Compositional Tokenization in Knowledge Graphs

Michael Galkin
Postdoctoral Fellow @ Mila & McGill

Ulm

Q3012

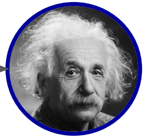
Nobel Prize

Q38104

University of Zurich

Q206702

Albert Einstein



Q937

The ImageNet Moment for KGs

Self-supervised pre-training → Fine-tuning on a downstream task

Graph ML

NLP

Vision

Mila

The ImageNet Moment for KGs

Self-supervised pre-training → **Fine-tuning on a downstream task**

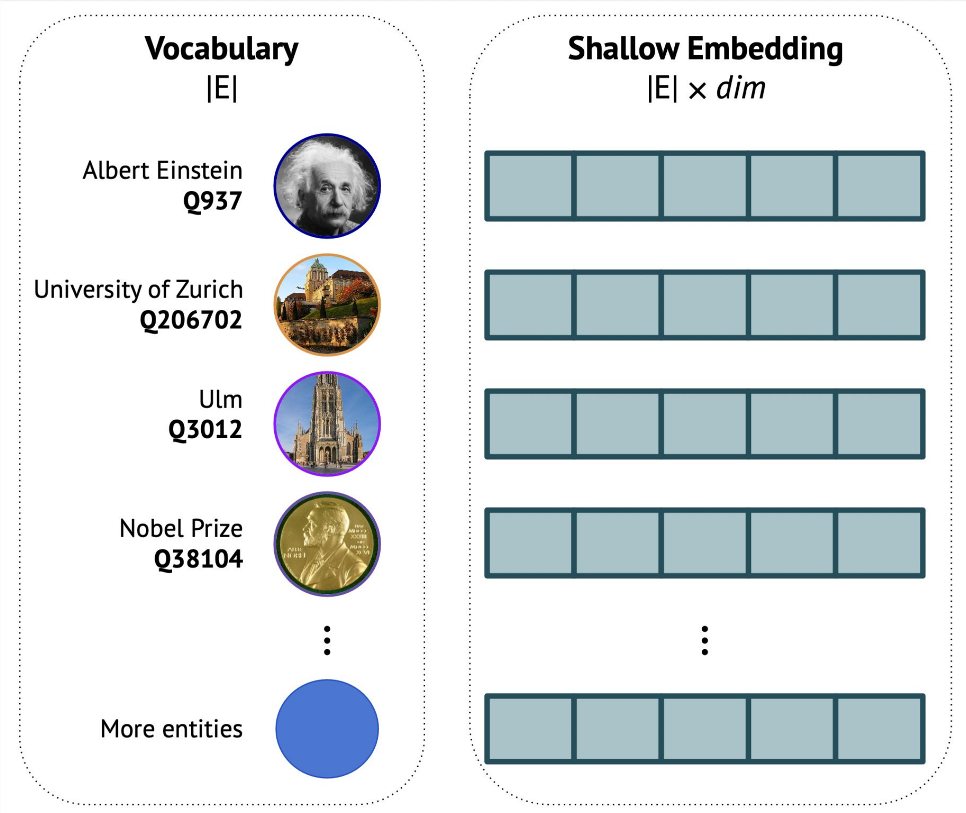
Wikidata: 100M nodes
Embs: [100M, dim] ?

PyTorch BigGraph

~200 GB

- Graph ML
- NLP
- Vision

Shallow Embedding



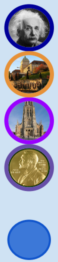
Transductive vs Inductive

Transductive

Inductive

Training

Vocab



Inference

New, unseen nodes (entities)

- Added to the seen graph
- Completely new inference graph

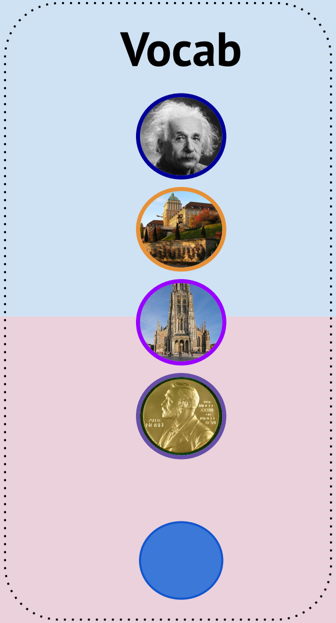
Transductive vs Inductive

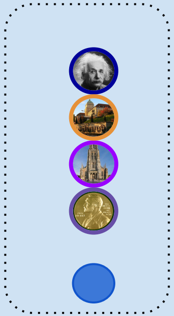
Shallow embeddings

Transductive

Inductive

Training





Inference

New, unseen nodes (entities)

- Added to the seen graph
- Completely new inference graph

OGB WikiKG: Just 2.5M nodes

Mila

Leaderboard for ogbl-wikikg2

The MRR score on the test and validation sets. The higher, the better.

Package: >=1.2.4

Deprecated `ogbl-wikikg` leaderboard can be found here.

BERT-Large is ~340M params

Rank	Method	Test MRR	Validation MRR	Contact	References	#Params	Hardware	Date
1	**PairRE (200dim)**	0.5208 ± 0.0027	0.5423 ± 0.0020	Linlin Chao	Paper, Code	500,334,800	Tesla P100 (16GB GPU)	Jan 28, 2021
2	RotatE (250dim)	0.4332 ± 0.0025	0.4353 ± 0.0028	Hongyu Ren – OGB team	Paper, Code	1,250,435,750	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021
3	TransE (500dim)	0.4256 ± 0.0030	0.4272 ± 0.0030	Hongyu Ren – OGB team	Paper, Code	1,250,569,500	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021
4	ComplEx (250dim)	0.4027 ± 0.0027	0.3759 ± 0.0016	Hongyu Ren – OGB team	Paper, Code	1,250,569,500	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021

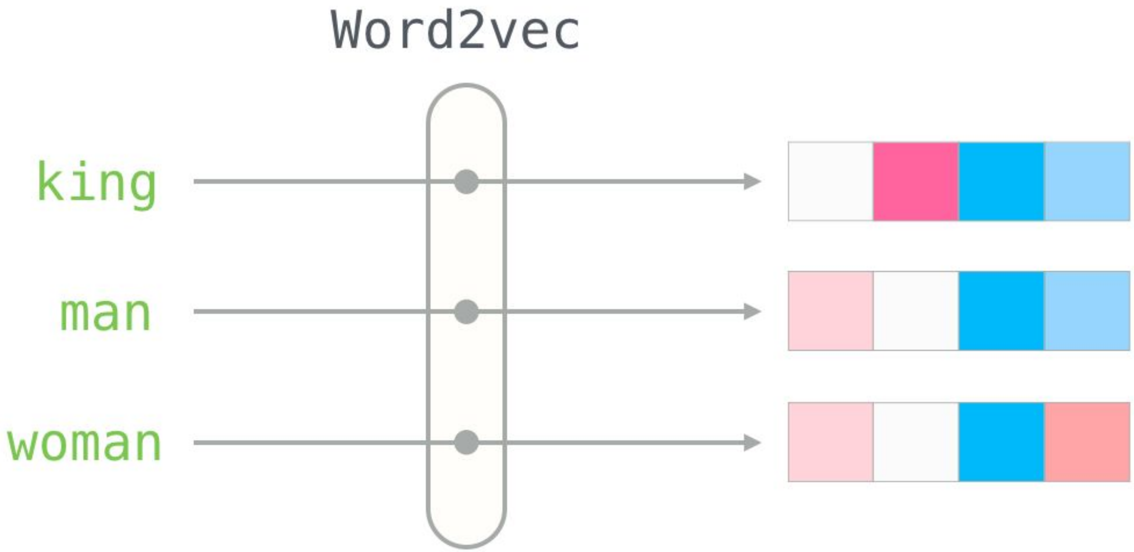
BERT	(340M params)	- disruption in NLP ✅
KG embs	(>1B params)	- 🥴

Life beyond shallow embedding?

