SAVeD:
Semantic Aware Version Detection

Artem Frenk (Data Science)
Roee Shraga (Computer Science, Data Science)

Motivation

Given tables, how can we determine if they are *versions* of each other?

id	name	score
1	Alice	83
2	Bob	91

id	full_name	score
1	Alice T.	84
2	Bob	91

id	name	has_cat?
1	Alice	True
2	Bob	False

What is a Version?

Let table T be composed of a set of attributes $T_A = \{A_1, \dots A_n\}$ and tuples $T_r = \{r_1 \dots r_m\}$.

Let each tuple defined by $r_i = \langle ri_0, ri_1 \dots ri_n \rangle$, s.t. r_{i0} can be easily recognized as the tuple identifier and $r_{ij} (j \neq 0)$ can represent a value assigned to the attribute A_j in the tuple r_i.

With that definition, let T and T' be two tables. We say that T' is a *version* of T iff there exists a transformation p such that $p(T) = T'$, where p belongs to a family of semantics-preserving transformations $\mathcal{P}$.

The version relationship is then defined as:

$$\text{Version}(T, T') \iff \exists p \in \mathcal{P}^* : p(T) = T'$$

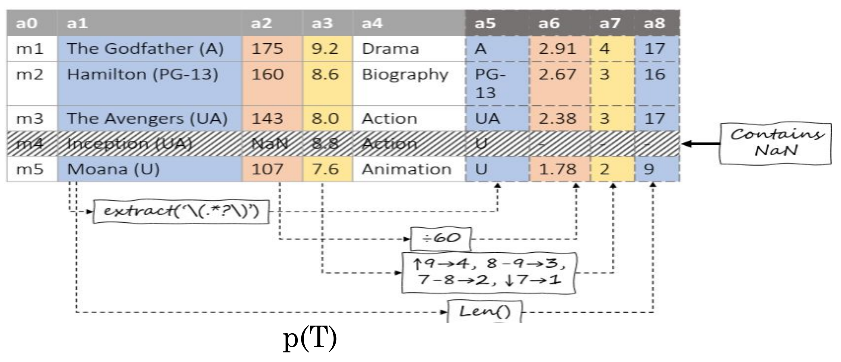
where $\mathcal{P}^*$ denotes the closure of $\mathcal{P}$ under composition.

a0	a1	a2	a3	a4
m1	The Godfather (A)	175	9.2	Drama
m2	Hamilton (PG-13)	160	8.6	Drama
m3	The Avengers (UA)	143	8.0	Action
m4	Inception (UA)	NaN	8.8	Action
m5	Moana (U)	107	7.6	Animation

T

a0	a1	a2	a3	a4	a5	a6	a7	a8
m1	The Godfather (A)	175	9.2	Drama	A	2.91	4	17
m2	Hamilton (PG-13)	160	8.6	Drama	PG-13	2.67	3	16
m3	The Avengers (UA)	143	8.0	Action	UA	2.38	3	17
m5	Moana (U)	107	7.6	Animation	U	1.78	2	9

T′



a0	a1	a2	a3	a4	a5	a6	a7	a8
m1	The Godfather (A)	175	9.2	Drama	A	2.91	4	17
m2	Hamilton (PG-13)	160	8.6	Biography	PG-13	2.67	3	16
m3	The Avengers (UA)	143	8.0	Action	UA	2.38	3	17
~~m4~~	~~Inception (UA)~~	~~NaN~~	~~8.8~~	~~Action~~	~~U~~	~~-~~	~~-~~	~~-~~
m5	Moana (U)	107	7.6	Animation	U	1.78	2	9

p(T)

