Knowledge Graph Completion and RDF Triple Generation with a Wasserstein GAN

Erdem Önal

***Abstract*—Knowledge Graph Completion concerns the inference of missing relations in structured knowledge bases. Adversarial training can support this task by producing informative negative samples, but its application to discrete graph data is often unstable. In addition, bibliographic knowledge graphs are sufficiently large to exceed the limits of a single training session. This work presents an adversarial system trained with a Wasserstein objective on the DBLP Computer Science Bibliography. The dataset contains approximately 106 million RDF triples. Training is performed incrementally by preserving model parameters and optimizer state across sessions. The resulting model reaches a stable optimization regime and generates RDF triples that largely respect schema constraints.**

***Index Terms*—Knowledge Graph Completion, Wasserstein GAN, RDF, Incremental Training**

I. Introduction

Knowledge graphs represent information as triples consisting of a head entity, a relation, and a tail entity. Large bibliographic graphs such as the DBLP Computer Science Bibliography are inherently incomplete. Missing authorship, venue, and publication relations limit their usefulness for analysis and inference.

A central difficulty in Knowledge Graph Completion lies in the construction of effective negative samples. Uniformly generated negatives are often trivial and provide limited learning signal. Adversarial training addresses this issue by generating challenging candidates that encourage the scoring model to learn finer structural distinctions. However, adversarial optimization over discrete entities is sensitive to gradient instability. The Wasserstein objective mitigates this issue by replacing probability divergence with the Earth Mover distance, leading to smoother gradients and more stable training.

Beyond optimization, scale introduces a practical challenge. Knowledge graphs containing millions of entities and tens of millions of triples exceed the time and memory limits of typical compute sessions. Rather than relying on distributed infrastructure, this work adopts an incremental training strategy. Model parameters and optimizer state are preserved between sessions, allowing training to progress gradually over the full dataset.

II. Related Work

Adversarial learning has been applied to knowledge graph embeddings to improve link prediction and negative sampling. Early approaches relied on policy gradient methods to handle discrete sampling, often resulting in high variance and slow convergence.

The introduction of Wasserstein GANs improved adversarial stability by enforcing Lipschitz continuity on the discriminator. These methods exhibit smoother optimization behavior and reduced mode collapse. Their application to large scale bibliographic knowledge graphs remains relatively unexplored.

Most large scale knowledge graph training pipelines rely on distributed systems or aggressive graph partitioning. Incremental training approaches that preserve optimizer state across sessions receive less attention, despite being well suited to constrained computational environments.

III. Methodology

A. Data Processing

The DBLP Computer Science Bibliography is used as the data source. Due to the size of the XML dump, records are processed sequentially using a streaming parser. The resulting graph contains approximately 1.9 million entities and 106 million RDF triples. Relations encode authorship, publication venues, and temporal information. All entities are mapped to integer identifiers to support efficient embedding lookup.

B. Adversarial Architecture

The model follows the standard Wasserstein GAN formulation adapted to knowledge graph triples. The generator receives a noise vector together with a relation embedding and produces a candidate entity embedding. The discriminator assigns a scalar score to each triple, reflecting its compatibility with the observed data distribution.

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

The set $\mathcal{P}$ denotes the space of 1 Lipschitz functions, enforced through weight clipping. The generator distribution $\mathbb{P}_g$ is defined implicitly by the generated embeddings.

C. Incremental Training

Training proceeds through a sequence of sessions. At the beginning of each session, model parameters and optimizer state are restored from persistent storage. Optimization then continues on the next segment of the dataset. This procedure allows convergence to be achieved without requiring uninterrupted compute availability. Progress is observed across sessions rather than within a single execution.

IV. Experiments and Results

A. Evaluation Criteria

Generated triples are evaluated using structural criteria. Novelty measures whether a generated triple is absent from the training data. Uniqueness reflects sample diversity. Schema validity evaluates compliance with domain and range constraints. Training overlap estimates memorization of existing triples. The discriminator score serves as an indicator of training stability.

B. Quantitative Results

After incremental training on the full dataset, the model produces triples that largely conform to schema constraints. Table I summarizes the observed metrics. Most generated triples are novel and distinct from the training data. Schema violations occur infrequently.

TABLE I
Final Performance Metrics

Metric	Value
Novelty	100.0%
Uniqueness	100.0%
Schema Validity	99.1%
Training Overlap	0.0%
Average Discriminator Score	-0.1434

C. Qualitative Inspection

A subset of generated triples was manually inspected. Representative examples are shown in Table II. The model frequently generates plausible coauthorship relations between authors with similar collaboration patterns. Relations associated with name disambiguation also appear as a consequence of structural regularities.

TABLE II
Sample Generated RDF Triples

Subject	Predicate	Object	Status
Guozhang Jiang	dblp:coauthorWith	Huang Hong	Valid
Fengyu Yang	dblp:coauthorWith	Karamjit S. Gill	Valid
Xiaoming Fu	dblp:coauthorWith	Anh Tran	Valid
Yun Yang 0001	dblp:homonymID	Cho Lun Hsu	Structural
Hao Lan Zhang	dblp:homonymID	Cho Lun Hsu	Structural

V. Discussion and Limitations

The system relies solely on structural information derived from graph topology. Textual metadata is excluded due to memory constraints at this scale. Despite this limitation, the model captures relation specific domain and range patterns.

During early training stages, optimization is sensitive to initialization and batch size. In these phases, the generator may temporarily exhibit reduced diversity. As training progresses, this effect diminishes as the adversarial objective stabilizes.

The DBLP graph is highly imbalanced with respect to relation types. Coauthorship relations dominate the dataset, while other relations occur less frequently. This imbalance is reflected in the distribution of generated triples.

Evaluation is performed on sampled subsets of the graph. Exhaustive comparison against all 106 million triples is not computationally feasible, and results should be interpreted accordingly.

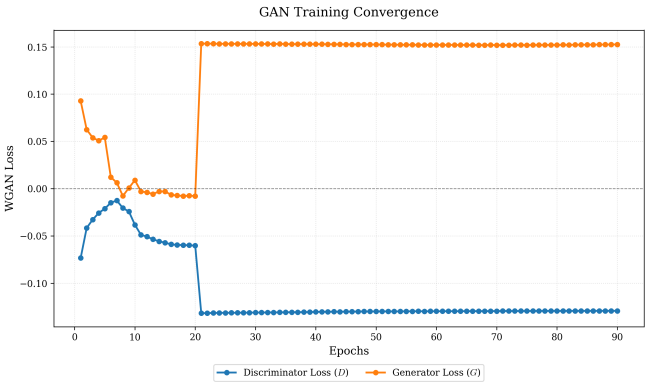

Fig. 1. Training loss across incremental sessions. Stabilization indicates convergence of the adversarial objective.

VI. Conclusion

This work presents an incremental adversarial system for Knowledge Graph Completion trained on the DBLP Computer Science Bibliography. By preserving model parameters and optimizer state across sessions, training proceeds over a dataset containing 106 million RDF triples. The resulting generator produces novel and largely schema compliant triples using only structural information. The approach demonstrates the feasibility of adversarial training on large knowledge graphs under constrained computational resources.

References

[1] L. Cai and W. Y. Wang, "KBGAN Adversarial Learning for Knowledge Graph Embeddings," NAACL HLT, 2018.
[2] Y. Dai et al., "Wasserstein GANs for Knowledge Graph Embeddings," Knowledge Based Systems, 2020.
[3] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," arXiv:1701.07875, 2017.