Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples

Peng Li*
Georgia Institute of Technology
pengli@gatech.edu

Yeye He, Cong Yan, Yue Wang, Surajit Chaudhuri
Microsoft Research
{yeyehe,coyan,wanyue,surajitc}@microsoft.com

ABSTRACT

Relational tables, where each row corresponds to an entity and each column corresponds to an attribute, have been the standard for tables in relational databases. However, such a standard cannot be taken for granted when dealing with tables "in the wild". Our survey of real spreadsheet-tables and web-tables shows that over 30% of such tables do not conform to the relational standard, for which complex table-restructuring transformations are needed before these tables can be queried easily using SQL-based tools. Unfortunately, the required transformations are non-trivial to program, which has become a substantial pain point for technical and non-technical users alike, as evidenced by large numbers of forum questions in places like StackOverflow and Excel/Tableau forums.

We develop an Auto-Tables system that can automatically synthesize pipelines with multi-step transformations (in Python or other languages), to transform non-relational tables into standard relational forms for downstream analytics, obviating the need for users to manually program transformations. We compile an extensive benchmark for this new task, by collecting 244 real test cases from user spreadsheets and online forums. Our evaluation suggests that Auto-Tables can successfully synthesize transformations for over 70% of test cases at interactive speeds, without requiring any input from users, making this an effective tool for both technical and non-technical users to prepare data for analytics.

PVLDB Reference Format:
Peng Li and Yeye He, Cong Yan, Yue Wang, Surajit Chaudhuri. Auto-Tables: Synthesizing Multi-Step Transformations to Relationalize Tables without Using Examples. PVLDB, 16(11): 3391 - 3403, 2023.
doi:10.14778/3611479.3611534

1 INTRODUCTION

Modern data analytics like SQL and BI are predicated on a standard format of relational tables, where each row corresponds to a distinct "entity", and each column corresponds to an "attribute" for the entities that contains homogeneous data-values. While such tables are de-facto standards in relational databases, such that as database people we may take this for granted, a significant fraction of tables "in the wild" actually fail to conform to such standards, making them considerably more difficult to query using SQL-based tools.

Non-relational tables are common, but hard to query. Real tables in the wild, such as spreadsheet-tables or web-tables, can often be "non-relational" and hard to query, unlike tables that we expect to find in relational databases. We randomly sampled hundreds of user spreadsheets (in Excel), and web tables (from Wikipedia), and found around 30-50% tables to have such issues. Figure 1 and Figure 2 show real samples taken from spreadsheets and the web, respectively, to demonstrate these common issues. (We emphasize that the problem is prevalent at a very large scale, since there are millions of tables like these in spreadsheets and on the web.)

Take Figure 1(a) for example. The table on the left is not a standard relational table, because each column marked in green contains sales numbers for only a single day (`19-Oct`, `20-Oct`, etc.), making these column values highly homogeneous in the horizontal direction (while in typical relational tables, we expect values in columns to be homogeneous in the vertical direction). Although this specific table format makes it easy for humans to eyeball changes day-over-day by reading horizontally, it is unfortunately hard to analyze using SQL. Imagine that one needs to compute the 14-day average of sales, starting from "`20-Oct`" – for this table, one has to write: `SELECT SUM("20-Oct", "21-Oct", "22-Oct", ...) FROM T`, across 14 different columns, which is long and unwieldy to write. Now imagine we need 14-day moving averages with every day in October as the starting date – the resulting SQL is highly repetitive and hard to manage.

In contrast, consider a transformed version of this table, shown on the right of Figure 1(a). Here the homogeneous columns in the original table (marked in green) are transformed into only two new columns: "`Date`" and "`Units Sold`", using a transformation operator called "`stack`" (listed in the first row of Table 1). This transformed table contains the same information as the original table, but is much easier to query – e.g., the same 14-day moving average can be computed using a succinct range-predicate on the '`Date`" column, where the starting date "`20-Oct`" is a literal parameter that can be easily changed into other values.

There are many such spreadsheet tables that require different kinds of transformations before they are ready for SQL-based analysis. Figure 1(b) shows another example where every 3 columns form a group, representing "`Revenue/Units Sold/Margin`" for a different year, repeating for many times (marked in red/green/blue in the figure). Tables with these repeating column-groups are also hard to query just like Figure 1(a), but in this case the required transformation operator is different and called "`wide-to-long`" (listed in the second row of Table 1).

Figure 1(c) shows yet another example, where each hotel corresponds to a column (whose names are in row-1), and each "attribute" of these hotels corresponds to a row. Note that in this case values in the same rows are homogeneous (marked in different colors),

---

*Part of work done while at Microsoft.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 16, No. 11 ISSN 2150-8097.
doi:10.14778/3611479.3611534