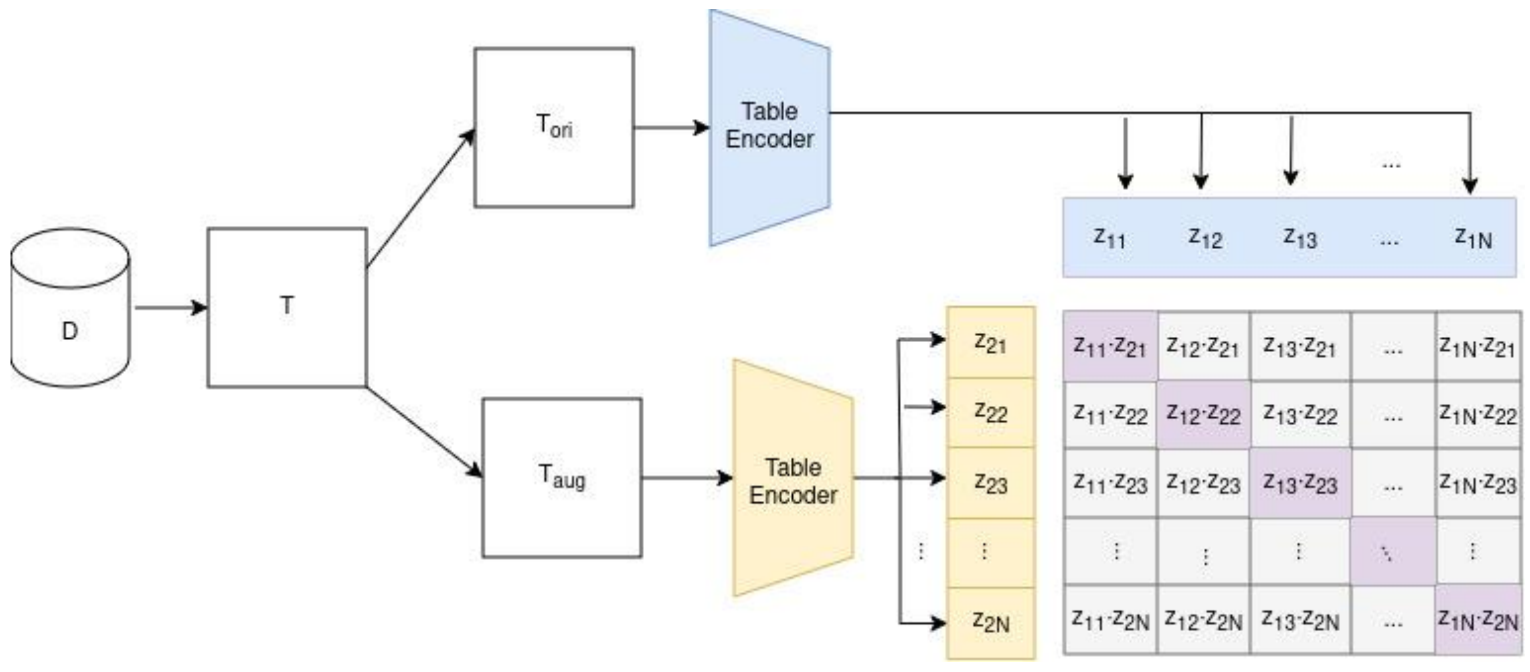
Related Work

- ➢ **Data Versioning**
  - ○ Storage, Scalability, Management, Version Control Systems (Git, Xet)
- ➢ **Data Discovery**
  - ○ Table Union Search, PBE, Related Table Search, Joinable Table Search
- ➢ **Table Representation Learning**
  - ○ Contrastive Pretraining, Permutation-Based Tabular Models

Preprint & Codebase

Contrastive Learning (SimCLR)

Augmentation at Train Time



$z_{11} \cdot z_{21}$	$z_{12} \cdot z_{21}$	$z_{13} \cdot z_{21}$	...	$z_{1N} \cdot z_{21}$
$z_{11} \cdot z_{22}$	$z_{12} \cdot z_{22}$	$z_{13} \cdot z_{22}$	...	$z_{1N} \cdot z_{22}$
$z_{11} \cdot z_{23}$	$z_{12} \cdot z_{23}$	$z_{13} \cdot z_{23}$	...	$z_{1N} \cdot z_{23}$
⋮	⋮	⋮	⋱	⋮
$z_{11} \cdot z_{2N}$	$z_{12} \cdot z_{2N}$	$z_{13} \cdot z_{2N}$	...	$z_{1N} \cdot z_{2N}$

Augmentation and Loss

- ➢ **Augmentation during training**
  - ○ All augmentations independently probable.
  - ○ Includes augmentations such as gaussian jitter, column & row drop and/or shuffle, NaN injection, and one-hot/dummy encoding
- ➢ **Normalized Temperature Scaled Cross-Entropy Loss**
  - ○ L2 Normalization so table embeddings lie on the unit hypersphere.
  - ○ For a batch of N table embeddings, positive pairs are pulled together in embedding space.
  - ○ Negative pairs are pushed farther away

Experimental Design

- ➢ **Benchmark**:
  - ○ 5 Datasets (IMDB, Titanic, Wine, NBA, Iris) collected into 1 large benchmark called SDVB (Semantic Data Versioning Benchmark)
- ➢ **Evaluation**:
  - ○ Evaluated on Separation (difference between version-version similarity and version to non-version similarity) and TPR.
  - ○ Evaluated against labels provided by benchmark authors to determine whether a table is a version.
  - ○ Compared to several pretrained small language models, as well as Starmie, a state of the art contrastive learning approach to dataset discovery.