A Retrieve-and-Read Framework for Knowledge Graph Link Prediction

Vardaan Pahuja
pahuja.9@osu.edu
The Ohio State University
Columbus, Ohio, United States

Boshi Wang
wang.13930@osu.edu
The Ohio State University
Columbus, Ohio, United States

Hugo Latapie
hlatapie@cisco.com
Cisco Research
San Jose, California, United States

Jayanth Srinivasa
jasriniv@cisco.com
Cisco Research
San Jose, California, United States

Yu Su
su.809@osu.edu
The Ohio State University
Columbus, Ohio, United States

ABSTRACT

Knowledge graph (KG) link prediction aims to infer new facts based on existing facts in the KG. Recent studies have shown that using the graph neighborhood of a node via graph neural networks (GNNs) provides more useful information compared to just using the query information. Conventional GNNs for KG link prediction follow the standard message-passing paradigm on the entire KG, which leads to superfluous computation, over-smoothing of node representations, and also limits their expressive power. On a large scale, it becomes computationally expensive to aggregate useful information from the entire KG for inference. To address the limitations of existing KG link prediction frameworks, we propose a novel retrieve-and-read framework, which first retrieves a relevant subgraph context for the query and then jointly reasons over the context and the query with a high-capacity reader. As part of our exemplar instantiation for the new framework, we propose a novel Transformer-based GNN as the reader, which incorporates graph-based attention structure and cross-attention between query and context for deep fusion. This simple yet effective design enables the model to focus on salient context information relevant to the query. Empirical results on two standard KG link prediction datasets demonstrate the competitive performance of the proposed method. Furthermore, our analysis yields valuable insights for designing improved retrievers within the framework.[1]

CCS CONCEPTS

• **Computing methodologies** → **Semantic networks**; **Statistical relational learning**.

KEYWORDS

Knowledge Graph Link Prediction, Knowledge Graph Completion, Graph Neural Networks, Transformers

[1]Code and data will be released on https://github.com/OSU-NLP-Group/KG-R3/.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3614769

ACM Reference Format:
Vardaan Pahuja, Boshi Wang, Hugo Latapie, Jayanth Srinivasa, and Yu Su. 2023. A Retrieve-and-Read Framework for Knowledge Graph Link Prediction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3583780.3614769

1 INTRODUCTION

Knowledge graphs encode a wealth of structured information in the form of *(subject, relation, object)* triples. The rapid growth of KGs in recent years has led to their wide use in diverse applications such as information retrieval [6, 71], question answering [22, 53], and data mining [73]. KG link prediction [2] that aims to infer new facts based on existing facts is a fundamental task on KGs. It finds applications in relation extraction [10, 62], question answering [25, 48], and recommender systems [29, 69].

Early methods for KG link prediction have focused on learning a dense embedding for each entity and relation in the KG, which is then used to calculate the plausibility of new facts via a simple scoring function [2, 15, 26, 37, 51]. The hope is that an entity's embedding will learn to compactly encode the structural and semantic information such that a simple scoring function would suffice for making accurate link predictions. However, it is challenging to fully encode the rich information of KGs into such shallow embeddings. Similar to how contextualized encoding models like BERT [16] have been supplanting static embeddings [40, 44] for natural language representation, several recent studies have adapted message-passing graph neural networks (GNNs) for KG link prediction [49, 50, 56]. By using a GNN to iteratively encode increasingly larger graph neighborhoods, GNN-based KG link prediction methods have shown great success. However, the reliance on message-passing over the entire KG leads to superfluous computation and limits their scalability on large-scale KGs. The same reason also leads to slow inference speed.

Intuitively, for each specific query, *e.g.*, *(Jack Ramsay, occupation, ?)*, only a small subgraph of the entire KG may be relevant for answering the query (Figure 1). If we could *retrieve* the relevant subgraph from the KG as *context*, we can then easily use a high-capacity model to *read* the query in the corresponding context to make the final inference. To this end, we propose a novel *retrieve-and-read framework* for KG link prediction (see Figure 1 for an overview and comparison with existing frameworks). It consists

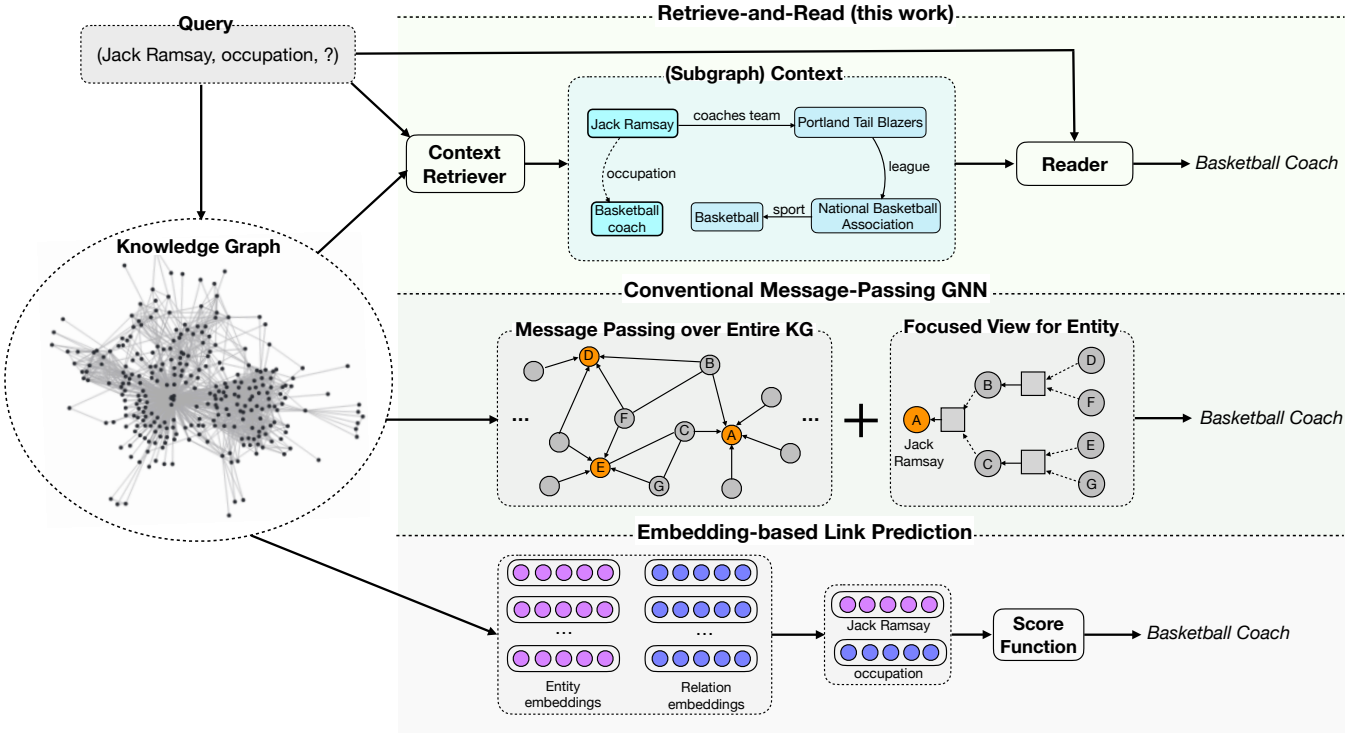

Figure 1: Overview of the proposed retrieve-and-read framework and comparison with existing frameworks for KG link prediction. Embedding-based methods try to encode all relevant information into shallow embeddings, while message-passing graph neural networks (GNNs) iteratively learn the representations through message-passing over the entire KG. In contrast, in our framework, we first retrieve a small context subgraph that is relevant to each input query, and then jointly encode the query and the context for the final prediction. Here for simplicity, we assume the context to be a connected subgraph, but being connected is not a necessary condition.

