ (or $\mathbf{m}_{co}^{\dashv}(v_j, v_i)$) $\in \mathbb{R}^l$ indicates the importance distribution of $\mathbf{D}_{v_i}^{\vdash}$ (or $\mathbf{D}_{v_j}^{\dashv}$). Then, we utilize a softmax function to normalize the importances and compute the context-modal representations:

$$\widetilde{\mathbf{e}}_i = \left(\mathbf{m}_{co}^{\vdash}(v_i, v_j)\right)^T \mathbf{D}_{v_i}^{\vdash} \quad \text{and} \quad \widetilde{\mathbf{e}}_j = \left(\mathbf{m}_{co}^{\dashv}(v_j, v_i)\right)^T \mathbf{D}_{v_j}^{\dashv}. \quad (7)$$

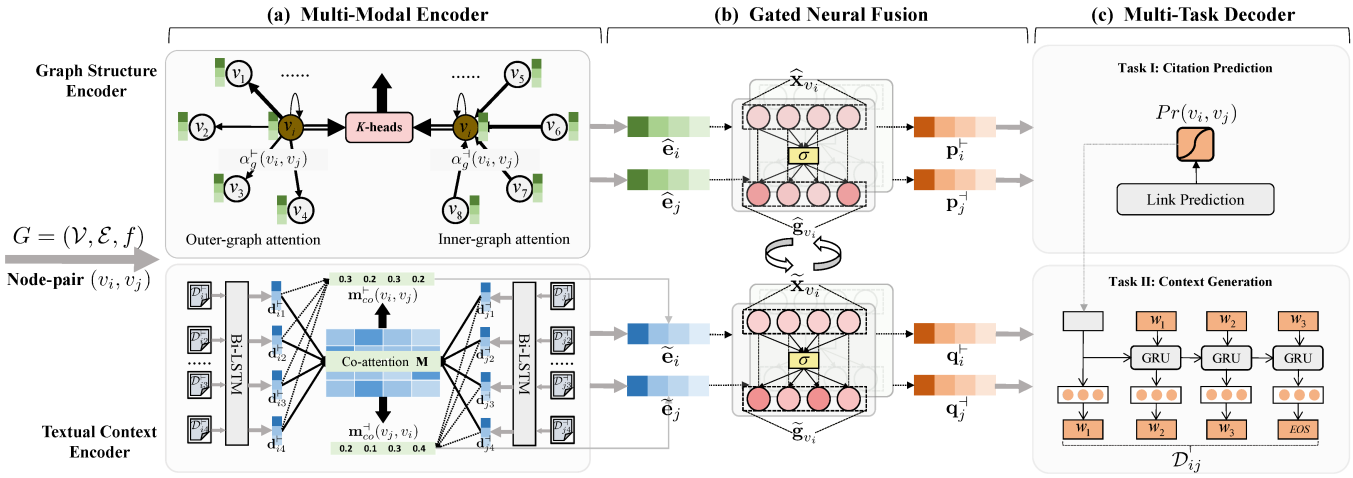


Figure 2: The framework of AutoCite. Given the contextual citation network G and a node-pair (v_i, v_j), we aim to predict the citation from outer-citer v_i to inner-citer v_j. (a) The multi-modal encoder first captures both the network structure and semantic information of v_i and v_j. (b) Then, a gated neural fusion integrates the multi-modal representation for different tasks adaptively. (c) Finally, the multi-task decoder predicts the probability of the citation link that exists and generates the corresponding context.

Here, $\widetilde{\mathbf{e}}_i \in \mathbb{R}^d$ and $\widetilde{\mathbf{e}}_j \in \mathbb{R}^d$ denote the context-modal representation of outer-citer v_i and inner-citer v_j, respectively. The textual context encoder is capable of capturing the semantic differences in the context of different roles.

3.2 Gated Neural Fusion

To solve auxiliary tasks, the information of different modalities should be integrated into a compact multi-modal representation [32]. Most existing fusion methods [19, 29] are oriented to a single task, ignoring the inherent feature cross between modalities and tasks, and cannot be directly applied to our model.