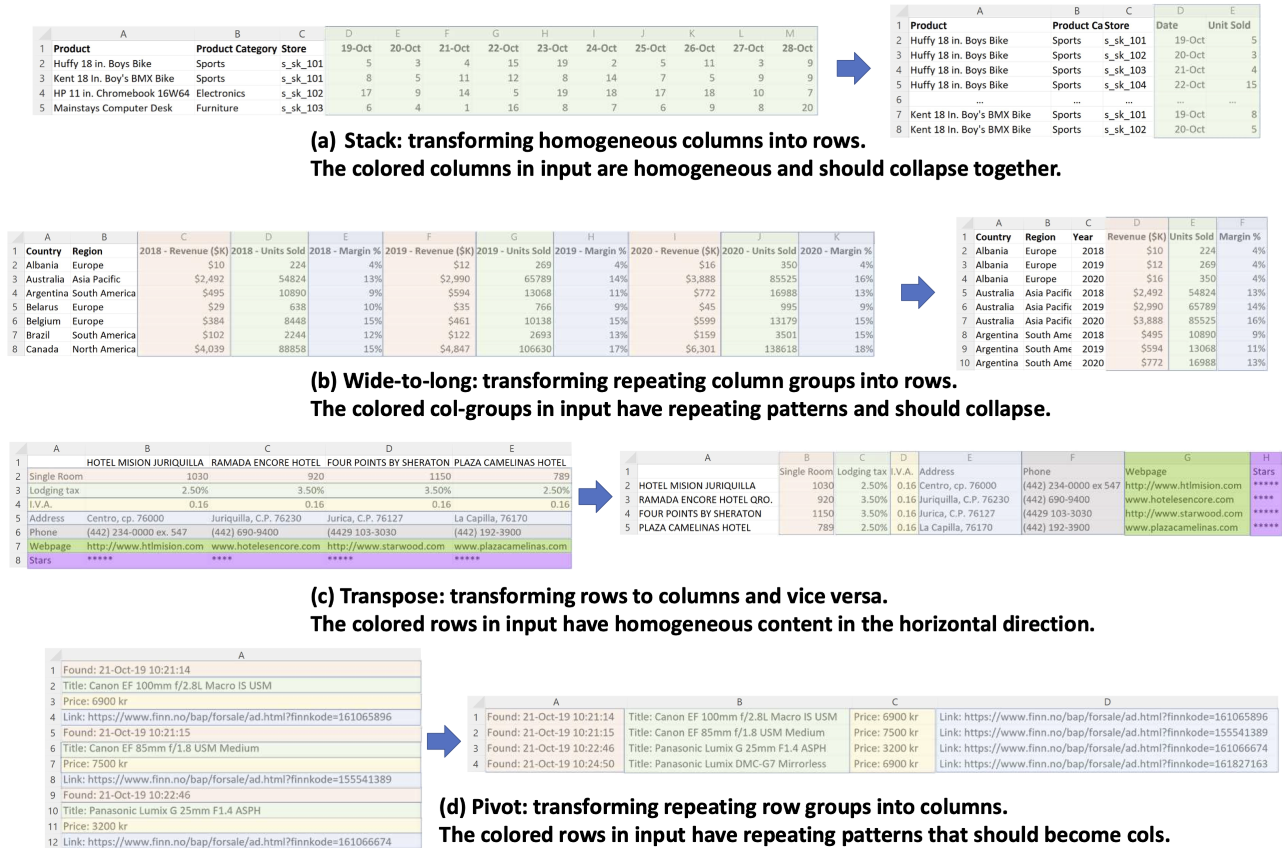


Figure 1: Example input/output tables for 4 operators in AUTO-TABLES: (a) Stack, (b) Wide-to-long, (c) Transpose, (d) Pivot. The input-tables (on the left) are not relational and hard to query, which need to be transformed to produce corresponding output-tables (on the right) that are relational and easy to query. Observe that the color-coded, repeating row/column-groups are "visual" in nature, motivating a CNN-like architecture like used in computer vision for object-detection.

Figure 2: Real Web tables from Wikipedia that are also non-relational, similar to the spreadsheet tables shown in Figure 1.

unlike relational tables where values in the same columns are homogeneous. A transformation called "transpose" is required in this case (listed in the third row of Table 1), to make the resulting table, shown on the right of the figure, easy to query – for instance, a query to sum up the total number of hotel rooms is hard to write on the original table, but can be easily achieved using a simple SUM query on the "Single Room" column in the transformed table.

Figure 1(d) shows another example where columns are represented as rows in the table on the left. This is similar to Figure 1(c), except that the rows in this case are "repeating" in groups, thus requiring a different transformation operator called "pivot" (listed in the fourth row of Table 1) as opposed to "transpose". The resulting table is shown on the right, which becomes easy to query.

While the examples so far are all taken from spreadsheets, we note that similar structural issues are also widespread in Web tables. Figure 2 shows real examples from Wikipedia, which share similar characteristics as the spreadsheet tables in Figure 1, which all require transformations before these tables can be queried effectively.

Non-relational tables are hard to "relationalize". We mentioned that the example tables in Figure 1 and Figure 2 require different transformation operators. Table 1 shows 8 such transformation operators commonly needed to relationalize tables (where the first 4 operators correspond to the examples we see in Figure 1).

The first column of Table 1 shows the name of the "logical operator", which may be instantiated differently in different languages

Table 1: Auto-Tables DSL: table-restructuring operators and their parameters to "relationalize" tables. These operators are common and exist in many different languages, like Python Pandas and R, sometimes under different names.