of two main components: a *retriever* that identifies the relevant KG subgraph as context for the input query and a *reader* that jointly considers the query and the retrieved context for inferring the answer. Such a retrieve-and-read framework has been widely used for the open-domain question answering (QA) problem [8, 74], which faces a similar fundamental challenge: it also needs to answer a question in a massive corpus where only a small fraction is relevant to each specific question. The modularization provided by this framework has enabled rapid progress on retriever and reader models separately [20, 28, 30, 64].

Embracing the retrieve-and-read framework for KG link prediction could bring multiple potential advantages: 1) It provides great flexibility to explore and develop diverse models for the retriever and reader separately. For example, we explore several different choices for the retriever, even including an existing KG link prediction model. Because the reader only needs to deal with a small subgraph instead of the entire KG, we could leverage the potential of high-capacity and more expressive models such as the Transformer [57], which has proven extremely successful for other tasks, but the application on KG link prediction has been limited. 2) Relatedly, separate and more focused progress can be made on each component, which can then be combined to form new KG link prediction models. 3) Instead of learning and leveraging the same *static* representation for all inferences as in existing frameworks, the reader can dynamically learn a contextualized representation for each query and context for more accurate prediction. 4) Finally, this framework could potentially support the development of GNN-based KG link prediction models for large-scale KGs, similar to how it has enabled open-domain QA to operate on web-scale corpora [8, 28, 61, 74].

To demonstrate the effectiveness of the proposed framework, we propose a novel instantiation of the framework, KG-R3 (KG Reasoning with Retriever and Reader). It uses an existing KG link prediction method, MINERVA [12] as retriever and a novel *Transformer-based GNN* as the reader. Existing GNNs are mostly based on message-passing [19], where the representation of a particular node is iteratively updated by its neighbors. Message-passing GNNs suffer from limited expressive power [65] and over-smoothing [36], *i.e.*, the representation of distinct nodes become indistinguishable as GNNs get deeper, which limits their model capacity. The time complexity of message-passing also grows exponentially with the number of layers, making it even harder to increase the capacity of such models. On the other hand, the Transformer model [57] has been driving the explosive growth of high-capacity models such as BERT [16]. The Transformer can support very high-capacity models [4, 11], which is one of the key reasons for its success. While

it is challenging to apply Transformer to the entire KG due to its limited context window, the small context in the retrieve-and-read framework makes it feasible. We design a novel Transformer-based GNN which has a two-tower structure to separately encode the query and the context subgraph and a cross-attention mechanism to enable deep fusing of the two towers. A graph-induced attention structure is also developed to encode the context subgraph.

The major contribution of this work is three-fold:

- We propose a novel retrieve-and-read framework for knowledge graph link prediction.
- We develop a novel instantiation, KG-R3, of the framework, which consists of a novel Transformer-based graph neural network for KG link prediction.
- We conduct empirical experiments on the standard FB15K-237 [54] and WN18RR [15] datasets and show that KG-R3 achieves competitive results with state-of-the-art methods.

2 RELATED WORK

Graph Neural Networks. Graphs naturally encode rich semantics of underlying data. Early models [5, 13, 33] extended the spectral convolution operation to graphs. Follow-up work [3, 27, 58] introduced attention and gating mechanisms to aggregate the salient information from a node's neighborhood. These aforementioned models are applicable only to homogeneous graphs. In our present work, we develop a novel Transformer-based GNN as the reader module for link prediction in multi-relational graphs like KGs.

KG Link Prediction Models. Early approaches for link prediction range from translation-based models [2, 37] and semantic matching models [43, 66] to the ones that leverage CNNs [15, 42]. These shallow embedding methods learn embeddings for each entity and relation and use a parameterized score function to predict the plausibility of a triple. To make use of the semantically rich graph neighborhood, several approaches have tried to adapt GNNs to multi-relational graphs for KG link prediction. R-GCN [49] and CompGCN [56] make use of relation-type dependent message aggregation. These methods use graph aggregation over the entire KG, thus limiting their scalability. In our framework, we propose to perform computation only on a relevant subgraph of the KG to mitigate this problem.

Path-based KG Reasoning. Another line of work uses multi-hop paths to synthesize information for predicting the missing facts in a KG. DeepPath [63] and MINERVA [12] formulate it as a sequential decision-making problem and use reinforcement learning to search paths to the target entity. For the retriever module, we use MINERVA as one of the baseline methods. Neural LP [67] and DRUM [47] use inductive logic programming to learn logical rules from KGs, which are used to weight different paths. Though these approaches are interpretable, they suffer from relatively poor performance compared to embedding-based KG link prediction methods. Our proposed framework can utilize the subgraphs generated by these approaches for improved performance.

Transformers for Graph ML Tasks. In contrast to NLP and vision, developing transformer models for graphs is challenging due to the absence of regularity in the data space, for example, 2D/3D grid in images and linear structure in sentences. Dwivedi and Bresson [17], Velickovic et al. [59] use Transformer-like self-attention for neighborhood aggregation in message-passing GNNs. Dwivedi et al. [18], Kim et al. [31], Kreuzer et al. [34] introduce positional encodings (PE) as an inductive bias for graph structure. Heterogeneous Graph Transformer [68] proposes a Levi graph based attention structure for encoding Abstract Meaning Representation (AMR) graphs for the task of AMR-to-text generation. GRAPH-BERT [70] uses the Transformer model for self-supervised learning of node representations, which can then be fine-tuned on downstream tasks. Similarly, Transformers have been applied for the task of KG link prediction. HittER [9] proposes a hierarchical Transformer model, which utilizes one-hop neighborhood context for the KG link prediction task. StAR [60] augments the contextual text representation obtained from a pre-trained language model with the structure information to learn better representations. Similarly, kgTransformer [38] pre-trains a Transformer model for out-of-domain generalization to unseen first-order logic queries. This work focuses on developing a Transformer-based GNN as the reader that can handle arbitrary subgraphs as context.