For the complex multi-task problem, we need to capture the multi-modal information differences for specific task-oriented information adaption requirements. Therefore, we propose a novel gated neural fusion, where the gates can control the features cross between modalities and help specific tasks capture critical information adaptively. Take outer-citer v_i as an example, for the citation link prediction task, the multi-modal representation of v_i is formally defined as:

$$\widehat{\mathbf{g}}_{v_i} = \sigma(\widehat{\mathbf{x}}_{v_i}), \tag{8}$$

$$\mathbf{p}_i = \widehat{\mathbf{g}}_{v_i} \otimes \widehat{\mathbf{e}}_i + (1 - \widehat{\mathbf{g}}_{v_i}) \otimes \widetilde{\mathbf{e}}_i, \tag{9}$$

where $\widehat{\mathbf{x}}_{v_i} \in \mathbb{R}^d$ is the node-independent link prediction oriented neural parameter of v_i to be learned. $\widehat{\mathbf{g}}_{v_i}$ denotes the task-driven gate, which controls the weight of uni-modal representation by bit-wise granularity. $\otimes$ is hadamard product operation and $\mathbf{p}_i \in \mathbb{R}^d$ indicates the multi-modal representation for node v_i. For context generation task, the multi-modal representation of v_i is formally defined as:

$$\widetilde{\mathbf{g}}_{v_i} = \sigma(\widetilde{\mathbf{x}}_{v_i}), \tag{10}$$

$$\mathbf{q}_i = (1 - \widetilde{\mathbf{g}}_{v_i}) \otimes \widehat{\mathbf{e}}_i + \widetilde{\mathbf{g}}_{v_i} \otimes \widetilde{\mathbf{e}}_i, \tag{11}$$

where $\widetilde{\mathbf{x}}_{v_i} \in \mathbb{R}^d$ is node-independent context generation oriented parameter of v_i, and $\widetilde{\mathbf{g}}_{v_i}$ controls the weight of uni-modal representation for context generation task adaptively.

The proposed gated neural fusion realizes either intra- or inter-modality feature cross, combining multiple information into a multi-modal representation without requiring manual tuning. Such task-driven fusion couples multiple tasks tightly, capturing critical information for specific task-oriented adaption requirements.

3.3 Multi-Task Decoder

To realize the citation link prediction and context generation simultaneously, we propose a multi-task decoder. Through gated neural fusion, we integrate four types of multi-task oriented multi-modal representations: $\mathbf{p}_i^{\vdash}$ and $\mathbf{q}_i^{\vdash}$ for outer-citer v_i, $\mathbf{p}_j^{\dashv}$ and $\mathbf{q}_j^{\dashv}$ for inner-citer v_j.

Citation Link Prediction. For node-pair (v_i, v_j), we define the probability of citation link from outer-citer v_i to inner-citer v_j that exists as:

$$Pr(v_i, v_j) = \sigma(\mathbf{p}_i^{\vdash} \cdot \mathbf{p}_j^{\dashv}). \tag{12}$$

Following previous methods [5, 18], we train the citation link prediction model with negative sampling, and the objective is defined as:

$$\mathcal{L}_g = \sum_{(v_i, v_j) \in \mathcal{E}} \Big(\log Pr(v_i, v_j) + \sum_{(\hat{v}_i, \hat{v}_j) \in \mathcal{N}} \log\left(1 - Pr(\hat{v}_i, \hat{v}_j)\right) \Big), \tag{13}$$

where $\mathcal{N}$ is the set of negative samples. Specifically, we randomly replace v_i (or v_j) in positive node-pair (v_i, v_j) among other nodes in $\mathcal{V}$ to construct the negative set.