Do we really need to learn & store the whole shallow embedding matrix $|E| \times dim$

Trying to fit a 100M x 200 tensor on a Tesla V100 ->

Back to 2014



Unseen words = [OOV] (out-of-vocabulary)

Source: The Illustrated Word2vec by Jay Alammar

Byte-Pair Encoding / WordPiece

Dictionary

5 l o w
2 l o w e r
6 n e w **est**
3 w i d **est**

Vocabulary

l, o, w, e, r, n, w, s, t, i, d, es, **est**

Add a pair (es, t) with freq 9

Byte-Pair Encoding vocabulary. Source: [Sennrich et al]() and [CS224N]()

Byte-Pair Encoding / WordPiece

"I love tacos, apples, and tea!"

i	love	tacos	,	app	##les	,	and	t	##e	##a	!
6	7	8	5	10	11	5	9	30	41	37	3

Byte-Pair Encoding / WordPiece

"I love tacos, apples, and tea!"

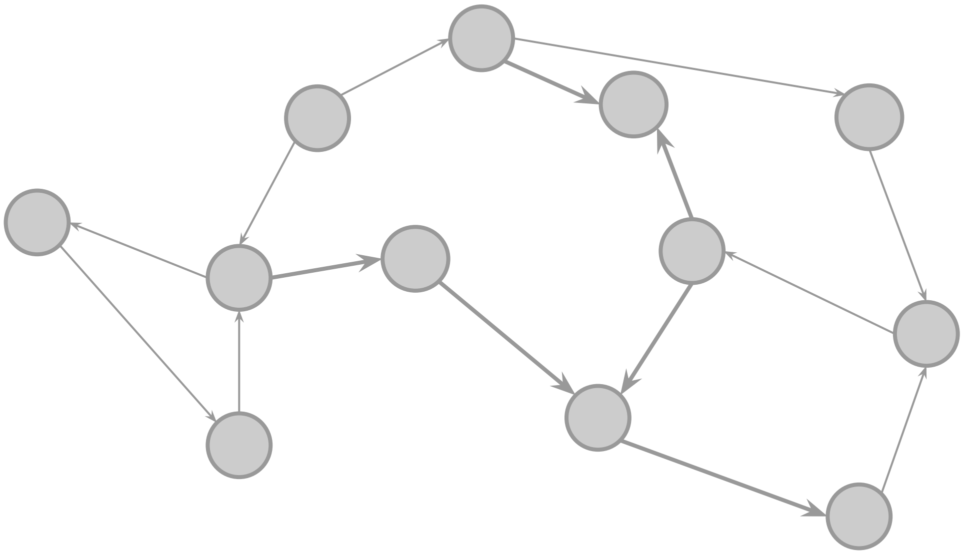
i	love	tacos	,	app	##les	,	and	t	##e	##a	!
6	7	8	5	10	11	5	9	30	41	37	3

- Fixed-size vocab of subword units (30-50K)
- We can tokenize any unseen word

Tokenizing KGs

BERT-Large (340M)	KG Embedding (1250M)
Encoder (Transformer) ~300M	Vocabulary 2.5M x 500d
Vocabulary 30K x 1024d	

Tokenization + Graphs?



If nodes in a graph are "**words**", can we design a fixed-size vocab of "**sub-word**" units?

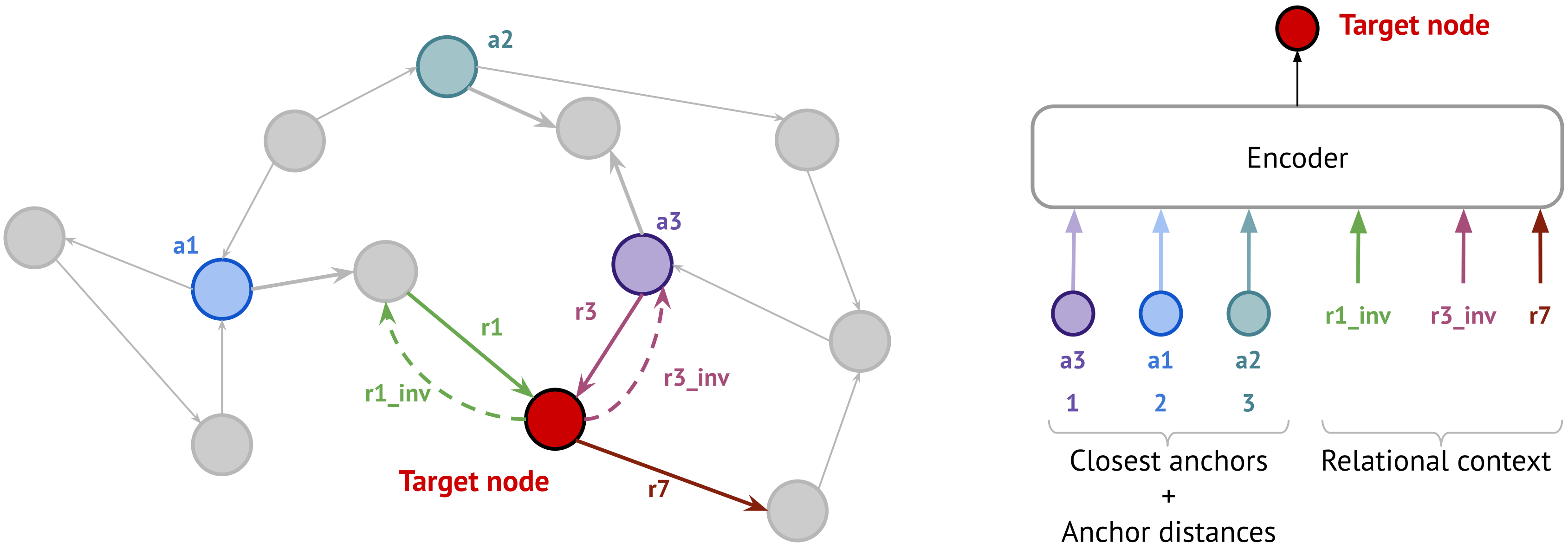
Tokenizing KGs

BERT-Large (340M)	NodePiece	KG Embedding (1250M)
Encoder (Transformer) ~300M	Encoder	Vocabulary 2.5M x 500d
Vocabulary 30K x 1024d	Vocab: K anchors, All relations	

Tokenizing KGs

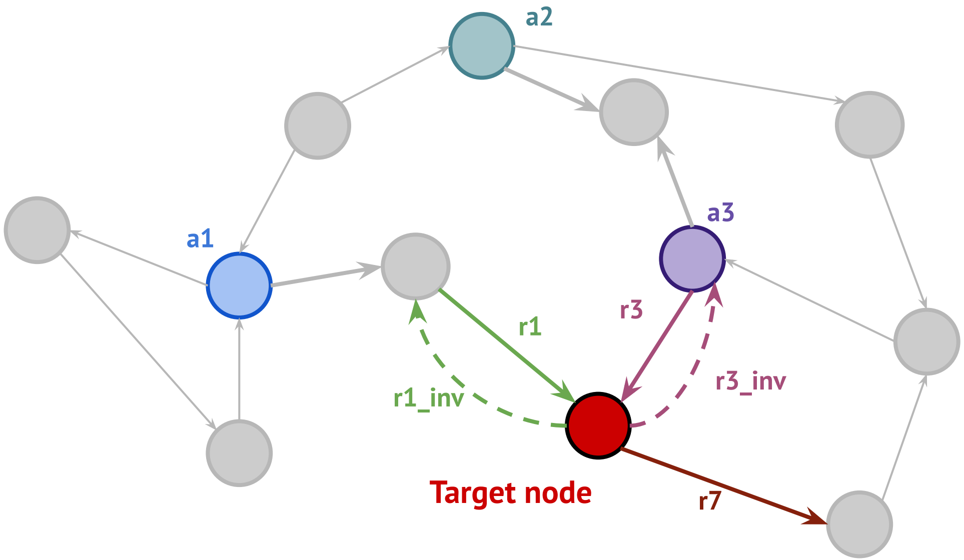
	Shallow embedding, only known words, otherwise OOV	Compositional representations, subword units
Language	Word2vec, GloVe	Byte-Pair Encoding, WordPiece
Graphs	All KG embedding algorithms (TransE, etc)	**NodePiece**