Open-domain Question Answering. The task of open-domain QA is to answer a question using knowledge from a massive corpus such as Wikipedia. A popular and successful way to address the challenge of large scale is through a two-stage retrieve-and-read pipeline [8, 74], which leads to rapid developments of retriever and reader separately [14, 28, 30, 64]. We draw inspiration from this pipeline and propose to use a retrieve-and-read framework for KG link prediction.

3 METHODOLOGY

3.1 Preliminaries

Knowledge Graph. Given a set of entities $\mathcal{E}$ and a set of relations $\mathcal{R}$, a knowledge graph can be defined as a collection of facts $\mathcal{F} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ where for each fact $f = (h, r, t),\ h, t \in \mathcal{E}, r \in \mathcal{R}$.

Link Prediction. The task of link prediction is to infer the missing facts in a KG. Given a link prediction query $(h, r, ?)$ or $(?, r, t)$, the model ranks the target entity among the set of candidate entities. For the query $(h, r, ?)$, h and r correspond to the source entity and the query relation, respectively.

3.2 Retriever

The function of the retriever module is to select a relevant subgraph of the KG as the query context. We use the following off-the-shelf methods to generate subgraph inputs for the Transformer-based reader module in our framework.

3.2.1 Uninformed Search-based Retrieval Strategies.

a) **Breadth-first search**: For breadth-first search, we sample edges starting from the source entity in breadth-first order till we reach the context budget.
b) **One-hop neighborhood**: The one-hop neighborhood includes edges in the immediate one-hop neighborhood of the source entity.

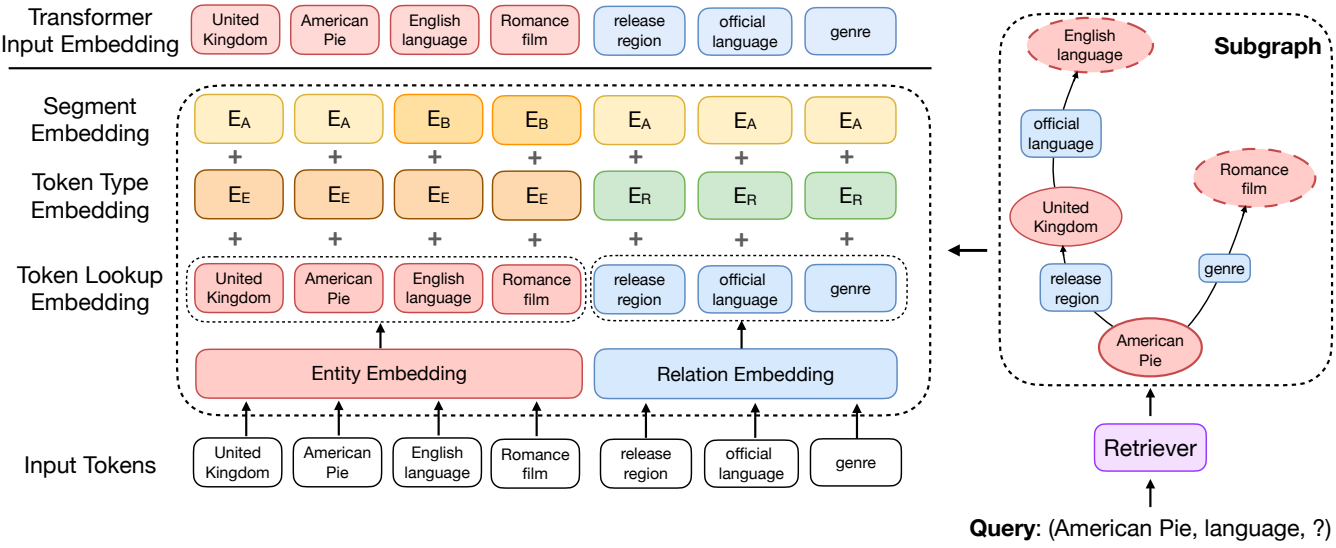

Figure 2: Schematic representation of embedding layer for subgraph input. The Transformer input is the sum of the token lookup embedding, the token type embedding, and the segment embedding.

3.2.2 Informed Search-based Retrieval Strategies. We use **MINERVA** [12] as our informed search-based retrieval strategy. It formulates KG link prediction as the generation of multi-hop paths from the source entity to the target entity. The environment is represented as a Markov Decision Process (MDP) on the KG, where the reinforcement learning agent gets a positive reward on reaching the target. The set of paths generated by MINERVA provides an interpretable provenance for KG link prediction. The retriever model utilizes the union of these paths decoded using beam search as the subgraph output.

Among these approaches, breadth-first search and one-hop neighborhood use uninformed search, *i.e.*, they only enrich the query using surrounding context without explicitly going toward the target entity. On the other hand, the subgraph obtained using the MINERVA retriever aims to provide context which encloses information towards reaching the target entity.

3.3 Reader Architecture

Embedding Layer. The input to the Transformer is obtained by summing the token lookup embedding, the token type embedding, and the segment embedding (Figure 2):

- **Token lookup embedding**: We use learned lookup embeddings for all entities and relations. These lookup embeddings store the global semantic information for each token.
- **Token type embedding**: Entity and relation tokens have different semantics, so we use token type embeddings to help the model distinguish between them.
- **Segment embedding**: It denotes whether a particular entity token corresponds to the terminal entity in a path beginning from the source entity. This helps the model to differentiate between the terminal tokens, which are more likely to correspond to the final answer vs. other tokens.

The input to the reader module is the query, *e.g.*, $(h, r, ?)$, and its associated context, a subgraph of the KG. The input sequence for the subgraph encoder is constructed by concatenating the sets of node and edge tokens. Each edge corresponds to a unique token in the input, though there might be multiple edges with the same predicate. At a high level, the query and subgraph are first encoded by their respective Transformer encoders. The query self-attention encoder takes "`[CLS], [source entity], [query relation]`" as the input sequence with a fully-connected attention structure. Then the cross-attention module is used to modulate the subgraph representation, conditioned on the query.

Graph-induced Self-Attention. The attention structure ($\mathcal{A}_i$) governs the set of tokens that a particular token can attend to in the self-attention layer of the subgraph encoder. It helps incorporate the (sub)graph structure into the transformer representations. Inspired by KG-augmented PLMs [21, 23], we define the attention structure (Figure 3) such that 1) all node tokens can attend to each other; 2) all edge tokens can attend to each other; 3) for a particular triple (h, r, t), the token pairs (h, r) and (r, t) can attend to each other, and 4) each token attends to itself. This design is motivated by the need to maintain a balance between the immediate graph neighborhood of a token vs. its global context in the subgraph.