Context Generation. We adopt the frequently used deep neural language model Gated Recurrent Unit (GRU) [6] to

generate citation contexts. For the citation from v_i to v_j, $\mathbf{q}_i^{\vdash}$ and $\mathbf{q}_j^{\dashv}$ are incorporated into the initial hidden state $\mathbf{h}_0$ first:

$$\mathbf{h}_0 = \tanh(\mathbf{W}_0^{\vdash}\mathbf{q}_i^{\vdash} + \mathbf{W}_0^{\dashv}\mathbf{q}_j^{\dashv} + \mathbf{b}_0). \tag{14}$$

Here, tanh is the activation function. $\mathbf{W}_0^{\cdot} \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_0 \in \mathbb{R}^d$ are parameters to be learned. The hidden state $\mathbf{h}_t$ at time t is calculated recursively:

$$\mathbf{h}_t = \mathrm{GRU}(\mathbf{h}_{t-1}, \mathbf{w}_t), \tag{15}$$

where $\mathbf{w}_t \in \mathbb{R}^d$ is the word embedding generated at time t. At each step, $\mathbf{h}_t$ is transformed into word probability distribution:

$$\mathbf{o}_t = \mathrm{Softmax}(\mathbf{W}_o\mathbf{h}_t + \mathbf{b}_o), \tag{16}$$

where $\mathbf{W}_o \in \mathbb{R}^{|V|\times d}$ and $\mathbf{b}_o \in \mathbb{R}^{|V|}$ are weight matrices, $|V|$ is the vocabulary size. $\mathbf{o}_t \in \mathbb{R}^{|V|}$ indicates the probability of words being selected as the generated word at time t. Beam search [26] is applied to generate a reasonable context $\mathcal{D}_{ij}$. Finally, the objective of context generation is defined as:

$$\mathcal{L}_{co} = \sum_{\mathcal{D}_{ij}} \Big(\sum_{t} -\log(\mathbf{o}_{t,w_t}) \Big). \tag{17}$$

The multi-modal encoder in AutoCite is sharing by different downstream tasks. We jointly train the multi-task objectives into a unified framework:

$$\mathcal{L}(\Theta) = \mathcal{L}_{co} + \lambda\mathcal{L}_g + \gamma\mathcal{L}_{reg}, \tag{18}$$

where Θ contains all parameters of AutoCite, and λ denotes the weight to balance the priorities of different losses. We also apply L_2 regularization $\mathcal{L}_{reg}$ of the model with parameter γ to avoid overfitting. By coupling citation link prediction and context generation through parameter sharing, both the two tasks can learn from each other effectively.

Table 1: Statistics of five public datasets.

	# Papers	# Citations	# Words
aan	7,019	11,911	77
peer	4,872	16,790	62
cora-pro	36,743	64,036	52
cora-os	25,605	45,014	53
cora-db	13,320	21,462	50

4 EXPERIMENTS

To investigate the effectiveness of AutoCite, we compare it with baselines on real-world citation network datasets.

4.1 Dataset

In our experiment, we have selected three representative publicly citation datasets as follows.

- **ACL Anthology Network** [1] provides extracted citations and collaboration networks of papers and authors. We utilize the citation network **aan** [12] with contexts constructed by Jeong et al.
- **PeerRead** [2] consists of over 14K paper drafts submitted in the artificial intelligence field. We use the public dataset **peer** [12] that includes citation contexts.
- **Cora** [3] contains numerous papers of seven categories, where the edges with context between nodes represent citations. We choose three categories datasets: programming (**pro**), operation systems (**os**), and database (**db**) according to different sizes and themes.

For each dataset, we intercept the sentences before and after each citation mark in raw manuscripts as its textual context. The statistics of the five datasets are summarized in Table 1.

4.2 Comparison Methods

To evaluate the performance of **citation link prediction**, we compare AutoCite with the following methods:

- **CTM** [10] models the citations and contexts as parallel data, through translation to predict citation links.
- **NMF** [18] learns the probability of citation link existing between node-pairs based on matrix factorization.
- **DeepWalk** [21] captures node-pairs via uniform random walks on a graph to learn node representations.
- **Node2Vec** [8] modifies a biased random walk strategy to explore the neighborhood architecture efficiently.
- **GAT** [25] utilizes graph neural networks to learn embeddings by leveraging contributions of neighbors.
- **CAML** [5] is an explainable recommendation that predicts the score and generates the user's review. We replace its score prediction to citation link prediction.