DSL operator	Python Pandas equivalent	Operator parameters	Description (example in parenthesis)
stack	melt [18]	start_idx, end_idx	collapse homogeneous cols into rows (Fig. 1a)
wide-to-long	wide_to_long [22]	start_idx, end_idx, delim	collapse repeating col-groups into rows (Fig. 1b)
transpose	transpose [21]	-	convert rows to columns and vice versa (Fig. 1c)
pivot	pivot [19]	repeat_frequency	pivot repeating row-groups into cols (Fig. 1d)
explode	explode [16]	column_idx, delim	convert composite cells into atomic values
ffill	ffill [17]	start_idx, end_idx	fill structurally empty cells in tables
subtitles	copy, ffill, del	column_idx, row_filter	convert table subtitles into a column
none	-	-	no-op, the input table is already relational

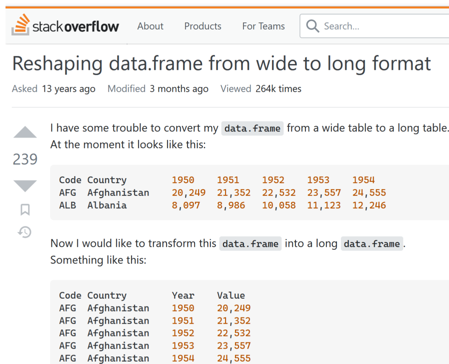


Figure 3: Example user question from StackOverflow, on how to restructure tables. Questions like this are common not only among technical users, but also non-technical users, as similar questions are commonly found on forums for Excel, Power-BI, and Tableau users too [6–9].

(e.g., in Python or R), with different names and syntax. The second column of the table shows the equivalent Pandas operator in Python [15], which is a popular API for manipulating tables among developers and data scientists, that readers may be familiar with.

While the functionalities listed in Table 1 already exist in languages such as R and Python, they are not easy for users to invoke correctly, because users need to:

- (1) Visually identify different structural issues in an input table that make it hard to query (e.g., repeating row-/column-groups shown in Figure 1(a-d)), which is not obvious to non-expert users;
- (2) Map the visual pattern identified from the input table, to a corresponding operator in Table 1 that can handle such issues. This is hard as users are often unfamiliar with the exact terminologies to describe these transformation operators (e.g., `pivot` vs. `stack`), often needing to search online for help;
- (3) Parameterize the chosen operator appropriately, using parameters tailored to the input table (e.g., which columns need to collapse into rows, what is the repeating frequency of column groups, etc.). This is again hard, as even developers need to consult the API documentation, which is often long and complex.
- (4) Certain input tables require more than one transformation step, for which users need to repeat steps (1)-(3) multiple times.

Completing these steps is a tall order even for technical users, as evidenced by a large number of related questions on forums like StackOverflow (e.g., [10–13]). Figure 3 shows such an example question (popular with many up-votes), where the developer provides example input/output tables to demonstrate the desired transformation, and seek help on what Pandas operators to invoke.

If technical users like developers find it hard to restructure their tables, as these StackOverflow questions would show, it comes as no surprise that non-technical enterprise users, who often deal with tables in spreadsheets, would find the task even more challenging. We find a large number of similar questions on Excel and Tableau forums (e.g., [6–9]), where users complain that without the required transformations it is hard to analyze data using SQL-based or Excel-based tools (e.g., [2–5]).

The prevalence of these questions confirms table-restructuring as a common pain point for both technical and non-technical users.

Auto-Tables: synthesize transformations without examples. In this work, we propose a new paradigm to automatically synthesize table-restructuring steps to relationalize tables, using the Domain Specific Language (DSL) of operators in Table 1, *without requiring users to provide examples.* Our key intuition of why we can do away with examples in our task, lies in the observation that given an input table, the logical steps required to relationalize it are *almost always unique* and with little ambiguity, as the examples in Figure 1 would all show. This is because the transformations required in our task only "restructure" tables, that do not actually "alter" the table content, which is unlike prior work that focuses on *row-to-row transformations* (e.g., TDE [36] and FlashFill [34]), or SQL-by-example (e.g. [26, 58]), where the output is "altered" that can produce many possible outcomes, which would require users to provide input/output examples to demonstrate the desired outcome.

For our task, we believe it is actually important *not* to ask users to provide examples, because in the context of table-to-table transformations like in our case, asking users to provide examples would mean users have to specify *an output table*, which is a substantial amount of typing effort, making it cumbersome to use.

As humans, we can "visually" recognize rows/columns patterns (e.g., homogeneous value groups, as color-coded vertically and horizontally in Figure 1), to correctly predict which operator to use. The question we ask in this paper, is whether an algorithm can "learn" to recognize such patterns by scanning the input tables alone, to predict suitable transformations, in a manner that is analogous to how computer-vision algorithms would scan a picture to identify common but more complex objects like dogs and cats.

We should note that like computer vision problems such as object detection, where hand-crafted heuristics are hard to write, the row/column-level patterns existing in our target tables are also data-dependent and subtle, which are hard to write as heuristic rules. Consider for example the table in Figure 1(b) – for ease of illustration we pick a case with three distinct groups of columns (currency, integers, and percentage-numbers, marked in different colors). One may hand-craft a heuristic "similarity function" between columns that may work for this simple example, but imagine the common scenario where all these columns have similar-looking integer numbers (e.g., with no dollar signs and percentage signs), which is much more challenging to predict using heuristics, as fine-grained differentiation is required to tell the subtle differences between columns (e.g., difference in column header semantics or column value ranges), which is best learned from the data. In fact, we tested

a baseline using heuristic rules to predict only the simple "stack", which has a low 0.38 accuracy, because of the subtle differences in data that are not captured by heuristics. We also tested an LLM-based approach using GPT-3.5, also without success (with more details in our experiments), further underlining the challenging nature of our task. These motivate us to develop a learning-based method specifically tailored to our table transformation task.