More formally, let the subgraph consist of m nodes and n edges. Let $\{\boldsymbol{h}_i^\ell\}_{i=1}^{m+n}$ denote the hidden representations of the tokens in layer ℓ of the subgraph self-attention encoder,

$$\boldsymbol{h}_i^{\ell+1} = \mathbf{O}^\ell \Big\|_{k=1}^{H} \Big(\sum_{j \in \mathcal{A}_i} w_{ij}^{k,\ell} \mathbf{V}^{k,\ell} \boldsymbol{h}_j^\ell \Big), \tag{1}$$

$$w_{ij}^{k,\ell} = \text{softmax}_{j \in \mathcal{A}_i} \Big(\frac{\mathbf{Q}^{k,\ell} \boldsymbol{h}_i^\ell \cdot \mathbf{K}^{k,\ell} \boldsymbol{h}_j^\ell}{\sqrt{d_k}} \Big). \tag{2}$$

Here, $\mathbf{Q}^{k,\ell}, \mathbf{K}^{k,\ell}, \mathbf{V}^{k,\ell} \in \mathbb{R}^{d_k \times d}$, $\mathbf{O}^{\ell} \in \mathbb{R}^{d \times d}$ are projection matrices for the k^{th} attention head in layer l, H denotes the number of attention heads per layer, d_k denotes the hidden dimension of keys and $\|$ denotes concatenation.

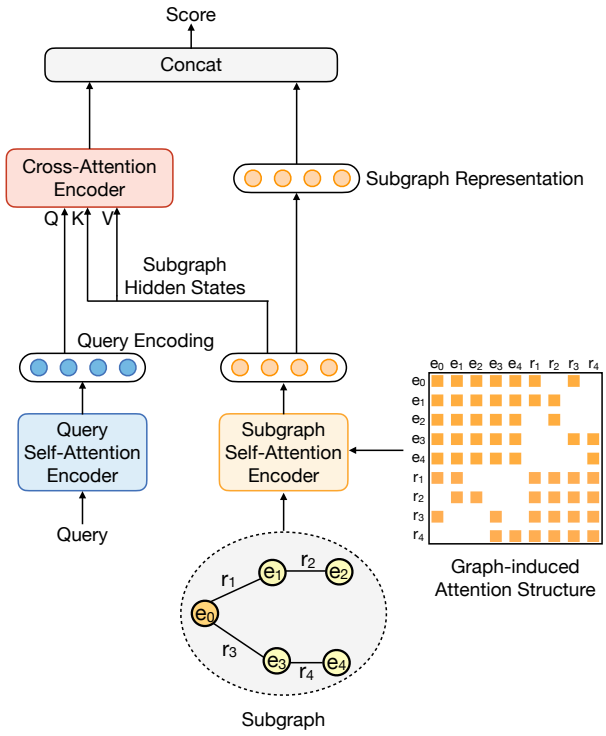

Figure 3: Reader module architecture. In the subgraph self-attention encoder, the graph-induced self-attention design governs the attention between tokens. In the cross-attention encoder, the link prediction query hidden states serve as the query input, while the subgraph hidden states serve as the key and value inputs.

Cross-Attention. To answer a link prediction query, the model needs a way to filter the context in the subgraph that is relevant to a particular link prediction query. To accomplish this, we introduce cross-attention from the query to the subgraph (Figure 3). Following Vaswani et al. [57], the queries come from query hidden states, whereas the keys and values come from the subgraph hidden states. The resultant representation encodes the subgraph information that is relevant to the query at hand. More formally, Let $\boldsymbol{e}_i^{\mathrm{q},\ell}$ and $\{\boldsymbol{e}_j^{\mathrm{sub}}\}_{j=1}^{m+n}$ denote the self-attention hidden representations of i^{th} query token and the subgraph tokens, respectively,

$$\text{Cross-Attention}(\{\boldsymbol{e}_i^{\mathrm{q},\ell}\}, \{\boldsymbol{e}_j^{\mathrm{sub}}\}_{j=1}^{m+n}) = \mathbf{O}^{\ell} \Big\|_{k=1}^{H} \left(\sum_{j=1}^{m+n} w_{ij}^{k,\ell} \mathbf{V}^{k,\ell} \boldsymbol{e}_j^{\mathrm{sub}} \right), \tag{3}$$

$$w_{ij}^{k,\ell} = \text{softmax}\left(\frac{\mathbf{Q}^{k,\ell} \boldsymbol{e}_i^{\mathrm{q},\ell} \cdot \mathbf{K}^{k,\ell} \boldsymbol{e}_j^{\mathrm{sub}}}{\sqrt{d_k}} \right). \tag{4}$$

This is concatenated with the contextualized representation of the source entity in the subgraph to output the feature vector for predicting the plausibility scores. For a link prediction query *e.g.*, $(h, r, ?)$, the model predicts a score distribution over all candidate entities. The model is trained using cross-entropy loss, framing it as a multi-class classification problem. Figure 3 illustrates the overall model architecture of the Transformer-based reader.

Table 1: Dataset Statistics

Dataset	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{R}\rvert$	# Facts		
			Train	Valid	Test
FB15K-237	14,541	237	272,115	17,535	20,466
WN18RR	40,943	11	86,835	3,034	3,134

4 EXPERIMENTS

4.1 Datasets

We use standard link prediction benchmarks FB15K-237 [54] and WN18RR [15] to evaluate our model. FB15K-237 is an encyclopedic KG derived from the FB15K dataset [2] after removing data contamination due to overlap between train and test sets. In a similar fashion, WN18RR is derived from the WN18 dataset [2] after deleting inverse relations to prevent inference by memorizing training facts. The FB15K-237 dataset is released under Microsoft Research Data License Agreement. The WN18RR dataset is released by Princeton University under their license. Both datasets are permitted to be used for academic research in their original form. Dataset statistics are reported in Table 1.

4.2 Evaluation Protocol

For a link prediction query, *e.g.*, $(h, r, ?)$, the model scores all candidate triplets $C = \{(h, r, t'), t' \in \mathcal{E}\}$ and ranks the correct entity in the list. Following Bordes et al. [2], we use the filtered evaluation setting *i.e.*, the rank of a target entity is not affected by alternate correct entities. We report results on standard evaluation metrics: Mean Reciprocal Rank (MRR), Hits@1, Hits@3, and Hits@10.

4.3 Baselines

We conduct a comprehensive comparison of our model with various KG link prediction baselines. These include embedding-based methods such as TransE [2], DistMult [66], and ComplEx [55], CNN-based approaches like ConvE [15] and ConvKB [42], path-based techniques such as MINERVA [12], Neural LP [67], and DRUM [47], GNN-based methods like R-GCN [49] and CompGCN [56], as well

Table 2: Comparison of our framework with baseline methods. For all metrics, higher is better. Missing values are denoted by –. Results of RESCAL, TransE, DistMult, ComplEx, and ConvE correspond to the best results obtained after extensive hyperparameter tuning [46]. Result of Neural LP and DRUM are taken from Qu et al. [45] following the standard evaluation setting. Results of other methods are taken from their original papers.