To evaluate the performance of **context generation**, we compare AutoCite with the following methods:

- **WordRNN** [23] builds only a word-level language generation model but ignoring the network structure.
- **Net2Text** [30] generates contexts for related papers conditioned on both words and network structure.
- **Seq2Seq** [2] is an encoder-decoder framework that translates the source context into a matching context.
- **CAML** [5] generates contexts based on the predicted score and nodes' textual contexts. It achieves both citation link prediction and context generation.

4.3 Experimental Settings

For citation link prediction, we regard it as a binary classification task, to judge whether there is a link between node-pairs. All existing links form the positive links set, and we randomly replace each one of the nodes in positive links to construct a negative set. Following previous works [8, 21], we use the Area Under ROC Curve (AUC) metric to evaluate the performance. For context generation, the model outputs the probability of $|V|$ words that may be selected at each time step. The generation process will stop when the generated context reaches the threshold length of T or when encountering a terminator. The widely used metric BLEU [2, 26, 30]

[1] http://aan.eecs.umich.edu

[2] https://github.com/allenai/PeerRead

[3] http://www.research.whizbang.com/data

Table 2: Performance comparison for citation link prediction on five datasets.

Datasets	Baselines						Our Methods	
	CTM	NMF	DeepWalk	Node2Vec	GAT	CAML	AutoCite-P	AutoCite
aan	70.30	83.28	86.70	87.52	89.25	86.94	91.94	**92.25**
peer	69.34	82.47	85.12	84.89	88.26	82.48	92.45	**92.74**
cora-pro	70.65	76.04	87.69	87.15	88.73	80.89	90.06	**91.76**
cora-os	69.07	75.20	83.61	84.01	86.34	80.54	89.46	**90.64**
cora-db	68.15	74.88	85.28	86.31	87.67	77.61	88.25	**89.23**

Table 3: Performance comparison for context generation on five datasets.

Datasets	Metrics	Baselines				Our Methods	
		WordRNN	Net2Text	Seq2Seq	CAML	AutoCite-G	AutoCite
aan	BLEU-1/2	23.8/19.8	33.2/25.8	49.6/39.5	53.5/44.2	56.2/44.7	**56.3**/**45.2**
	BLEU-3/4	16.0/10.6	16.7/12.2	30.8/21.8	33.5/24.3	33.1/24.8	**33.3**/**25.1**
peer	BLEU-1/2	20.6/17.9	31.2/24.0	45.9/36.6	52.7/41.5	53.7/42.5	**54.3**/**43.0**
	BLEU-3/4	13.3/9.19	17.4/13.9	23.4/19.2	29.4/21.3	30.6/22.0	**33.8**/**22.7**
cora-pro	BLEU-1/2	24.6/18.6	32.3/24.5	48.6/37.0	50.3/36.2	53.1/41.1	**53.2**/**41.4**
	BLEU-3/4	13.5/10.5	16.7/12.0	26.3/19.3	25.4/18.5	28.9/20.6	**21.2**/**20.8**
cora-os	BLEU-1/2	28.1/21.8	33.6/27.0	47.7/37.0	50.0/44.4	54.2/42.8	**54.9**/**42.3**
	BLEU-3/4	15.6/11.2	19.0/14.3	26.5/19.3	29.3/22.1	30.4/21.6	**30.9**/**22.3**
cora-db	BLEU-1/2	29.6/23.1	32.8/24.1	48.9/38.5	52.8/43.5	**56.9**/**45.2**	56.5/44.4
	BLEU-3/4	15.7/10.6	17.7/12.8	28.0/20.7	31.0/23.2	**32.3**/22.9	31.8/**23.5**

is applied to measure the similarity between the generated context and the ground-truth. We compute BLEU-1/2/3/4 to evaluate the performance in different granularities.