In computer vision, in order to pick up subtle clues from pictures, object detection algorithms are typically trained using large amounts of labeled data [31] (e.g., pictures of dogs that are manually labeled as such). In our task, we do not have such labeled datasets. Therefore, we devise a novel *self-training framework* that exploits the *inverse functional relationships* between operators (e.g., the inverse of "`stack`" is known as "`unstack`"), to automatically build large amounts of training data without requiring humans to label, as we will explain in Figure 6. Briefly, in order to build a training example for operator O (e.g., "`stack`"), we start from a relational table R and apply the inverse of O, denoted by O^{-1} (e.g., "`unstack`"), to generate a table $T = O^{-1}(R)$, which we know is non-relational. For our task, given T as input, we know O must be its ground-truth transformation, because by definition $O(T) = O(O^{-1}(R)) = R$, which turns T back to its relational form R. This makes (T, O) an (example, label) pair that we can automatically generate at scale, and use as our training data.

Leveraging training data so generated, we develop an Auto-Tables system that can "learn-to-synthesize" table-restructuring transformations, using a deep tabular model we develop inspired by CNN-like architectures popular in the computer vision literature. We show our approach is effective on real-world tasks, which can solve over 70% of test cases collected from user forums and spreadsheets, while being interactive with sub-second latency.

Contributions. We make these contributions in Auto-Tables:

- We propose a novel problem to automatically relationalize tables without examples, which addresses a common pain point for both technical and non-technical users, when they deal with tables in the wild outside of database settings.
- We develop Auto-Tables that learns-to-synthesize transformations, using a computer-vision inspired model architecture that exploits the common "visual" patterns in tables.
- We propose a self-supervision framework unique in our setting to overcome the lack of training data, by exploiting the inverse functional relationships between operators to auto-generate training data, obviating the expensive process of human labeling.
- We compile an extensive benchmark for this task by collecting 244 real test cases from user spreadsheets and online forums.[1] Our evaluation suggests that Auto-Tables can successfully synthesize transformations for over 70% of test cases at interactive speeds (with sub-second latency).

2 RELATED WORK

By-example transformation using program synthesis. There is a large body of prior work on using input/output examples to synthesize transformations. One class of techniques focuses on the so-called "row-to-row" transformations where one input row maps to one output row (e.g., TDE [36] and FlashFill [34]), which are orthogonal to the table-restructuring transformations in Auto-Tables, because these systems do not consider operators shown in Table 1 that can change the structure of tables. Other forms of row-to-row transformations using partial specifications (e.g., transform-by-pattern [27, 56], transform-by-target [39, 41], and transform-for-joins [46, 59]), are similarly also orthogonal to the problem we study in this work.

A second class of by-example transformation consider "table-to-table" operators, such as Foofah [38] and SQL-by-example techniques like PATSQL [52], QBO [53], and Scythe [54]. These techniques consider a subset of table-restructuring operators, which fall short in the Auto-Tables task as we will show experimentally. It is also worth pointing out that unlike Auto-Tables that takes no examples, these systems require users to provide *one example output table*, which is a significant amount of effort for users.

Computer vision models for object detection. Substantial progress has been made in the computer vision literature on object detection, with variants of CNN architectures being developed to extract salient visual features from pictures [35, 42, 51].

Given the "visual" nature of our problem shown in Figure 1, and the strong parallel between "pixles" in images and "rows/columns" in tables, both of which form two-dimensional rectangles, our model architecture is inspired by CNN-architectures for object detection, but specifically designed for our table transformation task.

Representing tables using deep models. Different techniques have been proposed to represent tables using deep models (e.g., TaBERT [57], Tapas [37], Turl [32], etc.). Most of these focus on natural-language (NL) aspects of tables, and tailor to NL-related tasks (e.g., NL-to-SQL and entity-linking [37, 57]), which we show are not suited for our table-transformation task, as it needs to exploit the structural homogeneity of tables (e.g., cell similarity in row/column-directions.).

Database schema design. There is a body of classical database research on schema design, which typically involves normalizing or decomposing one large table into multiple smaller tables, so that the decomposed tables satisfy relational "normal forms" (3NF, BCNF, etc.) [40], that can improve storage efficiency and avoid update anomalies, among other things. In contrast, our work has the goal of restructuring an input table to make it easy to query, which is always *single-table to single-table*, and thus both orthogonal and complementary to schema design (e.g., our transformed table can then be subject to schema-design steps if it needs to be stored in databases).

3 PRELIMINARY AND PROBLEM

In this section, we will introduce the table-restructuring operators considered in this work, and describe our synthesis problem.

3.1 Table-restructuring operators

We consider 8 table-restructuring operators in our DSL, which are listed in Table 1. Based on our analysis of tables in the wild (in user spreadsheets and on the web), these operators cover a majority of scenarios required to relationalize tables. Note that since our synthesis framework uses self-supervision for training that is not tied to the specific choices of operators, our approach can be easily extended to include additional operators for new functionalities.

[1] Available at https://github.com/LiPengCS/Auto-Tables-Benchmark.