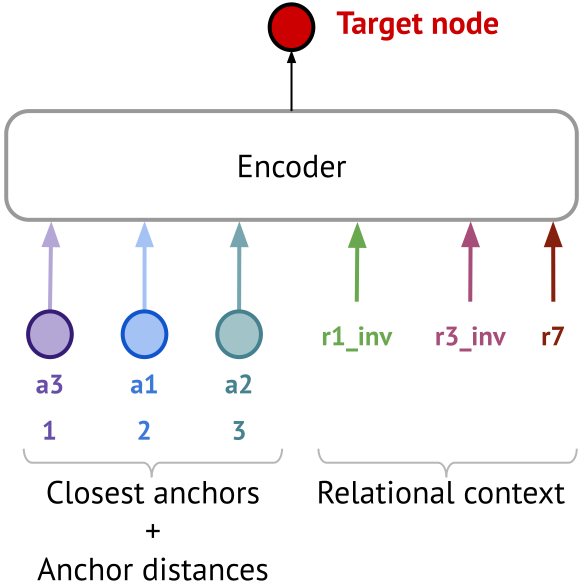
NodePiece - "*subword units*" for KGs



Vocabulary = Anchors + Relation types

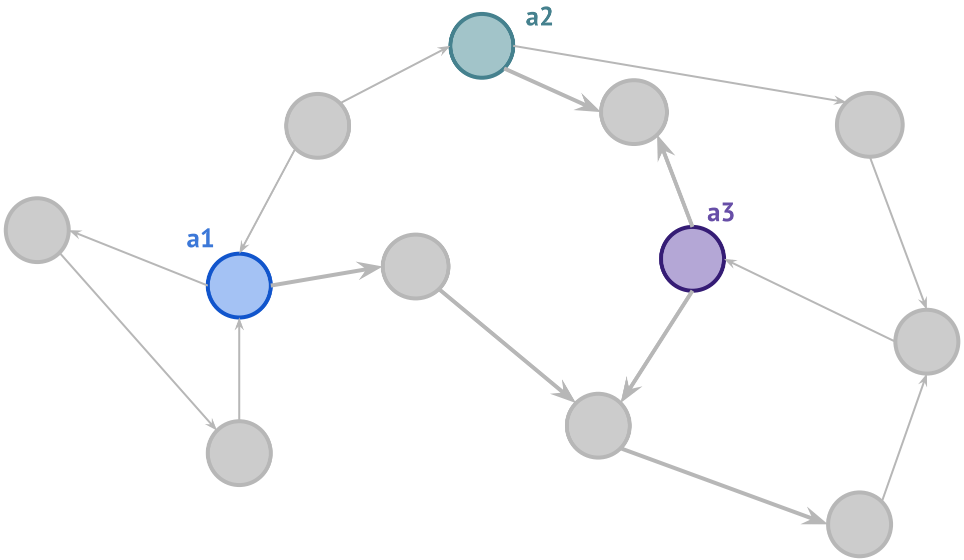
NodePiece - “*subword units*” for KGs





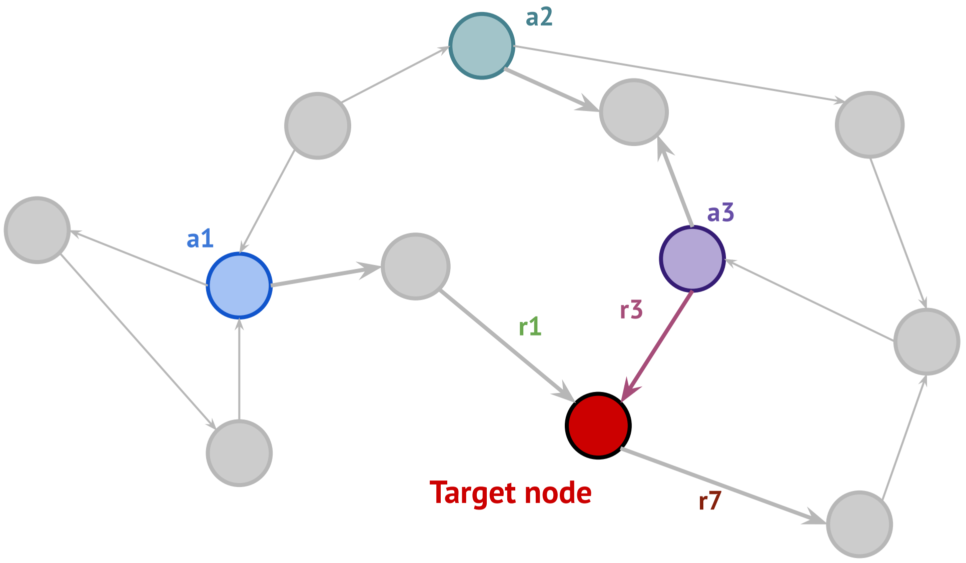
Inductive out-of-the-box: unseen nodes are “tokenized” with the same Vocab