For all baselines above, we have chosen the optimal hyper-parameters carefully or follow the original settings. We randomly choose 80% of samples as the training data and remaining as test data. The word(or node) embeddings are randomly initialized with dimension $d = 128$. We uniformly sample l textual contexts of each node, and the value of l is tested in [1,2,3,4]. The graph attention head $K = 4$, learning rate lr = 3e-3, coefficient γ = 5e-4 of L_2 regularization, and probability $p = 0.4$ of dropout for all parameters. The harmonic weight λ of loss function is set to 0.05, and the window size of beam search is set to 2. During training, we train for 50 epochs using Adam optimizer [15]. We will discuss how the key hyper-parameters affect performance in section 4.6. The source code is available at https://github.com/qqingwang/AutoCite.

4.4 Experimental Results

Table 2 and Table 3 respectively show the results of citation link prediction and context generation on five datasets, where the best results are boldfaced.

In Table 2, AutoCite-P means we train only a single link prediction task and remove the context generation loss. In general, our model has achieved the best performance on all datasets, and the variant of a single task decoder gets sub-optimal results. It shows that the multi-task learning decoder helps to improve both sub-tasks. NMF obtains better results than CTM in all cases. CAML integrates context information into matrix factorization and achieves a higher AUC. Still, none of the above methods capture the citation network structure, and they perform poorly than other methods based on network structure, e.g., Deepwalk, Node2vec and GAT. Among all the structure-based methods, GAT gets the best results on all datasets, only inferior to our model. This finding suggests that our dual-role graph attention can improve link prediction by exploring the characteristics differences of direction.

For context generation, the results are presented in Table 3, where AutoCite-G means that we consider only the context generation task. Our model explores multi-modal representations and has achieved a significant improvement over baseline methods. Among all the baselines, WordRNN generates context based on only a few words but ignoring the network structure and gets the worst performance. Net2Text captures the network structure and achieves slightly better results. However, Net2Text only considers the network structure and performs poorly than other models that integrate context information, e.g., Seq2Seq and CAML. It shows that semantic information of nodes is critical for generating contexts. In most cases, the performance of the multi-task decoder AutoCite exceeds the single-task decoder AutoCite-G. Experimental results confirm our hypothesis: the crossing of multi-modal representations and the sharing of information between multiple tasks are beneficial.

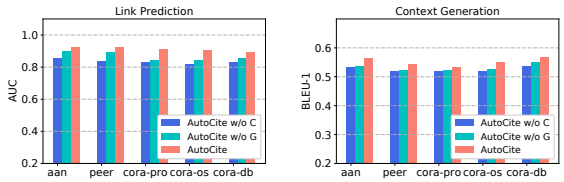


Figure 3: Influence of multi-modal representation.

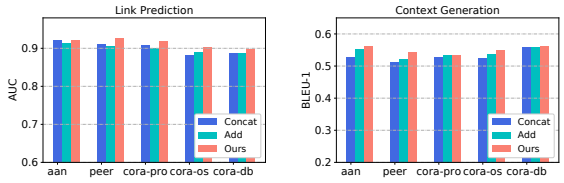


Figure 4: Effectiveness of gated neural fusion.

4.5 Study of Different Variants

Influence of multi-modal representation. We have conducted an ablation study to verify the effectiveness of multi-modal representation learning. As shown in Figure 3, "w/o C" denotes that we remove context-modal representations, and "w/o G" means the graph-modal representations have been removed. In both tasks, "w/o C" gets the worst effect, while "w/o G" achieves better results. The results indicate that semantic information may be more important, and the best results have been attained by considering both multi-modal information in network structure and contexts.