	B	C	D	E	F
1	Adams Elementary	Aki Kurose Middle School	Alki Elementary	B.F. Day Elementary	...
2	ES	MS	ES	ES	...
3	553	685	373	282	...
4	580	719	377	296	...
5	609	754	380	310	...
6	638	791	384	326	...
7	670	829	388	341	...
8	702	870	392	358	...

transpose

	A	B	C	D	E	F	G	H
1	School name	GradeID	2015	2016	2017	2018	2019	2020
2	Adams Elementary	ES	553	580	609	638	670	702
3	Aki Kurose Middle School	MS	685	719	754	791	829	870
4	Alki Elementary	ES	373	377	380	384	388	392
5	B.F. Day Elementary	ES	282	296	310	326	341	358
6	...	...	...	...	...	...	...	...

stack

	A	B	C	D
1	School name	GradeID	Year	Num
2	Adams Elementary	ES	2015	553
3	Adams Elementary	ES	2016	580
4	Adams Elementary	ES	2017	609
5	Adams Elementary	ES	2018	638
6	Adams Elementary	ES	2019	670
7	Adams Elementary	ES	2020	702
8	Aki Kurose Middle School	MS	2015	685
9	Aki Kurose Middle School	MS	2016	719
10	...	...	...	...

Figure 4: An example input table (on the left) that requires two transformation steps to relationalize: (1) a "`transpose`" step to swap rows and columns, (2) a "`stack`" step to collapse homogeneous columns (C to H) into two new columns. The resulting output table (on the right) becomes substantially easier to query with SQL (e.g., to filter and aggregate).

In this section, we will introduce the first 4 operators and their parameters shown in Table 1 (we will give additional details in our technical report [1] in the interest of space).

Stack. `Stack` is a Pandas operator [20] (also known as `melt` and `unpivot` in other contexts), that collapses contiguous blocks of homogeneous columns into two new columns. Like shown in Figure 1(a), column headers of the homogeneous columns ("`19-Oct`", "`20-Oct`", etc.) are converted into values of a new column called "`Date`", making it substantially easier to query (e.g., to filter using a range-predicate on the "`Date`" column).

Parameters. In order to properly invoke `stack`, one needs to provide two important parameters, `start_idx` and `end_idx` (listed in the third column of Table 1), which specify the starting and ending column index of the homogeneous column-group that needs to be collapsed. In the case of Figure 1(a), we should use `start_idx=3` (corresponding to column D) and `end_idx=12` (column M).

Note that because in Auto-Tables we aim to synthesize complete transformation steps that can execute on input tables, which requires us to predict not only the operators (e.g., `stack` for the table in Figure 1(a)), but also the exact parameters values correctly (e.g., slightly different parameters such as `start_idx=4` and `end_idx=12` would fail to produce the desired transformation).

Wide-to-long. `Wide-to-long` is an operator in Pandas [22], that collapses repeating column groups into rows (similar functionality can also be found in R [24]). Figure 1(b) shows such an example, where "`Revenue/Units Sold/Margin`" from different years form column-groups that repeat once every 3 columns. All these repeating column-groups can collapse into 3 columns, with an additional "`Year`" column for year info from the original column headers, as shown on the right in Figure 1(b). Observe that `wide-to-long` is similar in spirit to `stack` as both collapse homogeneous columns, although `stack` cannot produce the desired outcome when columns are repeating in groups, as is the case in this example.

Parameters. `wide-to-long` has 3 parameters, where `start_idx` and `end_idx` are similar to the ones used in `stack`. It has an additional parameter called "`delim`", which is the delimitor used to split the original column headers, to produce new column headers and data-values. For example, in the case of Figure 1(b), "`delim`" should be specified as "` - `" to produce: (1) a first part corresponding to values for the new "`Year`" column ("`2018`", "`2019`", etc.); and (2) a second part corresponding to the new column headers in the transformed table ("`Revenue`", "`Units Sold`", etc.). Like in `stack`, all 3 parameters here need to be instantiated correctly, before we can synthesize the desired transformation.

Transpose. `Transpose` is a table-restructuring operator that converts rows to columns and columns to rows, which is also used in other contexts such as in matrix computation. Figure 1(c) shows an example input table on the left, for which `transpose` is needed to produce the output table shown on the right, which would become relational and easy to query.

Parameters. Invoking `transpose` requires no parameters, as all rows and columns will be transposed.

Pivot. Like `transpose`, `pivot` also converts rows to columns, as the example in Figure 1(d) shows. However, in this case rows show repeating-groups (whereas in `wide-to-long` columns show repeating-groups), which need to be transformed into columns, like shown on the right of Figure 1(d).

Parameters. `Pivot` has one parameter, "`repeat_frequency`", which specifies the frequency at which the rows repeat in the input table. In the case of Figure 1(d), this parameter should be set to 4, as the color pattern of rows would suggest.

Additional operators. Table 1 has 4 additional table-restructuring operators, which we will briefly mention here. These include (1): "**`explode`**" [16], which converts columns with composite values (violating the First Normal Form [30]) into atomic values, so that the table can be queried using standard SQL; (2): "**`ffill`**" [17] that fills values in structurally empty cells so that the table can be queried; (3): "**`subtitle`**" that converts rows representing table sub-titles into separate columns for ease of queries; and finally (4): a "**`none`**" operator for input tables that are already relational, for which no transformation is needed, which is needed explicitly so that we do not "over-trigger" on tables that require no transformation.

