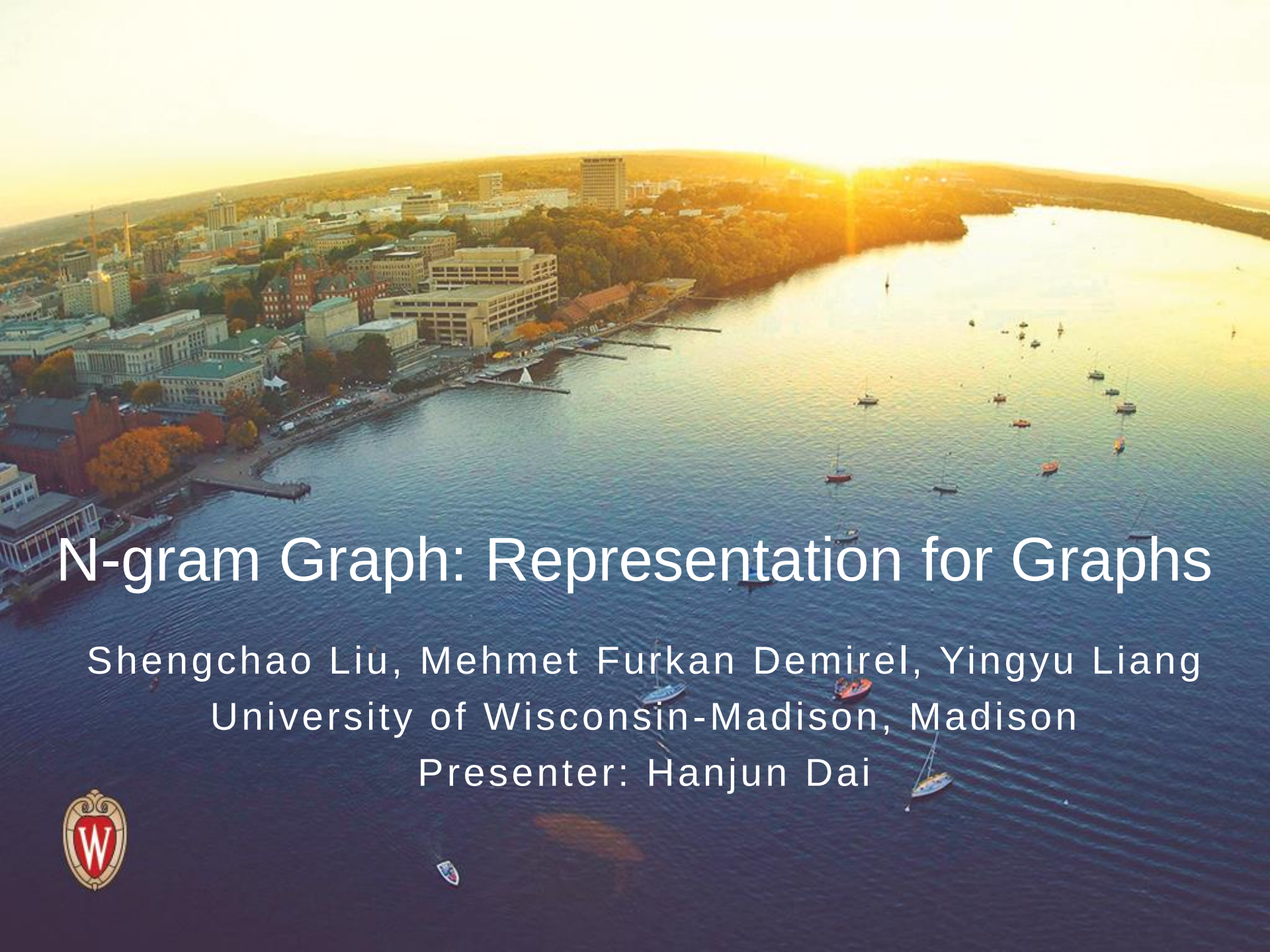
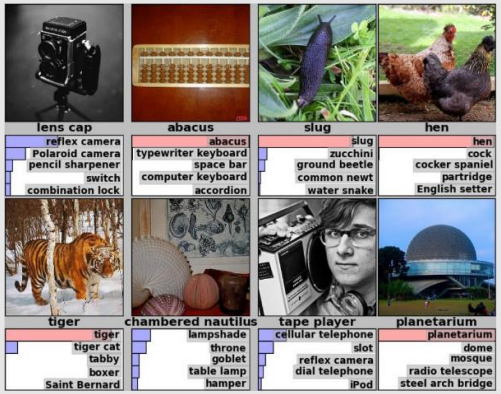