		FB15K-237				WN18RR			
Framework	**Model**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
Embedding-based	RESCAL [43]	.356	.263	.393	.541	.467	.439	.480	.517
	TransE [2]	.313	.221	.347	.497	.228	.053	.368	.520
	DistMult [66]	.343	.250	.378	.531	.452	.413	.466	.530
	ComplEx [55]	.348	.253	.384	.536	.475	.438	.490	.547
	ComplEx-N3 [35]	.37	–	–	.56	.48	–	–	.57
	RotatE [52]	.338	.241	.375	.533	.476	.428	.492	.571
CNN-based	ConvKB [42]	.243	.155	.371	.421	.249	.057	.417	.524
	ConvE [15]	.339	.248	.369	.521	.442	.411	.451	.504
Path-based	MINERVA [12]	.293	.217	.329	.456	.448	.413	.456	.513
	Neural LP [67]	.237	.173	.259	.361	.381	.368	.386	.408
	DRUM [47]	.238	.174	.261	.364	.382	.369	.388	.410
GNN-based	R-GCN [49]	.248	.151	–	.417	–	–	–	–
	CompGCN [56]	.355	.264	.390	.535	.479	.443	.494	.546
Transformer-based	HittER [9]	.373	.279	.409	**.558**	**.503**	**.462**	**.516**	**.584**
	KG-R3 (this work)	**.390**	**.315**	**.413**	.539	.472	.439	.481	.537

Table 3: Comparison of our Transformer-based GNN with other reader architectures using MINERVA subgraphs. Please refer to the Appendix for implementation details.

	FB15K-237				WN18RR			
Model	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**	**MRR**	**Hits@1**	**Hits@3**	**Hits@10**
CompGCN [56]	.272	.185	.302	.444	.335	.267	.382	.452
HetGT [68]	.264	.213	.274	.361	.365	.317	.396	.443
Ours	**.390**	**.315**	**.413**	**.539**	**.472**	**.439**	**.481**	**.537**

as transformer-based approaches like HittER [9].[2] To ensure a fair comparison, we exclude models that involve excessive computation, such as NBFNet [76], which relies on learning representations based on *all* paths between each pair of entities. This computationally expensive operation limits its scalability on large KGs. Additionally, we exclude approaches that utilize extra information or pre-trained language models [39, 60].

4.4 Implementation Details

We implement our models in Pytorch. We use $L = 3, A = 8, H = 320$ for the Transformer model (both self-attention and cross-attention layers), where L, A, and H denote the number of layers, the number of attention heads per layer, and the hidden size, respectively. The intermediate hidden dim. for the feedforward layer in Transformer is set to 1280. We use the Adamax [32] optimizer for training. The learning rate schedule includes warmup for 10% of the training steps followed by linear decay. The batch size is set to 512. For both datasets, we tune the learning rate on the validation set and report results on the test set with the best validation setting. We sort the training instances in ascending order based on subgraph size for better training efficiency. During training, we use early stopping based on the validation set to prevent overfitting. To prevent exposure bias, we omit subgraph edges that overlap with the query triple during training. A more comprehensive description of the hyperparameter setup is given in the Appendix.

[2]Due to space constraints, we only include the top performing KG link prediction methods in Table 2. For comprehensive results on all approaches in the literature, please refer to the surveys [1, 41].

5 RESULTS

In this section, we attempt to answer the following questions:

Q1. How does the performance of KG-R3 compare to other KG link prediction baselines? (§5.1)
Q2. How does our proposed Transformer-based GNN reader compare to other reader architectures? (§5.2)

Q3. How do factors like the presence of the target entity in the subgraph and the entity degree affect the link prediction performance? (§5.3)
Q4. How does KG-R3 compare to other GNN baselines in terms of theoretical complexity and inference speed? (§5.4)

5.1 Performance Comparison on KG Link Prediction

Table 2 shows the overall link prediction results. On the FB15K-237 benchmark, KG-R3 consistently outperforms all baselines in 3 out of 4 evaluation metrics. Our model obtains a 12.9% relative improvement (Hits@1) over HittER [9], a baseline Transformer model for KG link prediction. Compared to the best performing message-passing GNN baseline (CompGCN), it improves by 5.1% in terms of Hits@1. Our model, which uses the MINERVA subgraph retriever, improves over the MINERVA baseline by 9.8% and 2.6% (Hits@1) for FB15K-237 and WN18RR, respectively. For the WN18RR dataset, our model demonstrates superior performance compared to embedding-based approaches, CNN-based approaches, and path-based approaches in terms of Hits@1. However, the main goal of this work is not to achieve state-of-the-art performance on standard link prediction benchmarks. We hypothesize that this dataset is more sensitive to noisy subgraph inputs due to fewer relation types. Therefore, it is comparatively harder for the reader module to filter irrelevant context. It also leads to embedding-based approaches performing much better on this dataset.

5.2 Comparison to other Reader Architectures

We compare our proposed Transformer-based GNN reader module with other reader model architectures for the MINERVA retriever. We use a message-passing based GNN, CompGCN [56], and a Transformer-based GNN, Heterogeneous Graph Transformer (HetGT) [68], as baselines. For both datasets, our proposed Transformer-based GNN reader vastly outperforms them (Table 3). CompGCN uses the subgraph context to score candidate triplets. Instead of using a static context (entire KG), it now varies dynamically for each training instance. The dynamic contextualized embeddings output by the GCN encoder perform poorly with triplet-based scoring functions. Although similar to our proposed reader, HeGT lacks some important inductive biases which are key to a good performance — the token type embeddings, global attention between *e-e* and *r-r* tokens only rather than all tokens and cross-attention.

Given the context subgraph, a reader should learn to discard the superfluous context and reason using the relevant information to infer the correct answer. Our novel cross-attention design is a step in this direction. Expanding on this capability is a promising way to make the reader module more robust to noisy subgraph inputs.

5.3 Fine-grained Analyses

Effect of Target Entity Coverage in Subgraph. To gain further insights into the reader, we report a breakdown of the link prediction performance based on whether the target entity is present in the input subgraph (Table 4). When the target entity is present in the subgraph, the comparative performance is very high (Hits@1 is almost 8× compared to when it is absent for the MINERVA retriever). This can be explained by the fact that the coverage of the target entity provides the reader with some potentially correct reasoning paths to better establish the link between the source and the target entity. This also gives an insight that the coverage of the target entity in the context subgraph could be a useful indicator of the retriever module's performance.

Table 4: Performance breakdown based on whether the target entity is present in the input subgraph (FB15K-237 val. set). The performance is significantly better when the target entity is present in the subgraph.