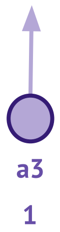
Anchor Node Selection



Current strategy:
40% top degrees
40% top PPR
20% random

Tokenization

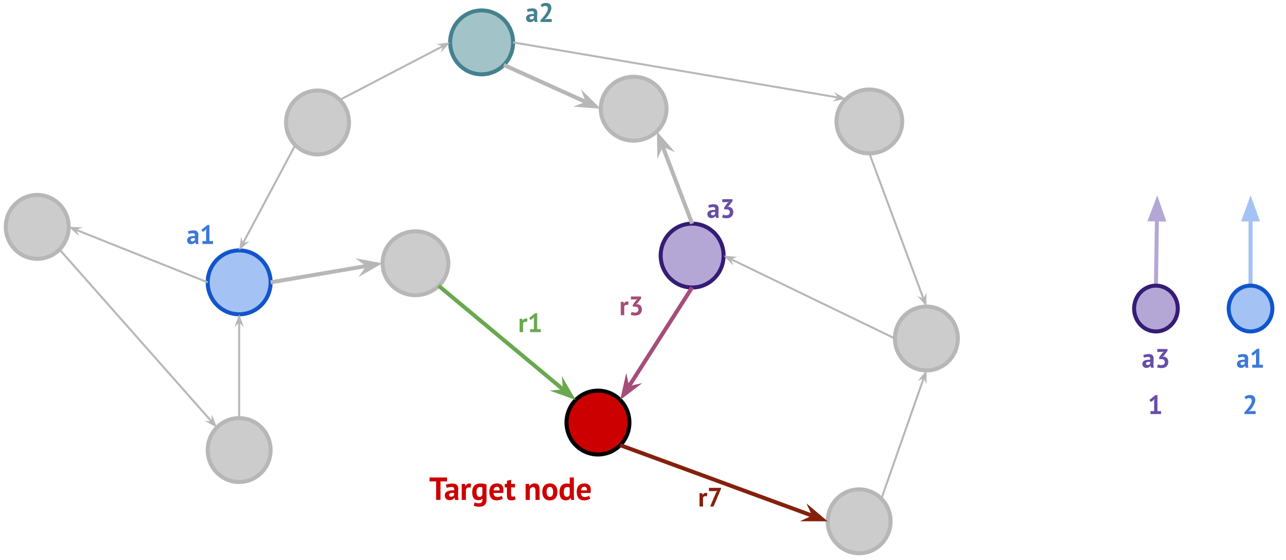




BFS from the target node until we reach |K| anchors

- Can be done in forward pass
- Can be pre-processed and saved

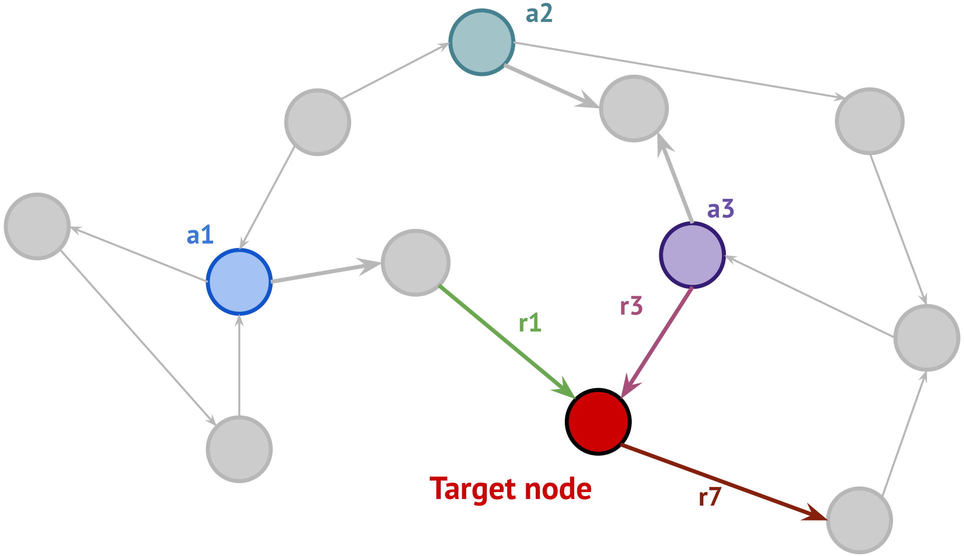
Tokenization

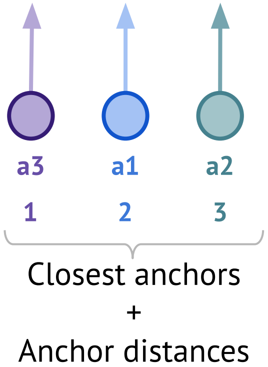


BFS from the target node until we reach |K| anchors

- Can be done in forward pass
- Can be pre-processed and saved

Tokenization

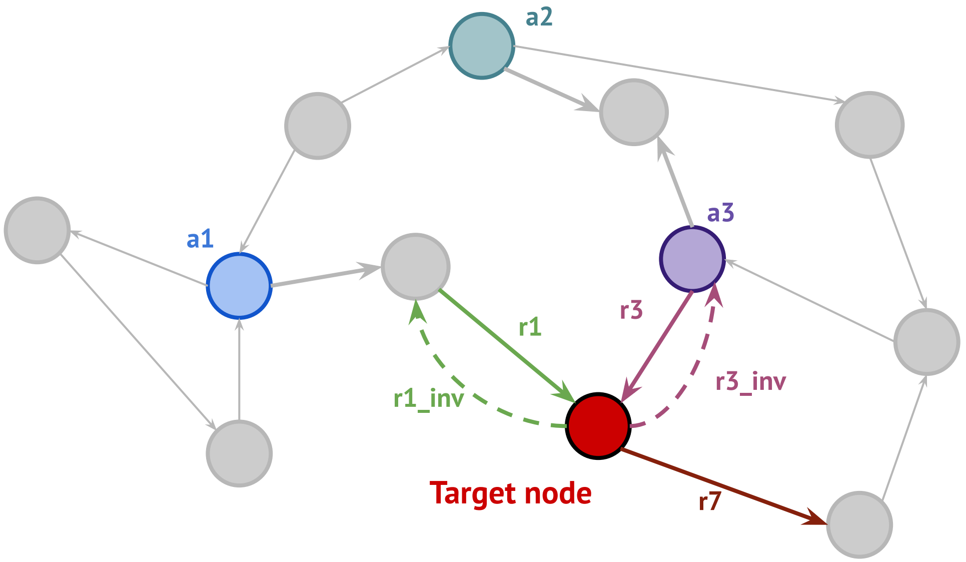


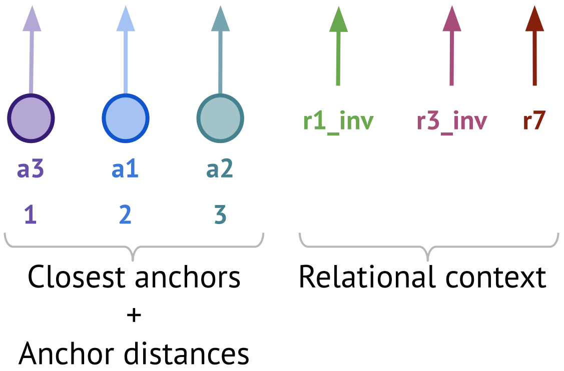


BFS from the target node until we reach |K| anchors

- Can be done in forward pass
- Can be pre-processed and saved

Tokenization

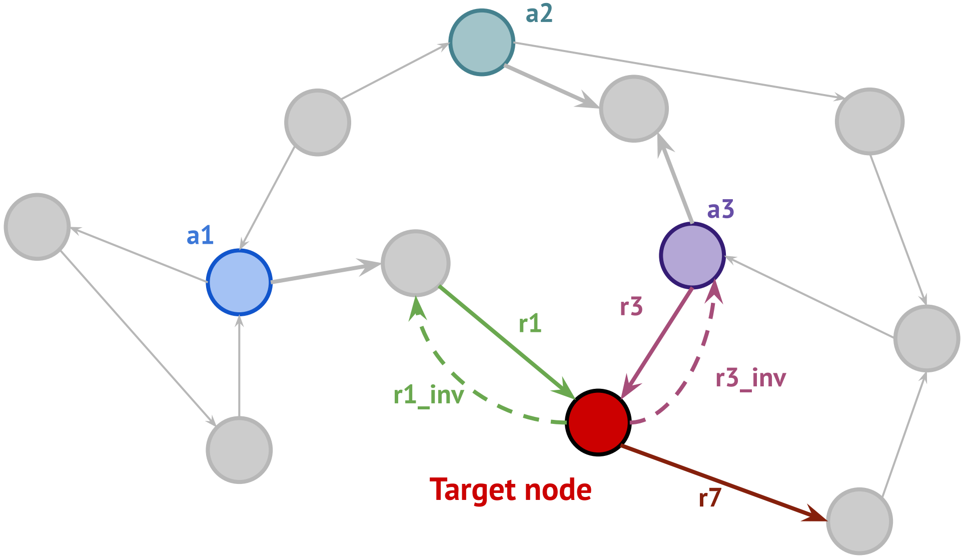


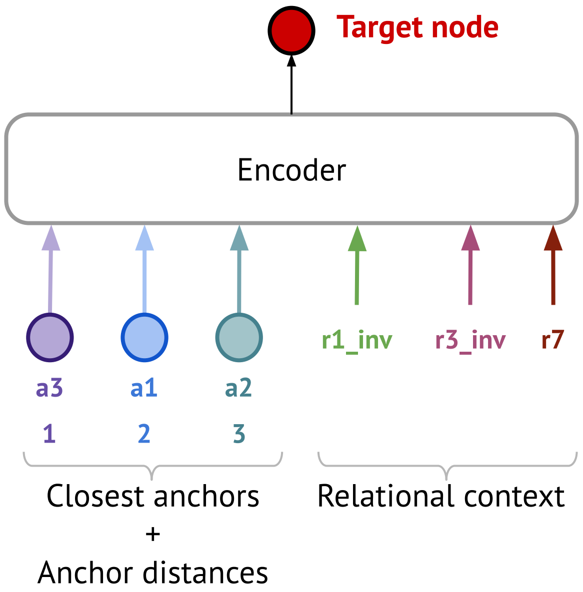


BFS from the target node until we reach |K| anchors

- Can be done in forward pass
- Can be pre-processed and saved

Tokenization

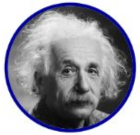




