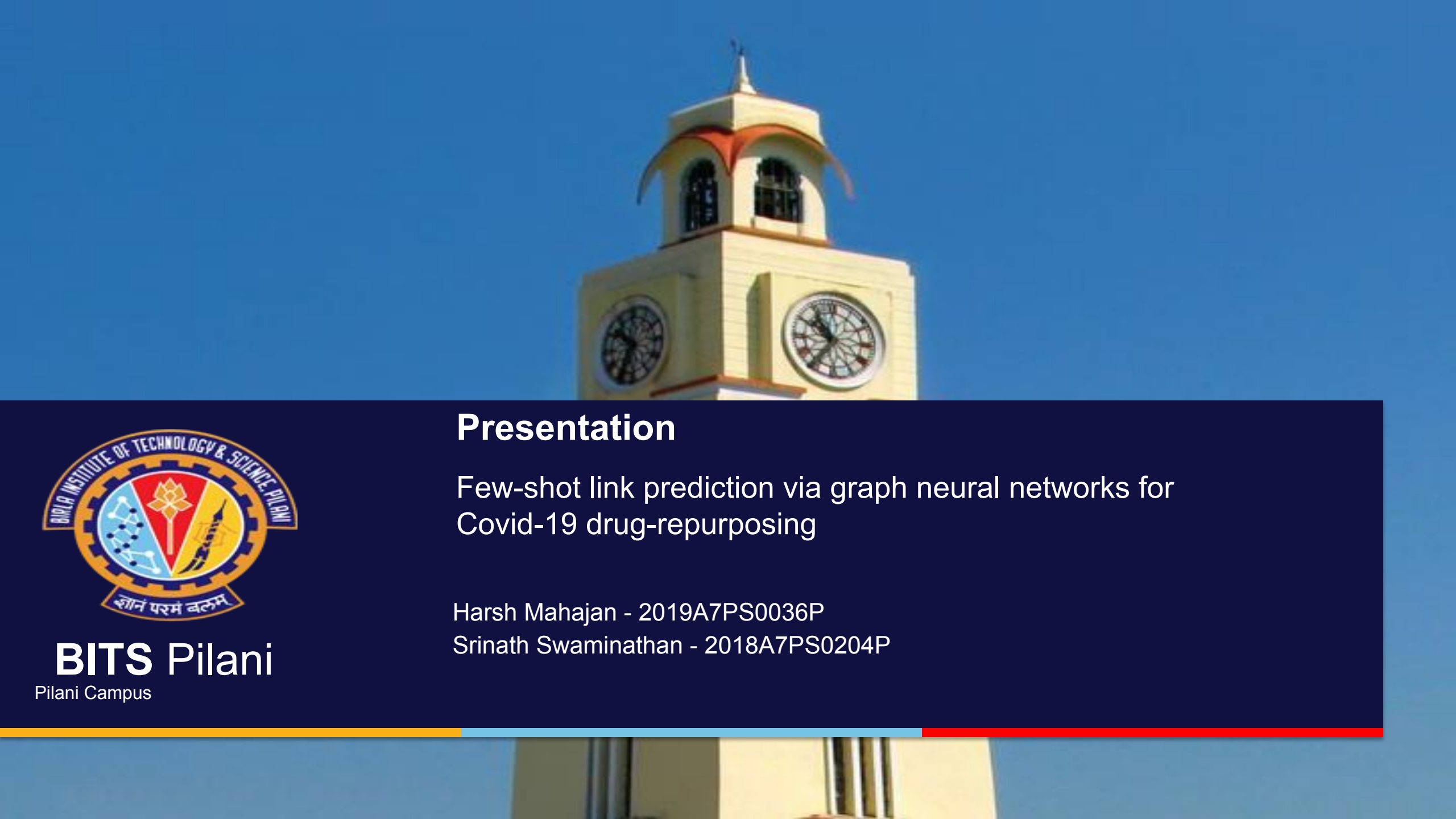

Presentation

Few-shot link prediction via graph neural networks for Covid-19 drug-repurposing

Harsh Mahajan - 2019A7PS0036P
Srinath Swaminathan - 2018A7PS0204P

BITS Pilani
Pilani Campus

Introduction

Drug Repurposing

- New drug discovery - slow pace and substantial costs
- Drug Repurposing - new medical uses for existing drugs
- Existing drugs have already undergone Phase 1 clinical trials (drug safety and side effects). This reduces the drug development timeline and considerably cuts R&D costs.