Retriever	Target ent. coverage	MRR	Hits@1	Hits@10
MINERVA	present	**.683**	**.599**	**.857**
	absent	.144	.078	.277
BFS	present	**.626**	**.515**	**.846**
	absent	.250	.167	.418
One-hop neigh.	present	**.949**	**.928**	**.978**
	absent	.339	.250	.518

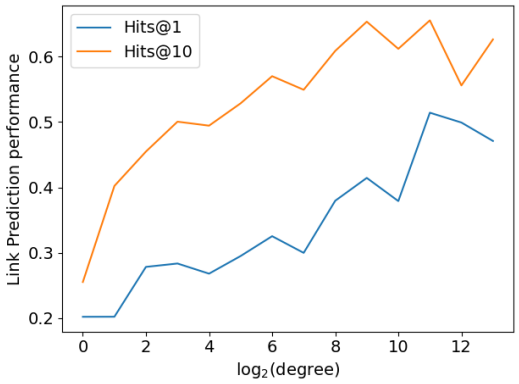

Figure 4: Link Prediction performance grouped by the logarithm of entity degree. A triple (h, r, t) belongs to a degree group if either h or t belongs to it. The performance is significantly better for higher entity degrees.

Effect of Entity Degree. We analyze the effect of the degree of source and target entities in the training graph on the overall link prediction performance for FB15K-237 (Figure 4). We observe that the performance increases gradually with an increase in the entity degree. Intuitively, a higher entity degree increases the retriever's likelihood of discovering useful evidential context in the form of logical reasoning paths, connecting the source entity to the target entity for link prediction.

5.4 Efficiency Analysis

Table 6 shows the comparison of training and inference complexity (per triplet) of our method with R-GCN, a prominent GNN baseline. The calculation includes the complexity of the MINERVA retriever $O(d^2 + d\frac{|\mathcal{E}|}{|\mathcal{V}|})$. Since the subgraph size $(n + e)$ is much smaller than

Table 5: Ablations for *Retriever* module (validation set). The MINERVA retriever significantly outperforms other retrievers.

		FB15K-237				WN18RR			
Model	**Target entity coverage (%)**	**MRR**	**H@1**	**H@3**	**H@10**	**MRR**	**H@1**	**H@3**	**H@10**
MINERVA	46.28	**.394**	**.319**	**.415**	**.546**	**.469**	**.437**	**.476**	**.529**
BFS	16.45	.311	.223	.343	.488	.377	.314	.416	.483
One-hop neigh.	.51	.340	.252	.371	.520	.441	.402	.456	.517

Table 6: Comparison of the complexity of our approach to baseline methods. Here, n and e denote the average number of nodes and edges in a subgraph, respectively. $|\mathcal{E}|$ and $|\mathcal{V}|$ denote the number of edges and nodes in the KG, respectively. T is the no. of iterations needed for convergence and d is the hidden dimension.

Model	**Training complexity**	**Inference complexity (per triplet)**
R-GCN [49]	$O(T(\|\mathcal{E}\|d^2))$	$O(\|\mathcal{E}\|d^2)$
KG-R3 (this work)	$O(T((n+e)^2d + (n+e)d^2 + d^2 + d\frac{\|\mathcal{E}\|}{\|\mathcal{V}\|}))$	$O((n+e)^2d + (n+e)d^2 + d^2 + d\frac{\|\mathcal{E}\|}{\|\mathcal{V}\|})$

Table 7: Ablations for *Reader* module (FB15K-237 val. set). Graph attention structure contributes the most towards final performance. Please refer to the Appendix for implementation details.

Model	**MRR**	**H@1**	**H@3**	**H@10**
Ours	**.394**	**.319**	**.415**	.546
– Cross-Attention	.387	.312	.410	.541
– Graph attention structure	.311	.228	.334	.482
– Subgraph embed	.370	.277	.406	**.556**
– Query embed	.359	.277	.384	.527

the number of entities in the KG (denoted by $|\mathcal{E}|$), our approach is more efficient than R-GCN both for training and inference.

Table 8: Comparison of wall time of predicting a single edge for ogbl-biokg. The wall time is measured on a server with 128 CPU cores and a single A6000 GPU.

Model	**Evaluation time per triplet (ms)**
CompGCN	6.94
KG-R3 (this work)	**1.19**

We further compare the wall time of predicting a single edge for ogbl-biokg [24]. Ogbl-biokg is a large-scale biomedical KG with 93.8K nodes and 5.1M edges. Our model achieves a speedup of 5.83× compared to CompGCN, a message-passing baseline (Table 8). This demonstrates that our model offers a significant advantage over message-passing GNNs in terms of inference efficiency.

5.5 Ablation Studies

Retriever. For the retriever, we experiment with three choices — MINERVA, breadth-first search, and one-hop neighborhood (Table 5). For both datasets, the MINERVA retriever outperforms BFS and one-hop neighborhood by a significant margin. This performance advantage can be attributed to the explicit training of MINERVA using reinforcement learning (RL) to discover paths leading to the target entity, whereas the other two approaches rely on uninformed search strategies. To gain further insights, we analyze the statistics for target entity coverage in the subgraph (Table 5).[3] As MINERVA outperforms the BFS retriever by a significant margin, this indicates that higher target entity coverage in the subgraph potentially contributes to better performance. However, the one-hop retriever outperforms BFS despite lower target entity coverage. This can be attributed to the fact that the reader better learns to ignore the noisy subgraph inputs due to the smaller context size provided by the one-hop retriever.

Reader. For the reader, we perform several ablations to evaluate the impact of different design choices. We investigate the impact of cross-attention, subgraph feature representation, query representation and using fully-connected attention in Transformer instead of the graph-induced attention structure (Table 7). The most significant drop in performance is caused by dropping the graph-induced attention structure, which shows that our novel attention design plays a key role in overall performance. Among the query and subgraph feature representations, the former has a greater contribution to the performance.

6 CONCLUSION AND FUTURE WORK

In this work, we propose a retrieve-and-read framework for knowledge graph link prediction. We develop a novel instantiation, KG-R3, of the framework, which consists of a novel Transformer-based graph neural network for KG link prediction. While being an initial exploration of our proposal, empirical experiments on standard benchmarks show that KG-R3 achieves competitive results with state-of-the-art methods, which indicates the great potential of the proposed framework. Our analysis offers valuable insights that can

[3]The one-hop neighborhood retriever has a low coverage as we used a small sample of one-hop edges following the HittER baseline.

aid in the design of better retrievers within the proposed framework. In principle, our proposed framework is scalable to large KGs, but it will require some nontrivial engineering optimizations, *e.g.*, mixed CPU-GPU training, storing entity embedding parameters on CPU and accessing them using a push-pull API [72, 75]. Presently, the reader component, even though outperforming existing alternatives, is still prone to noisy context subgraphs. How to further improve the robustness of the reader to noisy contexts is another major venue for future investigation. We believe this new framework will be a valuable resource for the research community to accelerate the development of high-performance and scalable graph-based models for KG link prediction. Future work will involve exploring the application of our framework in other knowledge graph tasks as well.

7 ETHICAL CONSIDERATIONS

KG link prediction aims to complete knowledge graphs by assigning higher scores to correct facts than incorrect ones. However, the top-ranked predictions might not be necessarily true, so a human verification of such predictions is required before the missing links are incorporated into the KG for public consumption. Together with human evaluation, such models have great potential to improve the coverage of KGs.

