Wasserstein GAN for Knowledge Graph Completion with Continuous Learning

Erdem Önal

***Abstract*—Knowledge Graph Completion addresses the problem of missing relations in structured knowledge bases. Adversarial models can generate informative negative samples, but many approaches suffer from unstable optimization when applied to discrete graph data. Most existing models also assume static datasets and do not adapt as new information becomes available. This work presents a Wasserstein based adversarial model for Knowledge Graph Completion. The method is evaluated on the DBLP Computer Science Bibliography and generates valid RDF triples. The use of the Wasserstein distance leads to stable training behavior. The model is updated daily through an automated pipeline. Experiments conducted on a large dataset containing over 3.9 million entities show that the generated triples are largely novel and free of duplication.**

***Index Terms*—Knowledge Graph Completion, Wasserstein GAN, Continuous Learning, RDF**

I. Introduction

Knowledge graphs represent facts as triples composed of a head entity, a relation, and a tail entity. Large graphs such as the DBLP Computer Science Bibliography are inherently incomplete. Missing authorship, venue, or publication relations reduce their analytical value.

Knowledge Graph Completion aims to infer these missing facts by learning from existing structure. A common challenge in this task is the construction of effective negative samples. Uniform random sampling often produces trivial negatives that provide little learning signal. Prior work has shown that adversarial sampling can improve this process by generating more informative candidates.

Applying adversarial learning to discrete graph data poses optimization challenges. Gradient information can degrade during training, which may lead to unstable convergence. Wasserstein based objectives alleviate this issue by maintaining smoother gradient behavior throughout the optimization process.

Knowledge graphs also evolve as new data becomes available. Static embedding models are not well suited to reflect such changes. This work studies a system that combines adversarial learning with periodic model updates in order to handle evolving graph content.

II. Related Work

Early adversarial approaches such as KBGAN demonstrated that generated negative samples can improve link prediction performance. These methods rely on policy gradient techniques to handle discrete sampling. Such techniques often increase variance and slow down convergence.

Wasserstein based adversarial models were later introduced to improve optimization stability. By replacing divergence based objectives with the Wasserstein distance, these models reduce gradient saturation effects. Previous studies applied this idea to knowledge graph embeddings in static settings.

Continual learning for knowledge graphs has received comparatively less attention. Existing studies highlight the limitations of models trained on fixed snapshots. They argue for mechanisms that allow incremental updates as new entities and relations appear. The system described in this work follows this direction by retraining the model on a regular schedule.

III. Methodology

A. Data Processing

The DBLP Computer Science Bibliography serves as the data source. The raw XML dump contains several million records and cannot be loaded into memory directly. A streaming parser is therefore used to process the data sequentially. After filtering incomplete and low quality entries, over three point nine million publications are retained. These records are transformed into RDF triples representing authorship, publication venues, and temporal information.

B. WGAN Architecture

The model follows the Wasserstein GAN formulation applied to knowledge graph triples. The generator receives random noise together with a relation embedding and produces a candidate tail representation. The discriminator assigns a scalar score to each triple and estimates its consistency with the data distribution. Training minimizes the Wasserstein distance between real and generated samples under a Lipschitz constraint on the discriminator.

$$\min_{G} \max_{D \in \mathcal{P}} \mathbb{E}_{x \sim \mathbb{P}_r}[D(x)] - \mathbb{E}_{\tilde{x} \sim \mathbb{P}_g}[D(\tilde{x})] \tag{1}$$

Here $\mathcal{P}$ denotes the set of one Lipschitz functions and $\mathbb{P}_g$ is the distribution induced by the generator.

C. Continuous Learning Pipeline

Model updates are performed once per day using an automated workflow built on GitHub Actions. Each update resumes training from the most recent checkpoint and incorporates new DBLP records when available. Model parameters and embedding matrices are versioned using Git Large File Storage. This incremental strategy limits the computational overhead associated with repeated full retraining as the knowledge graph expands.

IV. Experiments and Results

A. Evaluation Metrics

Generated triples are evaluated using several complementary criteria. Novelty indicates whether a generated triple is absent from the training set. Uniqueness reflects how often identical triples are produced during generation. Training overlap measures the fraction of generated triples that exactly match known facts. Relation coverage reflects the usage of schema relations. Average distance represents the mean discriminator score assigned to generated triples.

B. Quantitative Results

Training exhibits stable convergence behavior across runs as illustrated in Figure 1. In the final experiment, one thousand triples were generated. The results are summarized in Table I.

TABLE I
Performance Metrics

Metric	Value
Novelty	100.0%
Uniqueness	100.0%
Training Overlap	0.0%
Relation Coverage	100.0%
Average Distance	0.5133

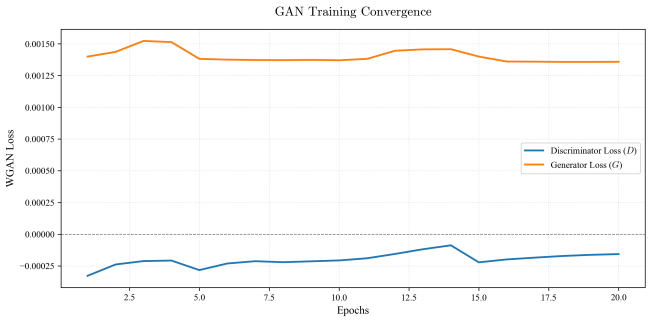

Fig. 1. Training loss convergence during WGAN optimization on the DBLP dataset

V. Discussion

The observed average distance of 0.5133 provides insight into the structure of the learned embedding space. In the Wasserstein setting, a low distance indicates that generated triples lie close to the distribution of observed facts. This suggests that the generator produces candidates that remain consistent with the relational patterns of the graph. At the same time, the high novelty score shows that these candidates are not replicas of training data. Together, these results indicate that the model explores regions of the latent space that correspond to plausible but previously unobserved relations. This behavior supports the use of the model for identifying likely missing links rather than reproducing known ones.

VI. Conclusion

This work presents a Wasserstein based adversarial approach to Knowledge Graph Completion with periodic model updates. The use of the Wasserstein distance leads to stable optimization on discrete graph data. Experiments on the DBLP dataset show that the model generates novel and unique triples. The results indicate that incremental retraining is a viable strategy for maintaining knowledge graph models as new data becomes available.

References

[1] L. Cai and W. Y. Wang, “KBGAN adversarial learning for knowledge graph embeddings,” in Proc. NAACL HLT, 2018.
[2] Y. Dai et al., “Generative adversarial networks based on Wasserstein distance for knowledge graph embeddings,” Knowledge Based Systems, 2020.
[3] A. Daruna et al., “Continual learning of knowledge graph embeddings,” IEEE Robotics and Automation Letters, 2021.