Drug Repurposing as Link Prediction

- The crux of drug repurposing is to identify interactions b/w biological entities such as genes present in drugs and diseases
- This can be modeled into a link prediction problem over the biological network
  - ❑ Nodes - Biological Entities
  - ❑ Edges - Interactions b/w the entities

Few-Shot Link Prediction

- For novel diseases like COVID-19, only a few interactions are available b/w viral proteins and chemical compounds that inhibit required genes
- Therefore, the edge-type b/w the compound and viral protein can be considered as rare.
- This gives rise to the few-shot link prediction framework, where a model is required to predict links of rare type.

Drug Repurposing Knowledge Graph (DRKG)

- Comprehensive biological knowledge graph relating genes, compounds, diseases, biological processes, side effects and symptoms.
- Created to assist various drug-repurposing techniques
- Consists of 13 entity/node types and 107 relation/edge types

Problem Formulation

Given

Consider the heterogeneous graph with T node types and R relation types defined as $\mathcal{G} := \{\{\mathcal{V}_t\}_{t=1}^{T}, \{\mathcal{E}_r\}_{r=1}^{R}\}$

Any node type t is defined as $\mathcal{V}_t := \{v_n^t\}_{n=1}^{N_t}$

Additionally, each node n_t is associated with a F × 1 feature vector x_{nt}

Any relation type r is defined as $\mathcal{E}_r := \{(v_n^t, v_n^{t'}) \in \mathcal{N}^t \times \mathcal{N}^{t'}\}$

Few-Shot Link Prediction

Given R − 1 sets of edges $\{\mathcal{E}_r\}_{r=1}^{R-1}$, a nodal attribute vector x_{nt} per node n_t, and a small set of links in the few-shot relation $\mathcal{E}_R$ with $|\mathcal{E}_R| \leq K$, the few-shot link prediction amounts to inferring the missing links of the rare type R

Limitations of SOTA Methods

RGCNs

- In the case of RGCNs, the relation embedding hr is directly trained from the loss function.
- However, this does not generalize well when only a few training links are present.
- This creates a need to learn the embeddings inductively rather than directly training from the loss function

Approach

Basic Idea

- When sufficient training examples are not available, a relation embedding can be thought of as the concatenation and aggregation of the node pairs participating in the relation

- The node embeddings are trained with sufficient examples from all its participating relations. This can be used to construct the embeddings for the rare relation type

I-RGCN - Node Embeddings

- I-RGCN is built on the standard RGCN network. The node embeddings are calculated in the same way as in RGCN

$$h_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r}\frac{1}{c_{i,r}}W_r^{(l)}h_j^{(l)} + W_0^{(l)}h_i^{(l)}\right)$$

- where N_i^r denotes the set of neighbor indices of node i under relation $r \in R$. $c_{i,r}$ is a problem-specific normalization constant.