ACKNOWLEDGEMENTS

The authors would like to thank colleagues from the OSU NLP group and Soumya Sanyal for their valuable feedback. This research was supported in part by Cisco, ARL W911NF2220144, NSF OAC 2112606, NSF OAC 2118240, and Ohio Supercomputer Center [7].

REFERENCES

[1] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Mikhail Galkin, Sahand Sharifzadeh, Asja Fischer, Volker Tresp, and Jens Lehmann. 2022. Bringing Light Into the Dark: A Large-Scale Evaluation of Knowledge Graph Embedding Models Under a Unified Framework. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 12 (2022), 8825–8845. https://doi.org/10.1109/TPAMI.2021.3124805

[2] Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 2787–2795. https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html

[3] Xavier Bresson and Thomas Laurent. 2017. Residual Gated Graph ConvNets. *arXiv e-prints*, Article arXiv:1711.07553 (Nov. 2017), arXiv:1711.07553 pages. arXiv:1711.07553 [cs.LG]

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[5] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral Networks and Locally Connected Networks on Graphs. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1312.6203

[6] Pablo Castells, Miriam Fernandez, and David Vallet. 2007. An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. *IEEE Transactions on Knowledge and Data Engineering* 19, 2 (2007), 261–272. https://doi.org/10.1109/TKDE.2007.22

[7] Ohio Supercomputer Center. 1987. Ohio Supercomputer Center. http://osc.edu/ark:/19495/f5s1ph73

[8] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1870–1879. https://doi.org/10.18653/v1/P17-1171

[9] Sanxing Chen, Xiaodong Liu, Jianfeng Gao, Jian Jiao, Ruofei Zhang, and Yangfeng Ji. 2021. HittER: Hierarchical Transformers for Knowledge Graph Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 10395–10407. https://doi.org/10.18653/v1/2021.emnlp-main.812

[10] Xiaoyu Chen and Rohan Badlani. 2020. Relation Extraction with Contextualized Relation Embedding (CRE). In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Association for Computational Linguistics, Online, 11–19. https://doi.org/10.18653/v1/2020.deelio-1.2

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. *arXiv e-prints*, Article arXiv:2204.02311 (April 2022), arXiv:2204.02311 pages. arXiv:2204.02311 [cs.CL]

[12] Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, Luke Vilnis, Ishan Durugkar, Akshay Krishnamurthy, Alex Smola, and Andrew McCallum. 2018. Go for a Walk and Arrive at the Answer: Reasoning Over Paths in Knowledge Bases using Reinforcement Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=Syg-YfWCW

[13] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 3837–3845. https://proceedings.neurips.cc/paper/2016/hash/04df4d434d481c5bb723be1b6df1ee65-Abstract.html

[14] Xiang Deng, Yu Su, Alyssa Lees, You Wu, Cong Yu, and Huan Sun. 2021. ReasonBERT: Pre-trained to Reason with Distant Supervision. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6112–6127. https://doi.org/10.18653/v1/2021.emnlp-main.494

[15] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Thirty-second AAAI conference on artificial intelligence*.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[17] Vijay Prakash Dwivedi and Xavier Bresson. 2021. A Generalization of Transformer Networks to Graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications* (2021).

[18] Vijay Prakash Dwivedi, Chaitanya K. Joshi, Anh Tuan Luu, Thomas Laurent, Yoshua Bengio, and Xavier Bresson. 2020. Benchmarking Graph Neural Networks. *arXiv e-prints*, Article arXiv:2003.00982 (March 2020), arXiv:2003.00982 pages. arXiv:2003.00982 [cs.LG]

[19] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. 2017. Neural message passing for quantum chemistry. In *International conference on machine learning*. PMLR, 1263–1272.

[20] Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, G P Shrivatsa Bhargav, Dinesh Garg, and Avi Sil. 2020. Span Selection Pre-training for Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2773–2782. https://doi.org/10.18653/v1/2020.acl-main.247

[21] Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2281–2290. https://doi.org/10.18653/v1/2020.findings-emnlp.207

[22] Gaole He, Yunshi Lan, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. Improving Multi-hop Knowledge Base Question Answering by Learning Intermediate Supervision Signals. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 553–561. https://doi.org/10.1145/3437963.3441753

[23] Lei He, Suncong Zheng, Tao Yang, and Feng Zhang. 2021. KLMo: Knowledge Graph Enhanced Pretrained Language Model with Fine-Grained Relationships. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4536–4542.

[24] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open Graph Benchmark: Datasets for Machine Learning on Graphs. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/fb60d411a5c5b72b2e7d3527cfc84fd0-Abstract.html

[25] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge Graph Embedding Based Question Answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 105–113. https://doi.org/10.1145/3289600.3290956

[26] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 687–696.

[27] Wei Jin, Tyler Derr, Yiqi Wang, Yao Ma, Zitao Liu, and Jiliang Tang. 2021. Node Similarity Preserving Graph Convolutional Networks. In *WSDM '21, The Fourteenth ACM International Conference on Web Search and Data Mining, Virtual Event, Israel, March 8-12, 2021*, Liane Lewin-Eytan, David Carmel, Elad Yom-Tov, Eugene Agichtein, and Evgeniy Gabrilovich (Eds.). ACM, 148–156. https://doi.org/10.1145/3437963.3441735

[28] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[29] Miriyala Kartheek and G P Sajeev. 2021. Building Semantic Based Recommender System Using Knowledge Graph Embedding. In *2021 Sixth International Conference on Image Information Processing (ICIIP)*, Vol. 6. 25–29. https://doi.org/10.1109/ICIIP53038.2021.9702632

[30] Omar Khattab, Christopher Potts, and Matei A. Zaharia. 2021. Baleen: Robust Multi-Hop Reasoning at Scale via Condensed Retrieval. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 27670–27682. https://proceedings.neurips.cc/paper/2021/hash/e8b1cbd05f6e6a358a81dee52493dd06-Abstract.html

[31] Jinwoo Kim, Dat Tien Nguyen, Seonwoo Min, Sungjun Cho, Moontae Lee, Honglak Lee, and Seunghoon Hong. 2022. Pure Transformers are Powerful Graph Learners. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=um2BxfgkT2_

[32] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[33] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=SJU4ayYgl

[34] Devin Kreuzer, Dominique Beaini, Will Hamilton, Vincent Létourneau, and Prudencio Tossou. 2021. Rethinking graph transformers with spectral attention. *Advances in Neural Information Processing Systems* 34 (2021), 21618–21629.