N-gram Graph: Representation for Graphs

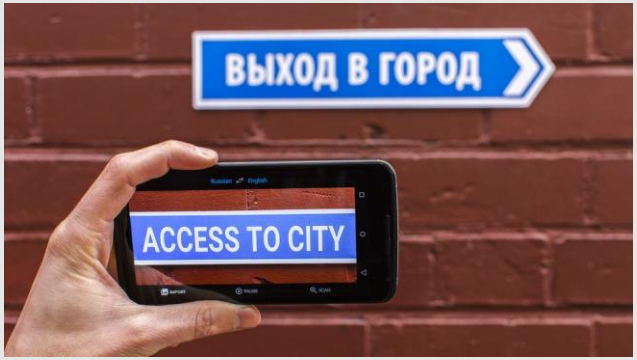
Shengchao Liu, Mehmet Furkan Demirel, Yingyu Liang
University of Wisconsin-Madison, Madison
Presenter: Hanjun Dai

Machine Learning Progress

- Significant progress in Machine Learning

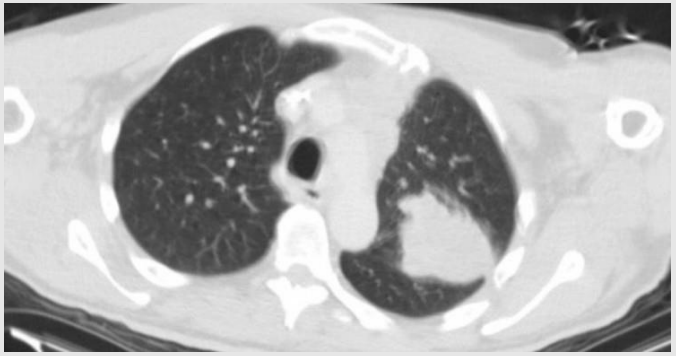


Computer vision



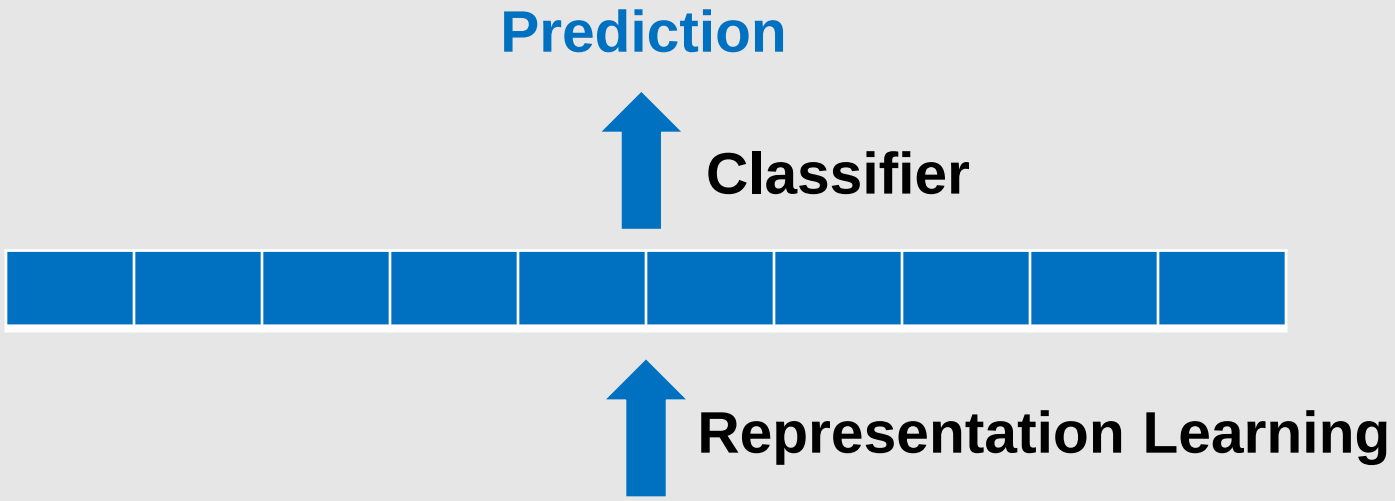
Machine translation

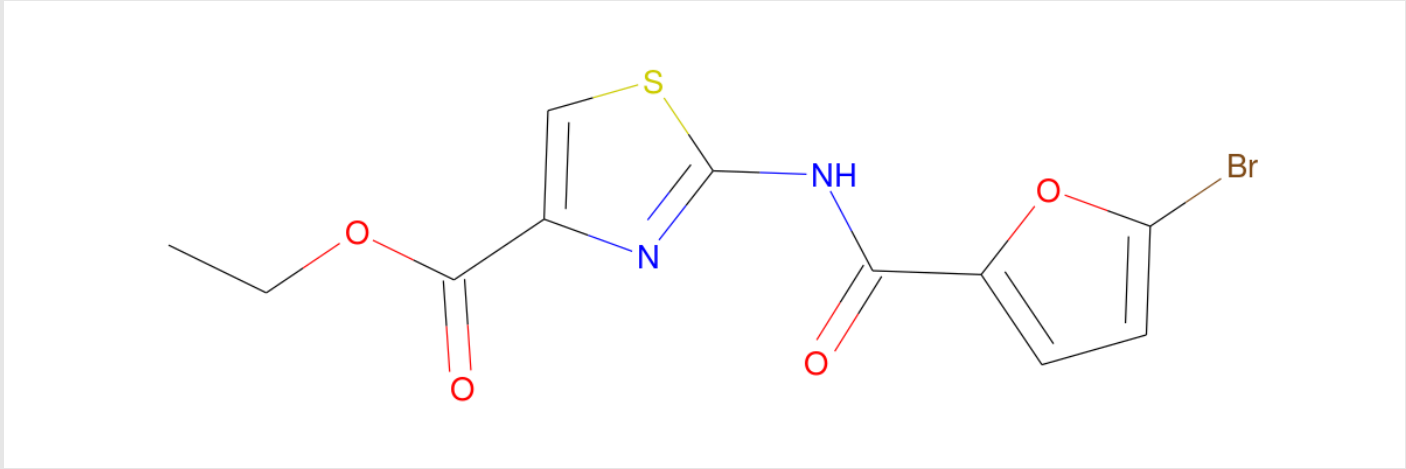
Game Playing