I-RGCN - Relational Embeddings

- The relation embedding h_r for a relation r is calculated through an MLP.

$$\mathbf{h}_r := \frac{1}{|\mathcal{E}_r|} \sum_{(n_t, n_{t'}) \in \mathcal{E}_r} \sigma(\mathbf{W}_1 \sigma(\mathbf{W}_2(\mathbf{h}_{n_t} \| \mathbf{h}_{n_{t'}}))$$

- where || denotes the vector concatenation. W_1 and W_2 are learnable parameters

Loss Function

- This model is supervised by a DisMult loss function similar to that in RGCN

$$\mathcal{L}_{\text{FSLP}} := \log(1 + \exp(-y \times \mathbf{h}_{n_t}^{\top} \text{diag}(\mathbf{h}_r) \mathbf{h}_{n_{t'}}))$$

- Where y is −1 for the negative triples and 1 for the positive ones.

$$\mathcal{L} = -\frac{1}{(1+\omega)|\hat{\mathcal{E}}|} \sum_{(s,r,o,y)\in\mathcal{T}} y \log l\big(f(s,r,o)\big) + (1-y) \log\big(1 - l\big(f(s,r,o)\big)\big)\,, \tag{7}$$

where $\mathcal{T}$ is the total set of real and corrupted triples, l is the logistic sigmoid function, and y is an indicator set to $y = 1$ for positive triples and $y = 0$ for negative ones.

Evaluation Metrics

Mean Reciprocal Rank (MRR)

- The model outputs a ranked list of possible candidates based on a target variable.
- For a single query, reciprocal rank = 1 / rank, where rank is the position of the highest-ranked valid candidate.
- For multiple queries Q, the Mean Reciprocal Rank is the mean of Q reciprocal ranks.

$$\mathrm{MRR} = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{\mathrm{rank}_i}$$

Hits@k

- Let n be the number of valid candidates in the list of top k candidates ranked in descending order of score.
- Then Hits@k = n / k
- This paper uses Hits@1, Hits@3, and Hits@10

Algorithm and Code

Improvement

Limitation of DistMult

- DRKG consists of asymmetric relations. For example, we can have a (drug, treats, disease) relation but not a (disease, treats, drug) relation

- However, DisMult cannot handle asymmetric relations as it is just an element wise product of the embeddings

Suggestion: ComplEx

- ComplEx decoder makes use of complex numbers and their conjugates

- ComplEx(z_u, r, z_v) = Re(z_u * r * conjugate(z_v)) where * denotes element-wise multiplication

ComplEx Implementation

Application to COVID-19 Drug Repurposing

- Drug-repurposing can be viewed as a few-shot link prediction task since only a few edges are available related to novel diseases in the DRKG

- The paper aims at predicting links among gene entities associated with the target disease and drug entities

- The paper uses corona-virus related diseases, including SARS, MERS and SARS-COV2, as target diseases representing Covid-19 as their functionality is similar

- FDA-approved drugs in Drugbank are selected as candidates, excluding for simplicity drugs with molecular weight less than 250 daltons, as many of certain drugs are actually health drugs. This amounts to 8104 candidate drugs.

- 442 target genes related to COVID are chosen. The model scores triplets and generates a rank list per target gene

- These ranked lists are checked the overlap among the top 100 predicted drugs and the drugs used in clinical trials per gene.

I-RGCN		RGCN	
Drug name	# hits	Drug name	# hits
Dexamethasone	240	Chloroquine	69
Ribavirin	142	Colchicine	41
Colchicine	128	Tetrandrine	40
Chloroquine	115	Oseltamivir	37
Methylprednisolone	86	Azithromycin	36
Tofacitinib	75	Tofacitinib	33
Thalidomide	70	Ribavirin	32
Losartan	64	Methylprednisolone	30
Hydroxychloroquine	48	Deferoxamine	30
Oseltamivir	46	Thalidomide	25
Deferoxamine	34	Dexamethasone	24
Ruxolitinib	23	Bevacizumab	21
Azithromycin	23	Hydroxychloroquine	19
Nivolumab	11	Losartan	19
Tradipitant	11	Ruxolitinib	13
Bevacizumab	10	Eculizumab	12
Eculizumab	7	Tocilizumab	11
Baricitinib	6	Anakinra	11
Sarilumab	6	Sarilumab	8
Tetrandrine	6	Nivolumab	6

Thank You