[35] Timothée Lacroix, Nicolas Usunier, and Guillaume Obozinski. 2018. Canonical Tensor Decomposition for Knowledge Base Completion. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 2869–2878. http://proceedings.mlr.press/v80/lacroix18a.html

[36] Qimai Li, Zhichao Han, and Xiao-Ming Wu. 2018. Deeper Insights Into Graph Convolutional Networks for Semi-Supervised Learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 3538–3545. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16098

[37] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, Blai Bonet and Sven Koenig (Eds.). AAAI Press, 2181–2187. http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9571

[38] Xiao Liu, Shiyu Zhao, Kai Su, Yukuo Cen, Jiezhong Qiu, Mengdi Zhang, Wei Wu, Yuxiao Dong, and Jie Tang. 2022. Mask and Reason: Pre-Training Knowledge Graph Transformers for Complex Logical Queries. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 1120–1130. https://doi.org/10.1145/3534678.3539472

[39] Justin Lovelace and Carolyn Rosé. 2022. A Framework for Adapting Pre-Trained Language Models to Knowledge Graph Completion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 5937–5955. https://aclanthology.org/2022.emnlp-main.398

[40] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 3111–3119. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

[41] Dat Quoc Nguyen. 2020. A survey of embedding models of entities and relationships for knowledge graph completion. In *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*. Association for Computational Linguistics, Barcelona, Spain (Online), 1–14. https://doi.org/10.18653/v1/2020.textgraphs-1.1

[42] Dai Quoc Nguyen, Tu Dinh Nguyen, Dat Quoc Nguyen, and Dinh Phung. 2018. A Novel Embedding Model for Knowledge Base Completion Based on Convolutional Neural Network. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 327–333. https://doi.org/10.18653/v1/N18-2053

[43] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A Three-Way Model for Collective Learning on Multi-Relational Data. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, Lise Getoor and Tobias Scheffer (Eds.). Omnipress, 809–816. https://icml.cc/2011/papers/438_icmlpaper.pdf

[44] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[45] Meng Qu, Junkun Chen, Louis-Pascal A. C. Xhonneux, Yoshua Bengio, and Jian Tang. 2021. RNNLogic: Learning Logic Rules for Reasoning on Knowledge Graphs. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=tGZu6DlbreV

[46] Daniel Ruffinelli, Samuel Broscheit, and Rainer Gemulla. 2020. You CAN Teach an Old Dog New Tricks! On Training Knowledge Graph Embeddings. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=BkxSmlBFvr

[47] Ali Sadeghian, Mohammadreza Armandpour, Patrick Ding, and Daisy Zhe Wang. 2019. DRUM: End-To-End Differentiable Rule Mining On Knowledge Graphs. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 15321–15331. https://proceedings.neurips.cc/paper/2019/hash/0c72cb7ee1512f800abe27823a792d03-Abstract.html

[48] Apoorv Saxena, Aditay Tripathi, and Partha Talukdar. 2020. Improving Multi-hop Question Answering over Knowledge Graphs using Knowledge Base Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4498–4507. https://doi.org/10.18653/v1/2020.acl-main.412

[49] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*. Springer, 593–607.

[50] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. 2019. End-to-End Structure-Aware Convolutional Networks for Knowledge Base Completion. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial*

Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. AAAI Press, 3060–3067. https://doi.org/10.1609/aaai.v33i01.33013060

[51] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (Eds.). 926–934. https://proceedings.neurips.cc/paper/2013/hash/b337e84de8752b27eda3a12363109e80-Abstract.html

[52] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=HkgEQnRqYQ

[53] Alon Talmor and Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 641–651. https://doi.org/10.18653/v1/N18-1059

[54] Kristina Toutanova and Danqi Chen. 2015. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality*. 57–66.

[55] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016 (JMLR Workshop and Conference Proceedings, Vol. 48)*, Maria-Florina Balcan and Kilian Q. Weinberger (Eds.). JMLR.org, 2071–2080. http://proceedings.mlr.press/v48/trouillon16.html

[56] Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based Multi-Relational Graph Convolutional Networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=BylA_C4tPr

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[58] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJXMpikCZ

[59] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rJXMpikCZ

[60] Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021. Structure-Augmented Text Representation Learning for Efficient Knowledge Graph Completion. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 1737–1748. https://doi.org/10.1145/3442381.3450043

[61] Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced Ranker-Reader for Open-Domain Question Answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 5981–5988. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16712

[62] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1591–1601.

[63] Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. DeepPath: A Reinforcement Learning Method for Knowledge Graph Reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 564–573. https://doi.org/10.18653/v1/D17-1060

[64] Wenhan Xiong, Xiang Lorraine Li, Srini Iyer, Jingfei Du, Patrick S. H. Lewis, William Yang Wang, Yashar Mehdad, Scott Yih, Sebastian Riedel, Douwe Kiela, and Barlas Oguz. 2021. Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=EMHoBG0avc1

[65] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2019. How Powerful are Graph Neural Networks?. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net. https://openreview.net/forum?id=ryGs6iA5Km

[66] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6575

[67] Fan Yang, Zhilin Yang, and William W. Cohen. 2017. Differentiable Learning of Logical Rules for Knowledge Base Reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 2319–2328. https://proceedings.neurips.cc/paper/2017/hash/0e55666a4ad822e0e34299df3591d979-Abstract.html

[68] Shaowei Yao, Tianming Wang, and Xiaojun Wan. 2020. Heterogeneous graph transformer for graph-to-sequence learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7145–7154.

[69] Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 353–362.

[70] Jiawei Zhang, Haopeng Zhang, Congying Xia, and Li Sun. 2020. Graph-Bert: Only Attention is Needed for Learning Graph Representations. *arXiv e-prints*, Article arXiv:2001.05140 (Jan. 2020), arXiv:2001.05140 pages. arXiv:2001.05140 [cs.LG]

[71] Lei Zhang, Michael Färber, and Achim Rettinger. 2016. XKnowSearch!: Exploiting Knowledge Bases for Entity-based Cross-lingual Information Retrieval. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (2016).

[72] Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. Dgl-ke: Training knowledge graph embeddings at scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 739–748.

[73] Shuangjia Zheng, Jiahua Rao, Ying Song, Jixian Zhang, Xianglu Xiao, Evandro Fei Fang, Yuedong Yang, and Zhangming Niu. 2021. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Briefings in bioinformatics* 22, 4 (2021), bbaa344.

[74] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. *arXiv e-prints*, Article arXiv:2101.00774 (Jan. 2021), arXiv:2101.00774 pages. arXiv:2101.00774 [cs.AI]

[75] Zhaocheng Zhu, Shizhen Xu, Jian Tang, and Meng Qu. 2019. GraphVite: A High-Performance CPU-GPU Hybrid System for Node Embedding. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, Ling Liu, Ryen W. White, Amin Mantrach, Fabrizio Silvestri, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (Eds.). ACM, 2494–2504. https://doi.org/10.1145/3308558.3313508

[76] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal A. C. Xhonneux, and Jian Tang. 2021. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 29476–29490. https://proceedings.neurips.cc/paper/2021/hash/f6a673f09493afcd8b129a0bcf1cd5bc-Abstract.html