Predicting Citation Count

Surya Teja, Akshita Mittel and
Sree Ram Sreekar

Problem Statement

- Given: A set of features $F = \{f_1, f_2, \ldots, f_n\}$
- Goal: Predict citation count of paper 'd' after time Δt of it's publication
- Formally, learn $\psi(d, F, \Delta t) \rightarrow C\ (d, \Delta t)$

- Seek to compare multiple approaches to solve this problem

The Dataset

- The citation data is extracted from DBLP, ACM, and other sources.
- Contains 629,814 papers and 632,752 citations.
- Each paper is associated with abstract, authors, year, venue, title and references.
- Found at https://cn.aminer.org/citation

Approach 1

Stratified Learning

The first approach – Stratified Learning

- **Two-stage prediction model**
  - In the first stage, the model maps a query paper into one of the six categories
  - In the second stage a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper.
- **Features used to train the regression module are broadly classified into three main categories:**
  - Author centric features
  - Venue centric features
  - Paper centric features

Results

	Our imp	Paper	Our imp	Paper
T: Num. of years after publication	R^2	R^2	MSE	MSE
T = 1	0.62	0.57	2.84	-(Not provided)
T = 2	0.51	0.55	3.11	-(Not provided)
T = 3	0.44	0.52	3.85	-(Not provided)
T = 4	0.38	0.50	4.32	-(Not provided)
T = 5	0.31	0.45	5.87	-(Not provided)

Approach 2:

Yan et al.
Citation Count Prediction:
Learning to Estimate Future Citations for Literature

Citation Count Prediction: Learning to Estimate Future Citations for Literature

1. Topic Rank
2. Diversity
3. Recency
4. H-Index
5. Author Rank
6. Productivity
7. Sociality
8. Venue Rank

Learning models:

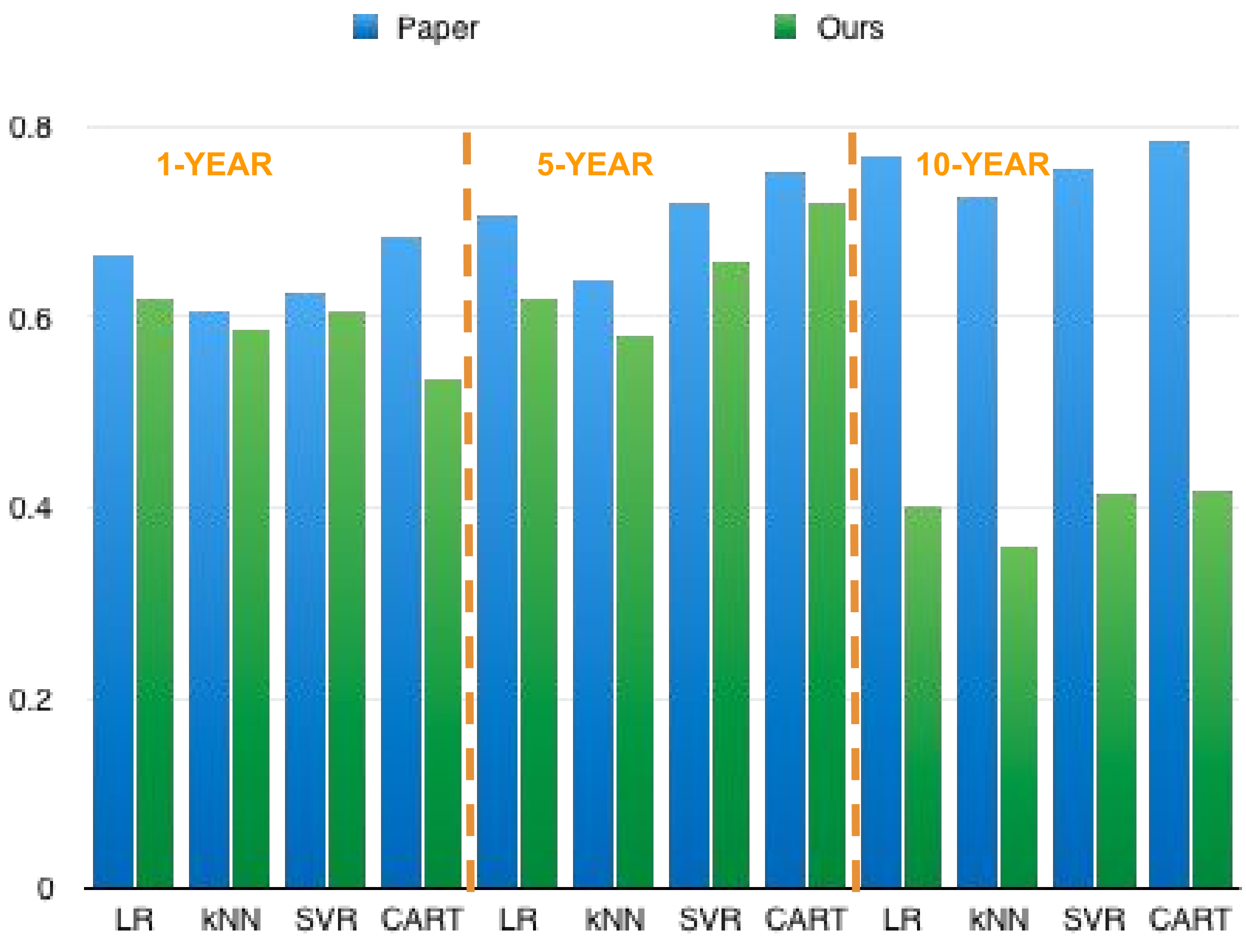
Linear regression(LR), k-NN, SVR, CART models were all used and compared against each other.