3.2 Problem statement

Given these table-restructuring operators listed in Table 1, we now introduce our synthesis problem as follows.

Definition 1. Given an input table T, and a set of operators $\mathbf{O} = \{stack, transpose, pivot, \ldots\}$, where each operator $O \in \mathbf{O}$ has a parameter space $P(O)$. Synthesize a sequence of multi-step transformations $M = (O_1(p_1), O_2(p_2), \ldots, O_k(p_k))$, with $O_i \in \mathbf{O}$ and $p_i \in P(O_i)$ for all $i \in [k]$, such that applying each step $O_i(p_i) \in M$ successively on T produces a relationalized version of T.

Note that in our task, we need to predict both the operator O_i and its exact parameters p_i correctly, each step along the way. This is challenging as the search space is large – even for a single-step transformation, there are thousands of possible operators/parameters to choose from (e.g., a table with 50 columns that requires "`stack`" will have 50x50 = 2500 possible parameters of start_idx and end_idx); for two-step transformations the search space is already in the millions (e.g., for "`stack`" alone it is $2500^2 \approx 6M$). Given the large search space, even a small difference in parameters can render the resulting transformation incorrect, as shown below.

Example 1. Given the input table T shown on the left of Figure 4, the ground-truth transformation M to relationalize T has two-steps: M = (transpose(), stack(start_idx:"2015", end_idx:"2020")). Here the first step "`transpose`" swaps the rows with columns, and the second step "`stack`" collapses the homogeneous columns (between column "`2015`" and "`2020`"). Note that this is the only correct sequence of steps – reordering the two steps, or using slightly different parameters (e.g., start_idx="2016" instead of "2015"), will all lead to incorrect output, which makes the problem challenging.

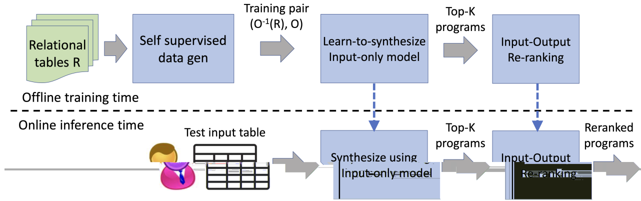


Figure 5: Architecture overview of AUTO-TABLES

Also note that although we show synthesized programs using our DSL syntax, the resulting programs can be easily translated into different target languages, such as Python Pandas or R, which can then be directly invoked. We should also note that two syntactically different programs M_1 and M_2 may be semantically equivalent, which can be verified under a set of algebraic rules. [2]

4 AUTO-TABLES: LEARN-TO-SYNTHESIZE

We now describe our proposed AUTO-TABLES system, which learns to synthesize transformations. We will start with an architecture overview before we delve into individual components.

4.1 Architecture overview

We represent our overall architecture in Figure 5. The system operates in two modes, with the upper-half of the figure showing the offline training-time pipeline, and the lower-half showing the online inference-time steps.

At offline training time, AUTO-TABLES uses three main components: (1) A "training data generation" component that consumes large collections of relational tables R, to produce (example, label) pairs; (2) An "input-only synthesis" module that learns-to-synthesize using the training data, and (3) An "input-output re-ranking" module that holistically considers both the input table and the output table (produced from the synthesized program), to find the most likely program.

The online inference-time part closely follows the offline steps, where we directly invoke the two models trained offline (the last two blue boxes shown in the figure). When given an input table from users, we pass the table through our input-only synthesis model, to identify top-k candidate programs, which are then re-ranked by the input-output model for final predictions.

We now describe these three modules in turn below.

4.2 Self-supervised training data generation

As discussed earlier, the examples in Figure 1 demonstrate that there are clear patterns in the input tables that we can exploit (e.g., repeating column-groups and row-groups) to predict required transformations for a given table. Note that these patterns are "visual" in nature, which can likely be captured by computer-vision-like algorithms.[3]

The challenge however, is that unlike computer vision tasks that typically have large amounts of training data (e.g., ImageNet [31])

[2] For example, pivot is equivalent to transpose followed by wide-to-long, and wide-to-long is equivalent to stack-split-pivot. Furthermore, the order of ffill and stack/wide-to-long can be swapped, as long as they operate on disjoint subsets of columns, etc. In our synthesis, we consider synthesized programs that are semantically equivalent to the ground-truth program also correct.

[3] Like computer vision problems such as object detection where hand-crafted heuristics are hard to write, the row/column-level patterns existing in our tables are also hard to write with heuristics, which makes a learning-based method necessary.

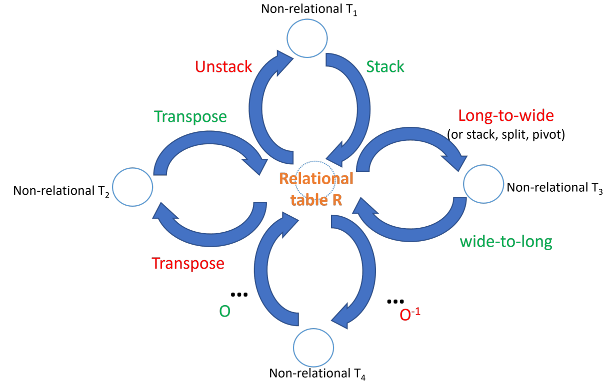


Figure 6: Leverage inverse operators to generate training data. In order to learn-to-synthesize operator O, we can start from any relational table R, apply its inverse operator O^{-1} to obtain $O^{-1}(R)$. Given $T = O^{-1}(R)$ as an input table, we know O must be its ground-truth transformation, because $O(O^{-1}(R)) = R$.

in the form of (image, label) pairs, in our synthesis task, there is no existing labeled data that we can leverage. Labeling tables manually from scratch are likely too expensive to scale.