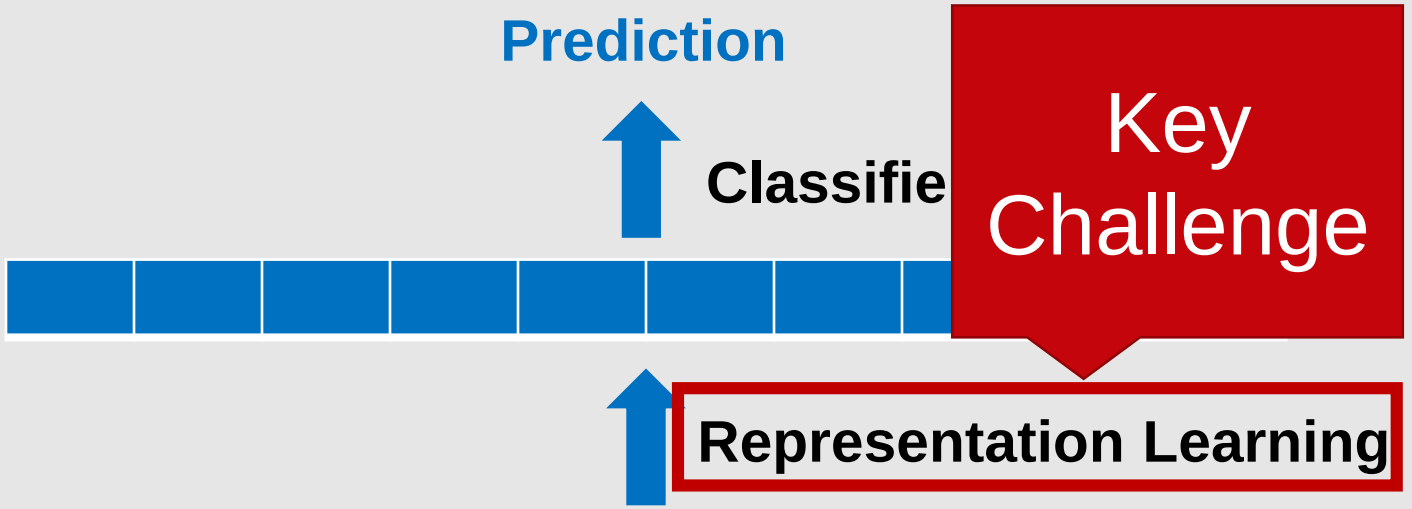
Medical Imaging

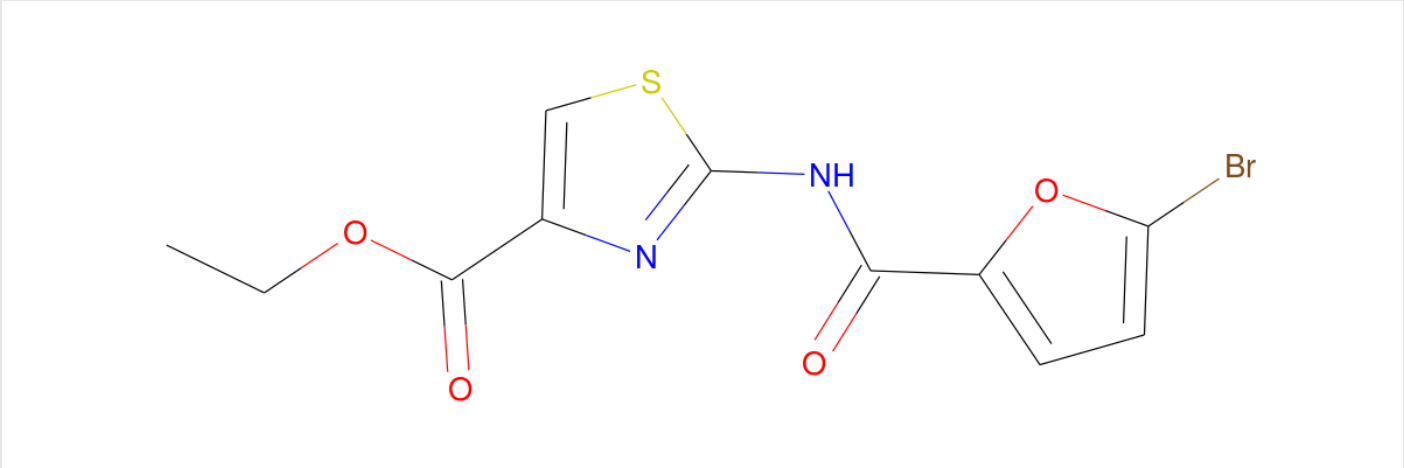
ML for Graph-structured Data like Molecules?





ML for Graph-structured Data like Molecules?





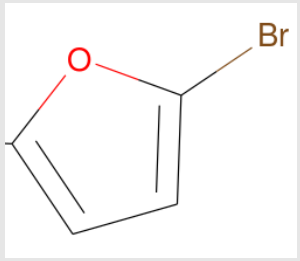
Our Method: N-gram Graphs

- Unsupervised, so can be used by various learning methods
- Simple, relatively fast to compute

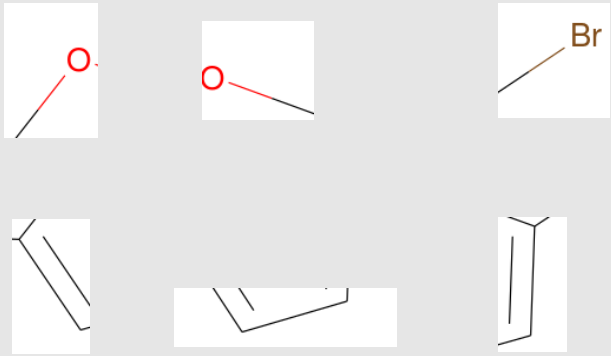
- Strong empirical performance
  - Outperforms traditional fingerprint/kernel and recent popular GNNs on molecule datasets