BFS from the target node until we reach |K| anchors

- Can be done in forward pass
- Can be pre-processed and saved

Tokenizing Einstein



Albert Einstein

$= f\Bigg($

Ulm ,

Nobel Prize ,

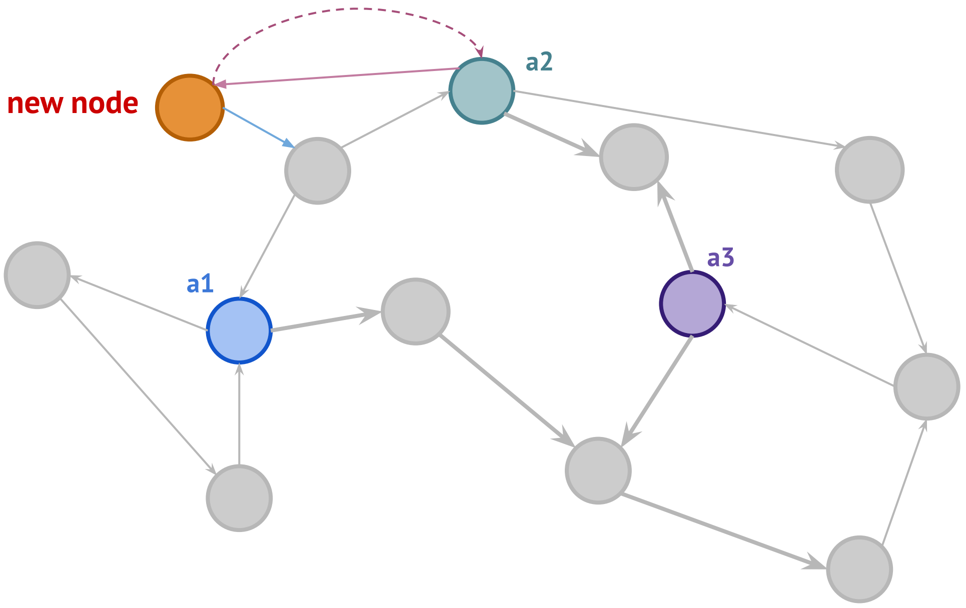
Theoretical Physics ,

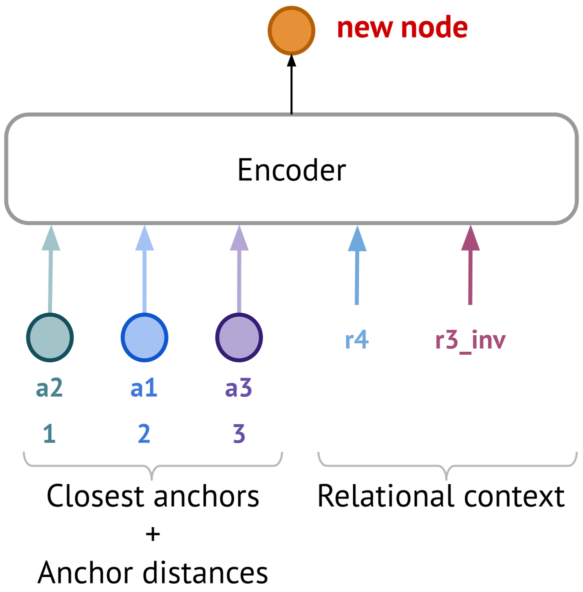
place of birth **P19** , award received **P166** , occupation **P106** , academic degree **P512** $\Bigg)$

3 nearest anchors

4 unique outgoing relations in the context

Unseen Node Tokenization





Tokenization Speed

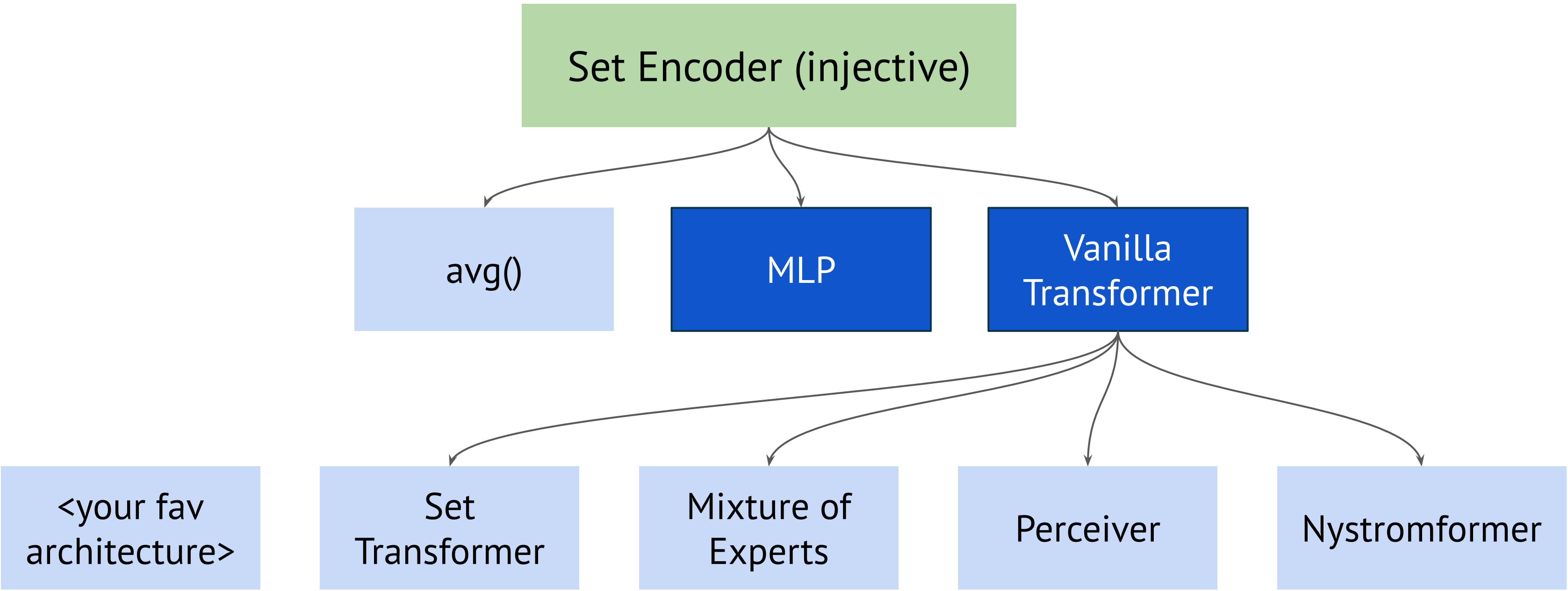
	15K nodes, 270K edges	40K nodes, 80K edges	120K nodes, 1M edges	2.5M nodes, 16M edges	5M nodes, 40M edges
Time	8 sec	30 sec	4.5 min	2-8 hours	3-9 hours
Size	7.5 MB	20 MB	40 MB	700 MB	1 GB

Single core, laptop CPU

METIS partitioning

Parallel pre-processing

Set Encoder


Set Encoder (injective)
avg()
MLP
Vanilla Transformer
<your fav architecture>
Set Transformer
Mixture of Experts
Perceiver
Nystromformer

Tasks

Link Prediction / Relation Prediction

$$score(h, r, t)$$

TransE

RotatE

...