The SVR model gave the best performance.

Results

- The metric followed in the paper and our presentation is R-squared.
- $$R^2 = \frac{\sum_{d \in D_T} (C_{T\,ccp}(d) - C_T(D_T))^2}{\sum_{d \in D_T} (C_T(d) - C_T(D_T))^2}$$

	Method	Value reported in paper	Value reported by our implementation	MSE
1-year CCP	LR	0.664	0.620	0.8197
	kNN	0.607	0.587	0.8976
	SVR	0.625	0.606	0.8597
	CART	0.683	0.534	1.8082
5-year CCP	LR	0.706	0.618	6.9491
	kNN	0.640	0.580	7.6715
	SVR	0.719	0.657	7.3733
	CART	0.752	0.718	11.8492
10-year CCP	LR	0.767	0.403	34.6031
	kNN	0.725	0.360	36.8839
	SVR	0.755	0.416	36.2493
	CART	0.786	0.418	45.2084



Paper
Ours
0.8
0.6
0.4
0.2
0
1-YEAR
5-YEAR
10-YEAR
LR
kNN
SVR
CART
LR
kNN
SVR
CART
LR
kNN
SVR
CART

OUR APPROACH:

Citation Count Prediction: Sequence to Sequence Learning

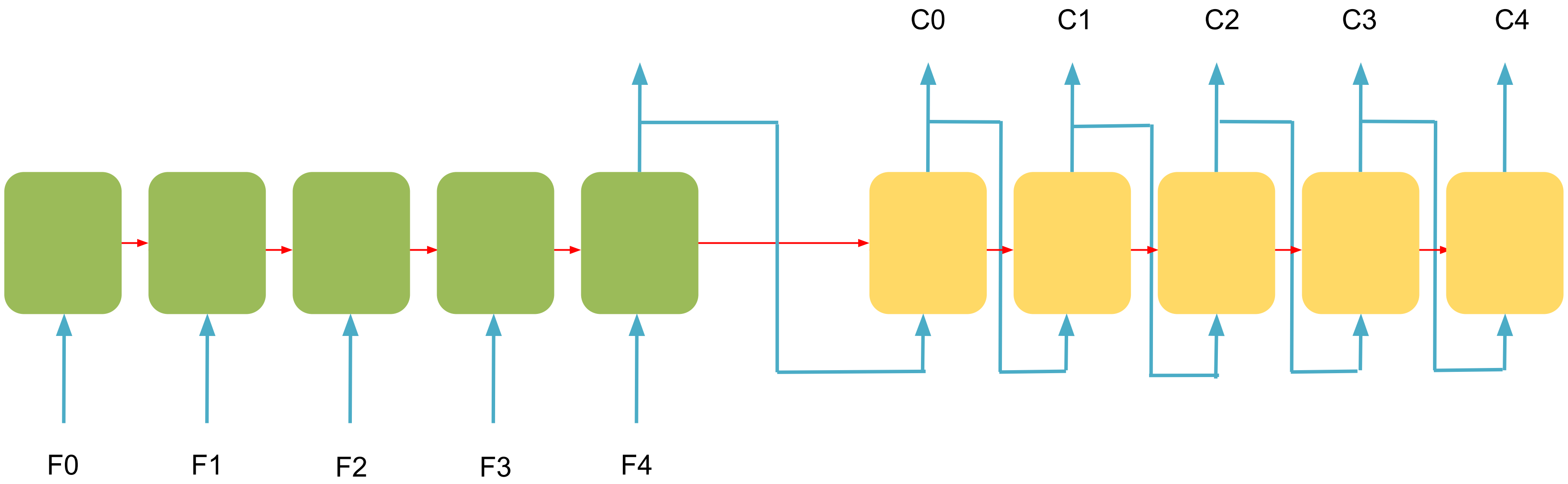
Approach

- Re-fashion the problem statement as a sequence-to-sequence mapping problem

- [F0, F1, F2, F3, F4] -> [C0, C1, C2, C3, C4]

- Apply LSTM encoder-decoder approach to solve the problem

Architecture



Results

V1-Aminer dataset

SET	Samples(papers)	MSE(our approach)	Yan et al MSE	KGP MSE
Train data	225,746	0.018	5.84	3.76
Validation data	111,189	0.014	6.21	4.44
Test data	165,952	0.021	6.94(best)	5.87

Future work

- Extend to larger data-set
- Qualitative discussion of results and interpretation
- Document/comment the code

Thank you!