Influence of diverse roles. We have designed comparative experiments to verify the effectiveness of diverse roles in the encoder. As shown in Table 4, "g-w/o" indicates that we replace the dual-role graph attention with a vanilla graph attention, and "t-w/o" means concatenating the semantic embeddings of nodes in context encoder. The results show that without considering the diverse roles of nodes in capturing network structure or semantic contexts will affect both tasks' performance. Besides, "t-w/o" achieves the best context generation on the cora-db dataset, but the proposed model gets the best performance in all other cases. Thus, it is necessary to capture the role diversity of nodes.

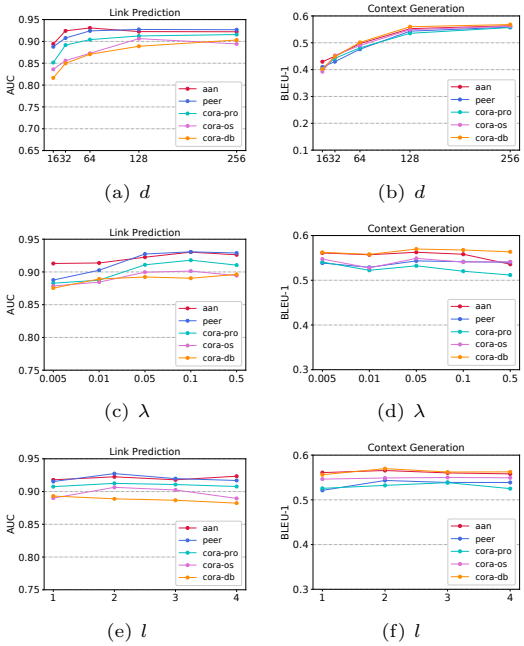
Effectiveness of gated neural fusion. To effectively join the encoder and decoder, we introduce the gated neural fusion to integrate multi-modal representations. Figure 4 shows the results with other widely used fusion methods. Concat [29] represents concatenation with a linear network, and Add [19] replaces it with an addition operation. Our gated neural fusion allies gates to control the contribution of uni-modality for multiple downstream tasks. In most cases, our gated neural fusion has achieved better and robust results. It suggests a simple but effective feature fusion method for multi-task that can be used in many other fields.

4.6 Parameters Analysis

In this section, we evaluate how different values of dimension size d, weight λ of link prediction, and sample number l of contexts affect the performance, while other parameters are fixed. Figure 5 shows the sensitivity analysis of parameters on five datasets. The first two subfigures present the results with different dimension sizes d. With the increase of d, the model gets better performance. After d increases to 64, the result of link prediction is relatively stable, while the result of context generation is still changing. Overall, our model can achieve a decent performance when $d = 128$. λ controls the harmonic weight of two tasks. It is observed that when λ is near 0.05 to 0.1, both citation link prediction and context generation are relatively stable and achieve better results. In general, as λ increases, the context generation task is mildly affected, but the citation link prediction is still stable. For the number l of sample contexts, the optimal value is in [2, 3] according to the last two subfigures. The change of l has a slight impact on the link prediction task but may stabilize the context generation task, which indicates the semantic information is important. Besides, since the cora-db dataset is sparse and the learned node representations may be smooth, which leads to the unstable performance. As the results show, although the above parameters have a slice of influence on results, they are still robust.

Table 4: Influence of diverse roles.

	Link Prediction			Context Generation		
Datasets	g-w/o	t-w/o	with	g-w/o	t-w/o	with
aan	92.68	90.02	**92.25**	55.8	52.2	**56.4**
peer	92.03	91.54	**92.26**	52.1	51.9	**54.3**
cora-pro	89.85	91.36	**91.83**	52.3	52.1	**53.2**
cora-os	87.44	89.12	**90.34**	53.7	54.0	**54.9**
cora-db	87.77	89.61	**89.67**	54.6	**56.3**	56.2



(a) d (b) d

(c) λ (d) λ

(e) l (f) l

Figure 5: Hyperparameter analysis.

Table 5: Comparison of generated contexts with ground-truth.