Any GNN works, too!

$$\begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_N \end{bmatrix} + \boxed{\text{GNN}}$$

Transductive Link Prediction

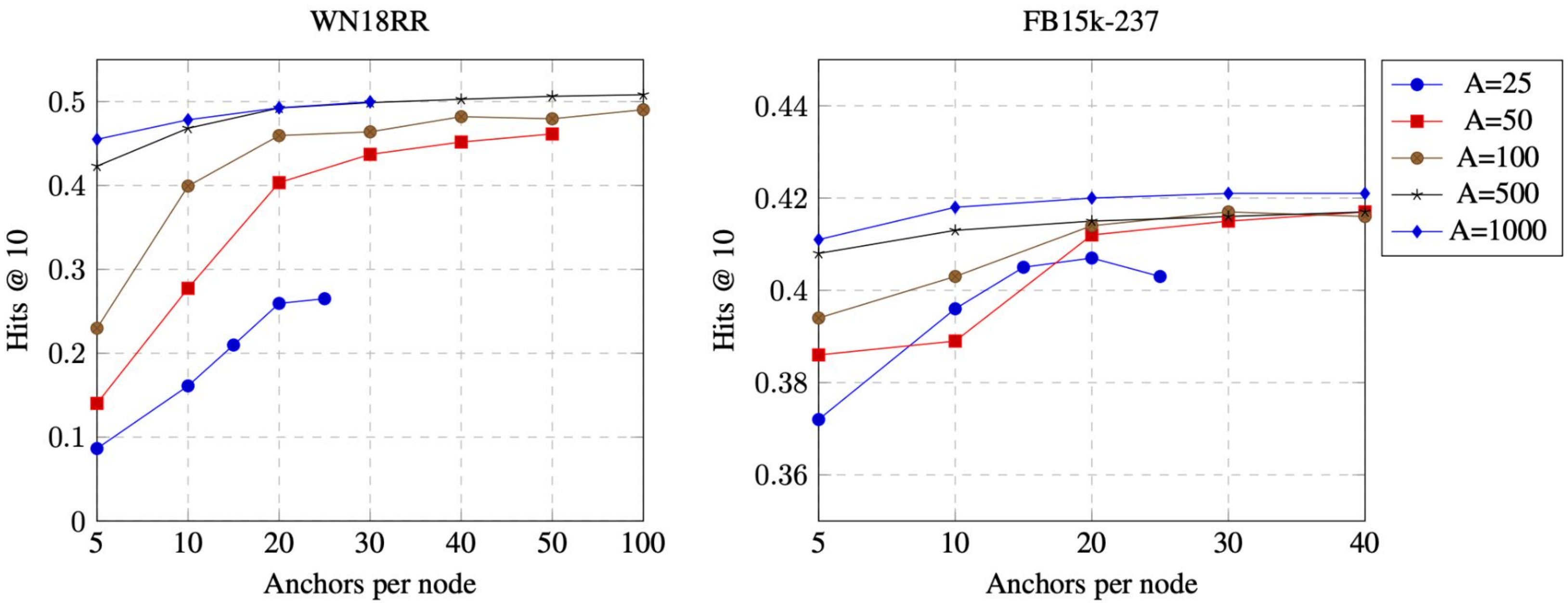


Figure 2: Combinations of total anchors A and anchors per node. Denser FB15k-237 saturates faster on smaller A while sparse WN18RR saturates at around 500 anchors.

Transductive Link Prediction

Table 3: Transductive link prediction on smaller KGs. † results taken from [38]. $|V|$ denotes vocabulary size (anchors + relations), #P is a total parameter count (millions). % denotes the Hits@10 ratio based on the strongest model.

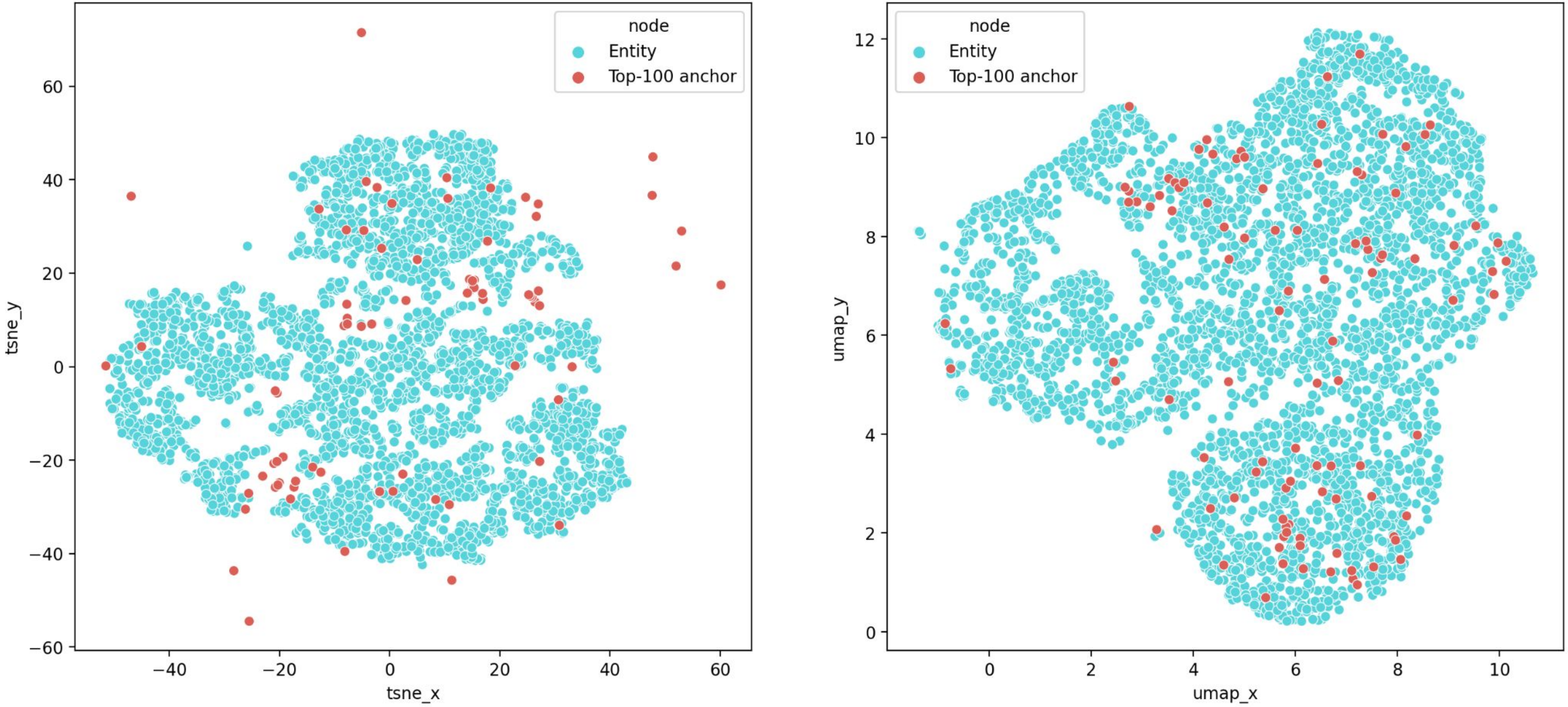
	FB15k-237					WN18RR				
	$\|V\|$	#P (M)	MRR	H@10	%	$\|V\|$	#P (M)	MRR	H@10	%
RotatE	15k + 0.5k	29	0.338†	0.533†	100	40k + 22	41	0.476†	0.571†	100
NodePiece + RotatE	1k + 0.5k	3.2	0.256	0.420	79	500 + 22	4.4	0.403	0.515	90
- no rel. context	1k + 0.5k	2	0.258	0.425	80	500 + 22	4.2	0.266	0.465	81
- no distances	1k + 0.5k	3.2	0.254	0.421	79	500 + 22	4.4	0.391	0.510	89
- no anchors, rels only	0 + 0.5k	1.4	0.204	0.355	67	0 + 22	0.3	0.011	0.019	0.3