  - Preliminary results on other types of data are also strong
- Strong theoretical power for representation/prediction

N-gram Graphs: Bag of Walks

- Key idea: view a graph as Bag of Walks
  - Walks of length n are called n-grams



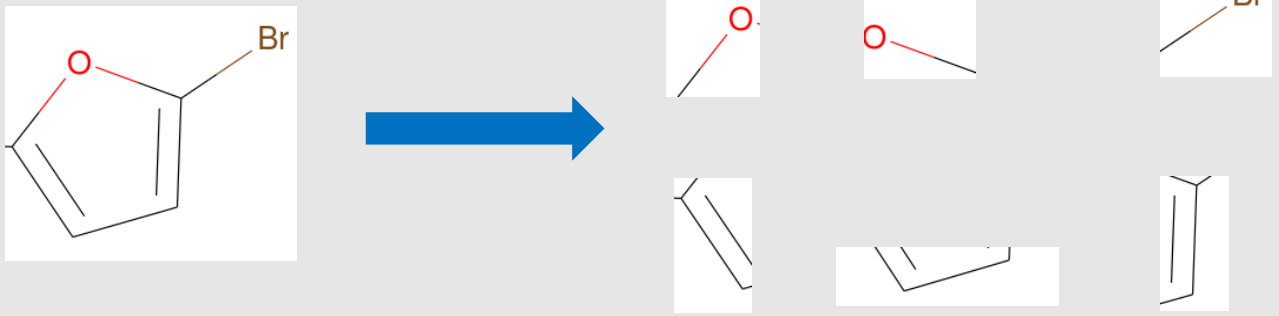
A molecular graph



Its 2-grams

N-gram Graphs: Bag of Walks

- Key idea: view a graph as Bag of Walks
  - Walks of length n are called n-grams