Case 1	Truth	In particular, convolutional neural networks and recurrent neural networks can efficiently capture the sequentiality of texts. And these methods are typically applied directly distributed embedding words or characters without any knowledge.
	AutoCite	To the first language model that translates texts' character level representations, convolutional neural networks have been proven capable of representing any various tasks such as syntactic analysis and sentence parse.
Case 2	Truth	Techniques modeling and analysis based classical algebraic topology conjunction with distributed simulation methods have brought about significant progress in our understanding computability problems asynchronous distributed setting.
	AutoCite	Regarding the algorithm of distributed agreement problem and mutual exclusion problem, many mathematical algorithms maintain the ability to interact with potentially equivalent provide a solution.
Case 3	Truth	This work has been directed towards parsing languages that allow specification pretty printing are rare, and they arise generators for software engineering environments the ergo support the programming system generator and the synthesizer generator.
	AutoCite	Ceres uses language flowcharts with the adl language, which first defines functional programming and functional programming methodology.

4.7 Case Study

To better understand the meaning of AutoCite, we present some contexts generated by it. As shown in Table 5, we randomly sampled three real cases from different datasets. "Truth" represents the ground-truth context, and red words indicate highly semantically related text.

Case 1 is sampled from the peer dataset, where the ground-truth context mainly introduces that neural networks can model the sequentiality of texts. In our generated context, AutoCite correctly predicts the same keywords (e.g., convolutional neural networks) and semantically related words (e.g., character level representations) as the original text. In case 2, the ground-truth context describes previous methods of techniques modeling and analysis. It is inferred that the citation may be a classic piece of related work. Although the conjunction of techniques is not described accurately, the output context captures the key point "distributed" skillfully. Next, the generated sentence in case 3 looks short in length. Still, the context also accurately provides important information, e.g., "language" and "programming". Moreover, both the generated and the original context focus on the same topic: "programming language". Through the analysis of the above real cases, our model can generate reasonable contexts. AutoCite can recommend comprehensive related work to authors, and help them write persuasive academic articles.

5 RELATED WORK

Link prediction is a widely used task that was applied in many scenarios[1, 13, 16, 28]. Link prediction in networks was usually based on node embeddings. Conventional graph representation learning methods [8, 21, 25] aimed to learn low-dimensional dense vectors for each node while preserving the original graph structure information. Such methods solve only the link prediction problem in simple networks, which cannot be directly applied to our task. For citation prediction, some works aimed at a given complete manuscript to provide a list of other references [14, 17, 22], and other research utilized a specific context as the query to generate short suggested lists [9, 24]. However, all of these methods only predicted citations without generating context.

Natural language generation models were widely used in machine translation [2], text summarization [3], and speech recognition [7]. Feed-forward Neural Network [4] was the first language model based on deep learning, which predicts the next word on a few previous words. Mikolov et al. [4] introduced the Recurrent Neural Network (RNN) to model text sequences. To alleviate the problem of gradient vanishing in RNN, Long-Short Term Memory (LSTM) [23] and Gated Recurrent Unit (GRU) [6] were proposed later. Recently, the sequence-to-sequence model [2] has received much attention, which models language in an encoder-decoder architecture.

In academic writing, Wang et al. [27] attempted to generate the abstract based on a given title. PaperRobot [26] was proposed to predict related entities for input and write a new paper. The above are all language models for uni-modal text inputs. However, many artificial intelligence problems involve more than one input modality [32]. In recent years, researchers began to devote to modeling multi-modal data. Combining image and text modality has attracted great attention. Yan et al. [31] proposed to generate an image for the specific text. The image caption [20] attempted to generate language texts based on image inputs. Most existing multi-modal research was image-oriented or text-oriented, which cannot be applied to networks. Net2Text [30] integrated node embeddings of network structure into personalized review generation. However, it ignored the differences in network characteristics and cannot integrate context effectively.