Transductive Link Prediction

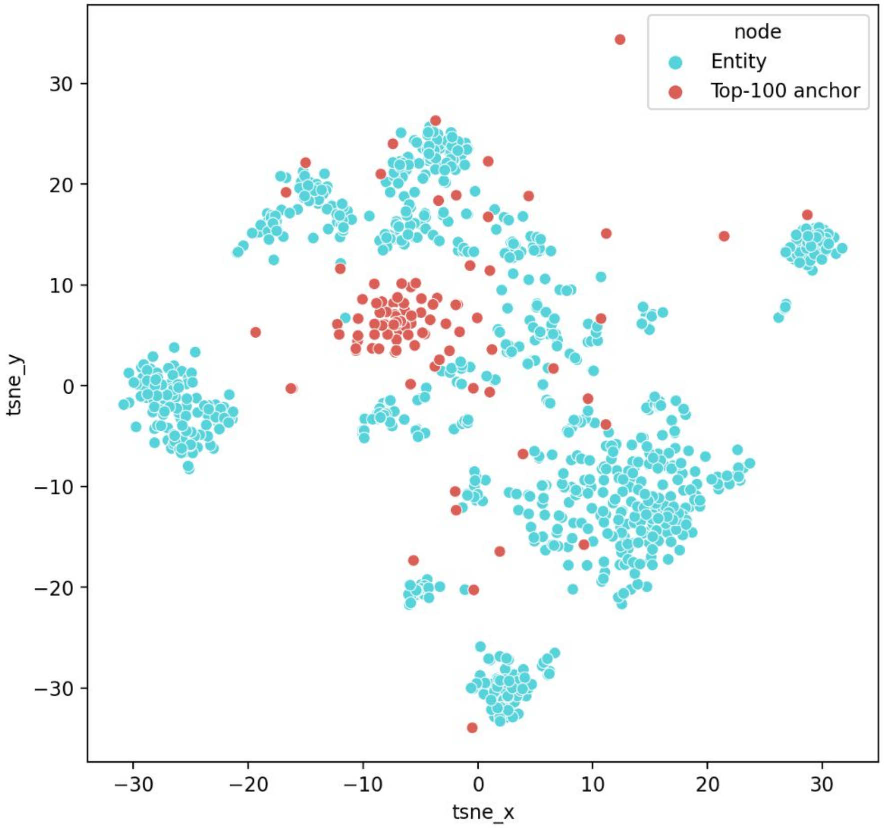
Table 3: Transductive link prediction on smaller KGs. † results taken from [38]. $|V|$ denotes vocabulary size (anchors + relations), #P is a total parameter count (millions). % denotes the Hits@10 ratio based on the strongest model.

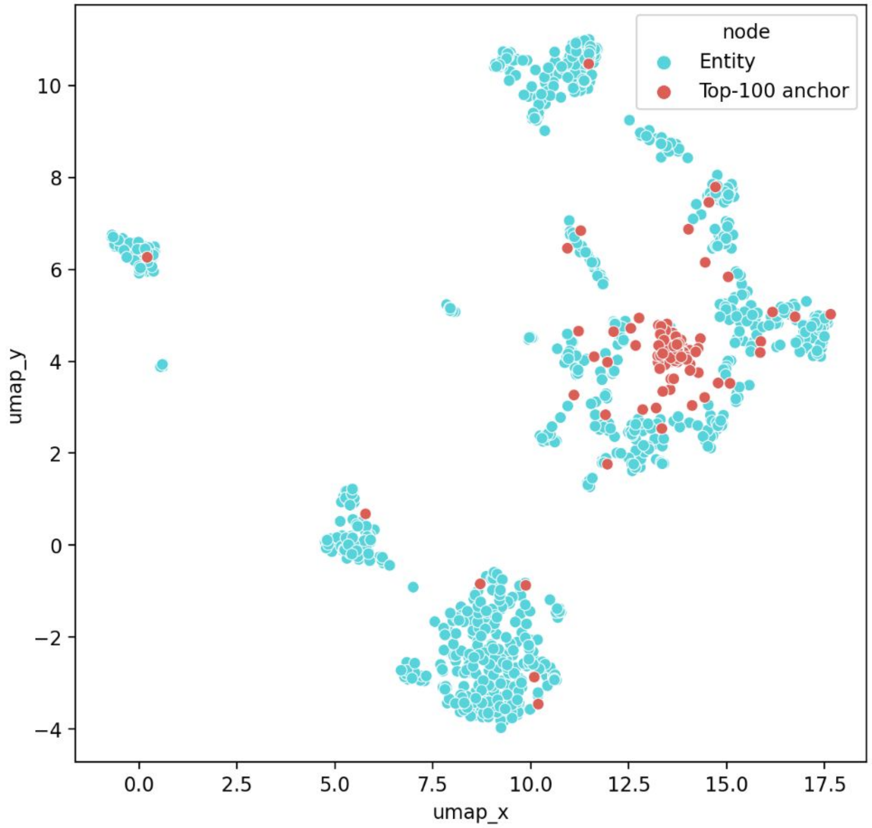
	FB15k-237					WN18RR				
	$\|V\|$	#P (M)	MRR	H@10	%	$\|V\|$	#P (M)	MRR	H@10	%
RotatE	15k + 0.5k	29	0.338†	0.533†	100	40k + 22	41	0.476†	0.571†	100
NodePiece + RotatE	1k + 0.5k	3.2	0.256	0.420	79	500 + 22	4.4	0.403	0.515	90
- no rel. context	1k + 0.5k	2	0.258	0.425	80	500 + 22	4.2	0.266	0.465	81
- no distances	1k + 0.5k	3.2	0.254	0.421	79	500 + 22	4.4	0.391	0.510	89
- no anchors, rels only	0 + 0.5k	1.4	0.204	0.355	67	0 + 22	0.3	0.011	0.019	0.3

WN18RR anchors + entities



FB15k-237 anchors + entities





Transductive Link Prediction

Table 4: Transductive link prediction on bigger KGs. The same denotation as in Table 3. Second RotatE has a similar parameter budget as a NodePiece-based model.