Leverage inverse operators. To overcome the lack of data, we propose a novel self-supervision framework leveraging the inverse functional-relationships between operators, to automatically generate large amounts of training data without using humans labels.

Figure 6 shows the overall idea of this approach. For each operator O in our DSL that we want to learn-to-synthesize, we can find its inverse operator (or construct a sequence of steps that are functionally equivalent to its inverse), denoted by O^{-1}. For example, in the figure we can see that the inverse of "`transpose`" is "`transpose`", the inverse of "`stack`" is "`unstack`", while the inverse of "`wide-to-long`" can be constructed as a sequence of 3 steps ("`stack`" followed by "`split`" followed by "`pivot`").

The significance of the inverse operators, is that it allows us to automatically generate training examples. Specifically, to build a training example for operator O (e.g., "`stack`"), we can sample any relational table R, and apply the inverse of O, or O^{-1} (e.g., "`unstack`"), to generate a non-relational table $T = O^{-1}(R)$. For our task, given T as input, we know O must be its ground-truth transformation, since by definition $O(T) = O(O^{-1}(R)) = R$, and R is known to be relational. This thus allows us to generate (T, O) as an (example, label) pair, which can be used for training.

Furthermore, we can easily produce such training examples at scale, by sampling: (1) different relational tables R; (2) different operators O; and (3) different parameters associated with each O, therefore addressing our lack of data problem in AUTO-TABLES.

The overall steps of the data generation process are shown in Algorithm 1, where Line 2, Line 3, Line 6 correspond to the sampling of operators (O), tables (R), and parameters (p), respectively, that together creates diverse training examples. We note that in Line 4, we perform an additional "data augmentation" step to create even more diversity in training, which we explain below.

Data Augmentation. Data augmentation [50] is a popular technique in computer vision and related fields, to enhance training data and improve model robustness. For example, in computer vision tasks, it is observed that training using additional data generated from randomly flipped/rotated/cropped images, can lead to improved model performance (because an image that contains

Algorithm 1: Auto-gen training examples

input : DSL operators **O**, large collections of relational tables **R**
output : Training table-label pairs: (T, O_p)

```
1  E ← {}
2  foreach O in O do
3  |  foreach R in R do
4  |  |  foreach R' in Augment(R) // Crop rows and columns
5  |  |  do
6  |  |  |  p ← sample valid parameter from space P(O)
7  |  |  |  O^{-1}_{p'} ← construct the inverse of O_p
8  |  |  |  T ← O^{-1}_{p'}(R')
9  |  |  |  E ← E ∪ {(T, O_p)}
10 return all training examples E
```

an object, say dog, should still contain the same object after it is flipped/rotated, etc.) [50].

In the same spirit, we augment each of our relational table R by (1) Cropping, or randomly sampling contiguous blocks of rows and columns in R to produce a new table R'; and (2) Shuffling, or randomly reordering the rows/columns in R to create a new R'. In Auto-Tables, we start from over 15K relational tables crawled from public sources (Section 5), and create around 20 augmented tables for each relational table R. This further improves the diversity of our training data and end-to-end model performance, as we will show in the experiments.

4.3 Input-only Synthesis

After obtaining large amounts of training data in the form of (T, O_p) using self-supervision, we now describe our "input-only" model that takes T as input, to predict a suitable transformation O_p.

4.3.1 *Model architecture.*

We develop a computer-vision inspired model specifically designed for our task, which scans through rows and columns to extract salient tabular features, reminiscent of how computer-vision models extract features from image pixels for object detection.

Our model architecture in Figure 7 consists of four sets of layers: (1) table embedding, (2) dimension reduction, (3) feature extraction, and (4) output layers. We will describe each in turn below.

Table embedding layers. Given an input table T, the embedding layer encodes each cell in T into a vector, to obtain an initial representation of T for training. At a high level, for each cell we want to capture both (1) the "*semantic features*" (e.g., people-names vs. company-names), and (2) the "*syntactic feature*" (e.g., data-type, string-length, punctuation, etc.), because both semantic and syntactic features provide valuable signals in our task, e.g., in determining whether rows/columns are homogeneous or similar.

For semantic features, we use the pre-trained Sentence-BERT [48] (a state-of-the-art embedding in NLP), which maps each cell into a 384-dimension vector that encodes its semantic meaning. For syntactic features, we encode each cell using 39 pre-defined syntactic attributes (data types, string lengths, punctuation, etc.). Concatenating the syntactic and semantic features produces a 423-dimension vector for each cell. For an input table T with n rows and m columns, this produces a $n \times m \times 423$ tensor as its initial representation.[4]

The left half of Figure 8 shows a simple sketch of this embedding step, which we will explain in more detail later.