6 CONCLUSION

In this paper, we present an academic writing assistant, AutoCite, which can provide authors with more comprehensive related work with generated high-quality citation context. Specifically, AutoCite involves a novel multi-modal encoder and multi-task decoder architecture. The multi-modal encoder in AutoCite captures the characteristic differences when a paper node cites others, or it is cited by others. The multi-task decoder in AutoCite jointly learns citation link prediction and context generation in a unified manner. Furthermore, to effectively join the encoder and decoder, we propose a gated neural fusion to integrate multi-modal representations for downstream tasks adaptively. Extensive experiments are conducted on five real-world datasets, and the results demonstrated the superiority of AutoCite.

7 ACKNOWLEDGEMENT

This work is funded in part by the National Natural Science Foundation of China Projects No. U1636207, No. U1936213 and No. 61671157. This work is also partially supported by NSF through grant IIS-1763365 and by FSU. This work is funded by Ant Financial through the Ant Financial Science Funds for Security Research.

REFERENCES

[1] Luca Maria Aiello, Alain Barrat, Rossano Schifanella, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Friendship prediction and homophily in social media. *TWEB*, 6(2):1–33, 2012.
[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
[3] Regina Barzilay and Michael Elhadad. Using lexical chains for text summarization. *Advances in automatic text summarization*, pages 111–121, 1999.
[4] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.
[5] Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In *IJCAI*, pages 2137–2143, 2019.
[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
[7] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *ICASSP*, pages 6645–6649. IEEE, 2013.
[8] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks. In *SIGKDD*, pages 855–864, 2016.
[9] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *WWW*, pages 421–430. ACM, 2010.
[10] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C Lee Giles, and Lior Rokach. Recommending citations: translating papers into references. In *CIKM*, pages 1910–1914. Citeseer, 2012.
[11] Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C Lee Giles. A neural probabilistic model for context based citation recommendation. In *AAAI*. ACM, 2015.
[12] Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Park, and Sungchul Choi. A context-aware citation recommendation model with bert and graph convolutional networks. *arXiv preprint arXiv:1903.06464*, 2019.
[13] Yizhu Jiao, Yun Xiong, Jiawei Zhang, and Yangyong Zhu. Collective link prediction oriented network embedding with hierarchical graph attention. In *CIKM*, pages 419–428, 2019.
[14] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. Utilizing context in generative bayesian models for linked corpus. In *AAAI*. ACM, 2010.
[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
[16] Xiangnan Kong, Jiawei Zhang, and Philip S Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, pages 179–188, 2013.
[17] Onur Küçüktunç, Kamer Kaya, Erik Saule, and Ümit V Çatalyürek. Fast recommendation on bibliographic networks. In *ASONAM*, pages 480–487. IEEE, 2012.
[18] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2001.
[19] Jie Liu, Na Li, and Zhicheng He. Network embedding with dual generation tasks. In *IJCAI*, pages 5102–5108, 2019.
[20] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016.
[21] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *SIGKDD*, pages 701–710, 2014.
[22] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. Cluscite: Effective citation recommendation by information network-based clustering. In *SIGKDD*, pages 821–830. ACM, 2014.
[23] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *INTER-PEECH*, 2012.
[24] Jie Tang and Jing Zhang. A discriminative approach to topic-based citation recommendation. In *PAKDD*, pages 572–579. Springer, 2009.
[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
[26] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. Paperrobot: Incremental draft generation of scientific ideas. In *ACL*, 2019.
[27] Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. Paper abstract writing through editing mchanism. In *ACL*, 2018.
[28] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *SIGIR*, pages 165–174, 2019.
[29] Chuhan Wu, Fangzhao Wu, Tao Qi, Suyu Ge, Yongfeng Huang, and Xing Xie. Reviews meet graphs: Enhancing user and item representations for recommendation with hierarchical attentive graph neural network. In *EMNLP-IJCNLP*, pages 4886–4895, 2019.
[30] Shaofeng Xu, Yun Xiong, Xiangnan Kong, and Yangyong Zhu. Net2text: An edge labelling language model for personalized review generation. In *DASFAA*, pages 484–500. Springer, 2019.
[31] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *ECCV*, pages 776–791. Springer, 2016.
[32] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 2019.