A molecular graph

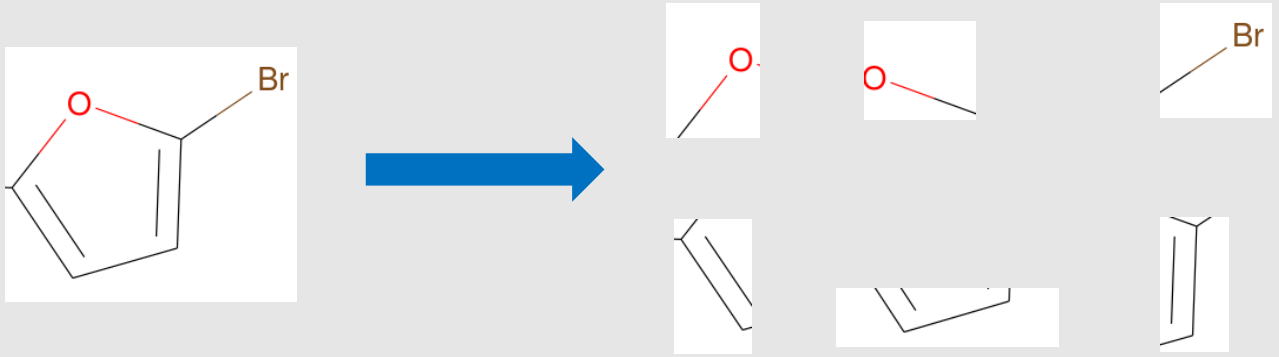
Its 2-grams

N-gram Graph (suppose the embeddings for vertices are given):

1. Embed each n-gram: entrywise product of its vertex embeddings
2. Sum up the embeddings of all n-grams: denote the sum as $f_{(n)}$
3. Repeat for $n = 1, 2, \ldots, T$, and concatenate $f_{(1)}, \ldots, f_{(T)}$

N-gram Graphs: Bag of Walks

- Key idea: view a graph as Bag of Walks
  - Walks of length n are called n-grams



A molecular graph

Its 2-grams

Equivalent to a simple Graph Neural Network!

N-gram G… …are given):

1. Embe… …ex embeddings
2. Sum up the … …sum as $f_{(n)}$
3. Repeat for $n = 1, 2, \ldots, T$, and concatenate $f_{(1)}, \ldots, f_{(T)}$

Experimental Results

- 60 tasks on 10 datasets (predict molecular properties)
- Compared to classic fingerprint/kernel and recent GNNs

Experimental Results

- 60 tasks on 10 datasets (predict molecular properties)
- Compared to classic fingerprint/kernel and recent GNNs

Dataset	# Task	Eval Metric	WL SVM	Morgan RF	Morgan XGB	GCNN	Weave	GIN	N-Gram RF	N-Gram XGB
Delaney	1	RMSE					1, 1	–	0, 1	0, 1
Malaria	1	RMSE		1, 1				–	0, 1	0, 1
CEP	1	RMSE		1, 1				–	0, 1	0, 1
QM7	1	MAE					0, 1	–	0, 1	1, 1
QM8	12	MAE		1, 4	0, 1	7, 12	2, 6	–	0, 2	2, 11
QM9	12	MAE	–		0, 1	4, 7	1, 8	–	0, 8	7, 12
Tox21	12	ROC-AUC	0, 2	0, 7		0, 2	0, 1		3, 12	9, 12
clintox	2	ROC-AUC	0, 1			1, 2	0, 1			1, 2
MUV	17	PR-AUC	4, 12	5, 11	5, 11			0, 7	2, 4	1, 6
HIV	1	ROC-AUC		1, 1					0, 1	0, 1
Overall	60		4, 15	9, 25	5, 13	12, 23	4, 18	0, 7	5, 31	**21, 48**

- N-gram+XGBoost: top-1 for 21 tasks, and top-3 for 48 tasks
- Overall better than the other methods

Theoretical Analysis

- N-gram graph ~= compressive sensing of the count statistics (i.e., histogram of different types of n-grams)
- Thus has strong representation and prediction power

Come to Poster # 70 for details!

- Code published: https://github.com/chao1224/n_gram_graph