	CoDEx-L					YAGO 3-10				
	$\|V\|$	#P (M)	MRR	H@10	%	$\|V\|$	#P (M)	MRR	H@10	%
RotatE (500d)	77k + 138	77	0.258	0.387	100	123k + 74	123	$0.495^{\dagger}$	$0.670^{\dagger}$	100
RotatE	77k + 138	3.8	0.196	0.322	83	123k + 74	4.8	0.121	0.262	39
NodePiece + RotatE	7k + 138	3.6	0.190	0.313	81	10k + 74	4.1	0.247	0.488	73
- no rel. context	7k + 138	3.1	0.201	0.332	86	10k + 74	3.7	0.249	0.482	72
- no distances	7k + 138	3.6	0.179	0.302	78	10k + 74	4.1	0.250	0.491	73
- no anchors, rels only	0 + 138	0.6	0.063	0.121	31	0 + 74	0.5	0.025	0.041	6

Out-of-sample (inductive) LP

Table 7: Out-of-sample link prediction. † results are taken from [1]. $|V|$ denotes vocabulary size (anchors + relations), #P is a total parameter count (millions).

	oFB15k-237					oYAGO 3-10 (117k)				
	$\|V\|$	#P (M)	MRR	H@10	%	$\|V\|$	#P (M)	MRR	H@10	%
oDistMult-ERAvg	11k + 0.5k	2.4	0.256†	0.420†	100	117k + 74	23.4	OOM	OOM	-
NodePiece + DistMult	1k + 0.5k	1	0.206	0.372	88	10k + 74	2.7	0.133	0.261	100
- no rel. context	1k + 0.5k	1	0.173	0.329	78	10k + 74	2.7	0.125	0.245	94
- no distances	1k + 0.5k	1	0.208	0.372	88	10k + 74	2.7	0.133	0.260	99
- no anchors, rels only	0 + 0.5k	0.8	0.069	0.127	30	0 + 74	0.7	0.015	0.017	6

Encoder: Transformer

Relation Prediction

Table 5: Relation prediction results. $|V|$ denotes vocabulary size (anchors + relations).

	FB15k-237			WN18RR			YAGO 3-10		
	$\|V\|$	MRR	H@10	$\|V\|$	MRR	H@10	$\|V\|$	MRR	H@10
RotatE	15k + 0.5k	0.905	0.979	40k + 22	0.774	0.897	123k + 74	0.909	0.992
NodePiece + RotatE	1k + 0.5k	0.874	0.971	500 + 22	0.761	0.985	10k + 74	0.951	0.997
- no rel. context	1k + 0.5k	0.876	0.968	500 + 22	0.541	0.958	10k + 74	0.898	0.993
- no distances	1k + 0.5k	0.877	0.970	500 + 22	0.746	0.975	10k + 74	0.943	0.997
- no anchors, rels only	0 + 0.5k	0.873	0.971	0 + 22	0.545	0.947	0 + 74	0.951	0.998

Node Classification

Table 6: Node classification results. $|V|$ denotes vocabulary size (anchors + relations), #P is a total parameter count (millions).

	$\|V\|$	#P (M)	WD50K (5% labeled) ROC-AUC	WD50K (5% labeled) PRC-AUC	WD50K (5% labeled) Hard Acc	WD50K (10% labeled) ROC-AUC	WD50K (10% labeled) PRC-AUC	WD50K (10% labeled) Hard Acc
MLP	46k + 1k	4.1	0.503	0.016	0.001	0.510	0.017	0.002
CompGCN	46k + 1k	4.4	0.836	0.280	0.176	0.834	0.265	0.161
NodePiece + GNN	50 + 1k	0.75	0.981	0.443	0.513	0.981	0.450	0.516
- no rel. context	50 + 1k	0.64	0.982	0.446	0.534	0.982	0.449	0.530
- no distances	50 + 1k	0.74	0.981	0.448	0.516	0.981	0.448	0.513
- no anchors, rels only	0 + 1k	0.54	0.984	0.453	0.532	0.984	0.456	0.533

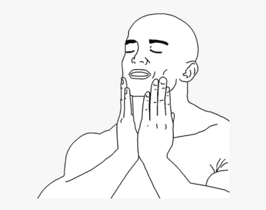
OGB WikiKG 2 : New SOTA

Leaderboard for ogbl-wikikg2

The MRR score on the test and validation sets. The higher, the better.

Package: >=1.2.4

Deprecated `ogbl-wikikg` leaderboard can be found here.



Rank	Method	Test MRR	Validation MRR	Contact	References	#Params	Hardware	Date
1	**NodePiece + AutoSF**	0.5703 ± 0.0035	0.5806 ± 0.0047	Mikhail Galkin (Mila)	Paper, Code	6,860,602	Tesla V100 (32 GB)	Jul 17, 2021
2	**AutoSF**	0.5458 ± 0.0052	0.5510 ± 0.0063	Yongqi Zhang (4Paradigm)	Paper, Code	500,227,800	Quadro RTX 8000 (45GB GPU)	Apr 2, 2021
3	**PairRE (200dim)**	0.5208 ± 0.0027	0.5423 ± 0.0020	Linlin Chao	Paper, Code	500,334,800	Tesla P100 (16GB GPU)	Jan 28, 2021
4	RotatE (250dim)	0.4332 ± 0.0025	0.4353 ± 0.0028	Hongyu Ren – OGB team	Paper, Code	1,250,435,750	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021
5	TransE (500dim)	0.4256 ± 0.0030	0.4272 ± 0.0030	Hongyu Ren – OGB team	Paper, Code	1,250,569,500	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021
6	ComplEx (250dim)	0.4027 ± 0.0027	0.3759 ± 0.0016	Hongyu Ren – OGB team	Paper, Code	1,250,569,500	Quadro RTX 8000 (45GB GPU)	Jan 23, 2021

OGB WikiKG 2 : New SOTA

Input graph: 2.5M nodes, 16M edges

- 20K anchors (< 1% total nodes) -> 4M params
  - 0 anchors / 0 node embeddings -> 0.47 MRR (Top-4)
- 1070 relation types (535 x2 with inverses) -> 200K params
- “Word length” - 32 tokens
  - 20 anchors per node
  - 12 relations in context
  - Tokenization in pre-processing (METIS + igraph) ~ 8 hours
- 2-layer MLP encoder -> ~2M params
- 250K training steps, 1 Tesla V100, 15 hours overall

Total: 6.8M params

Inductive Link Prediction

Inference graphs are disjoint with training (new nodes)
NodePiece + CompGCN encoder = SOTA on many tasks on relation-rich graphs

Table 14: Inductive Link Prediction Results, Hits@10

Class	Method	FB15k-237				WN18RR				NELL-995			
		V1	V2	V3	V4	V1	V2	V3	V4	V1	V2	V3	V4
Path	Neural LP	0.529	0.589	0.529	0.559	0.744	0.689	0.462	0.671	0.408	0.787	0.827	0.806
	DRUM	0.529	0.587	0.529	0.559	0.744	0.689	0.462	0.671	0.194	0.786	0.827	0.806
	RuleN	0.498	0.778	0.877	0.856	0.809	0.782	0.534	0.716	0.535	0.818	0.773	0.614
GNN	GraIL	0.642	0.818	0.828	0.893	0.825	0.787	0.584	0.734	0.595	**0.933**	0.914	0.732
	NBFNet	0.692	0.858	0.898	0.923	**0.942**	**0.895**	**0.900**	**0.881**	-	-	-	-
	NP + CompGCN	**0.873**	**0.939**	**0.944**	**0.949**	0.830	**0.886**	0.785	0.807	**0.890**	0.901	**0.936**	**0.893**

📜 > [Paper PDF]() < 📜

Code & Data https://github.com/migalkin/NodePiece

Contact mikhail.galkin@mila.quebec

Socials @michael_galkin

Mila

ods.ai

- Треки
- Соревнования
- Мероприятия
- Проекты
- Хабы

← KG Course 2021

Новости

KG Course 2021

Курс по графам знаний (Knowledge Graphs)
и как их готовить в 2021 году.

Авторы

Михаил Галкин
Mila Quebec & McGill University

Вадим Сафронов
Key Points

Сергей Иванов
Criteo

https://ods.ai/tracks/kgcourse2021
https://migalkin.github.io/kgcourse2021/

KGC

Program Speakers Partners About FAQs

Home / KGC 2022 Call for Presentations

[KGC 2022 CALL FOR PRESENTATIONS]

The Knowledge Graph Conference 2022

Call for Presentations is now open!

KGs + NLP chair

https://www.knowledgegraph.tech/kgc-2022-call-for-presentations/