Dimension reduction layers. Since the initial representation from the pre-trained Sentence-BERT has a large number of dimensions (with information likely not needed for our task, which can slow down training and increase the risk of over-fitting), we add dimension-reduction layers using two convolution layers with 1×1 kernels, to reduce the dimensionality from 423 to 64 and then to 32, to produce $n \times m \times 32$ tensors. Note that we explicitly use 1×1 kernels so that the trained weights are shared across all table-cells, to produce consistent representations after dimension reduction.

Feature extraction layers. We next have feature extraction layers that are reminiscent of CNN [43] but specifically design for our table task. Recall from Figure 1 that the key signals for our task are:

- (1) identify whether values in row or column-directions are "similar" enough to be "homogeneous" (e.g., Figure 1(b) vs. Figure 1(c));
- (2) identify whether entire rows or columns are "similar" enough to show repeating patterns (e.g., Figure 1(b) vs. Figure 1(d)).

Intuitively, if we were to hand-write heuristics, then signal (1) above can be extracted by comparing the representations of adjacent cells in row- and column-directions. On the other hand, signal (2) can be extracted by computing the average representations of each row and column, which can then be used to find repeating patterns.

Based on this intuition, and given the strong parallel between the row/columns in tables and pixels in images, we design feature-extraction layers inspired by *convolution filters* [43] that are popular in CNN architectures to extract visual features from images [42, 51]. Specifically, as shown in Figure 7, we use 1x2 and 1x1 convolution filters followed by average-pooling, in both row- and column-directions, to represent rows/columns/header. Unlike general *nxm* filter used for image tasks (e.g., 3x3 and 5x5 filters in VGG [51] and ResNet [35]), our design of filters are tailored to our table task, because:

- (a) 1x2 filters can easily learn-to-compute signal (1) above (e.g., 1x2 filters with +1/-1 weights can identify the representation differences between neighboring cells, which when averaged, can identify homogeneity in row/column directions).
- (b) 1x1 filters can easily learn-to-compute signal (2) above (e.g., 1x1 filters with +1 weights followed by average-pooling, correspond to representations for entire rows/columns, which can be used to find repeating patterns in subsequent layers).

We use an example below to demonstrate why these 1x1 and 1x2 filters are effective for extracting tabular features.

Example 2. Figure 8(a) shows a simplified example, when using Column-B of Figure 1(a) as input, which has a list of values "`Sports`", "`Electronics`", etc. These raw cell values first pass through the embedding step, which produces a row of features for each value, with both *syntactic features* (under the headers "`is-string`", "`str-length`", etc.), and *semantic features* (under the header "`s-BERT`" for sentence-BERT). This results in an embedding table, where each row corresponds to an input cell.

[4]Like in computer vision problems that use a fixed "window", we take the first 100 data-rows (plus a header) and 50 columns at the top-left of each input T (producing a $101 \times 50 \times 423$ tensor), which is sufficient to identify table patterns and predict transformations, like the examples in Figure 1 would show.

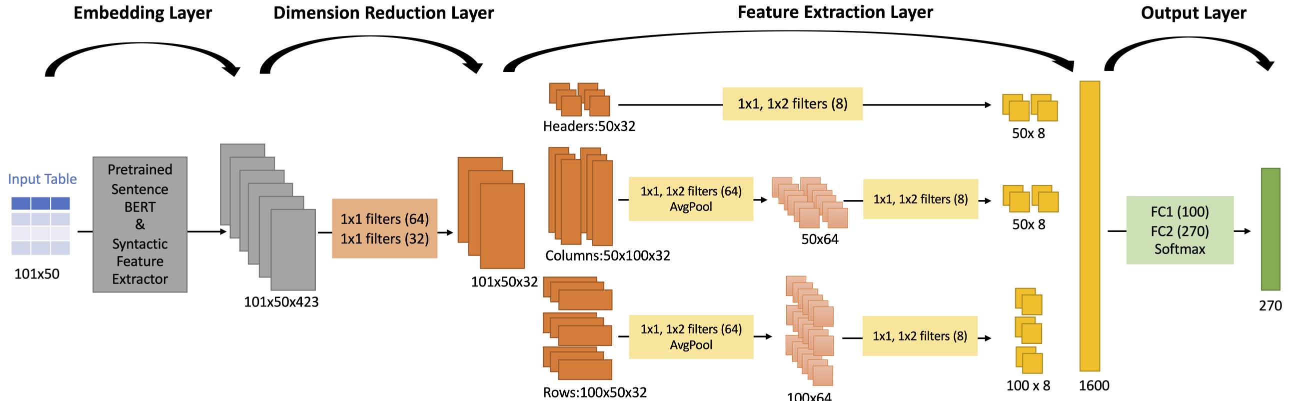


Figure 7: Input-only synthesis: model architecture.

Next, we pass this embedding table through 1x1 and 1x2 convolution filters, which performs element-wise dot-product [43]. Assuming we have a simple 1x1 filter shown at the top of the figure, with weights $[1, 0, \ldots]$. Because only the first bit of this simple filter is 1 and the rest is 0, performing a dot-product on the embedding table essentially only extracts the "`is-string`" type information of each cell, which in this case is all 1, leading to a matrix of $[1, 1, 1, 1]$ (since all cells are of type string). After average pooling, this results in a single feature-value 1 to represent a specific aspect of this entire column (in this case, type information).

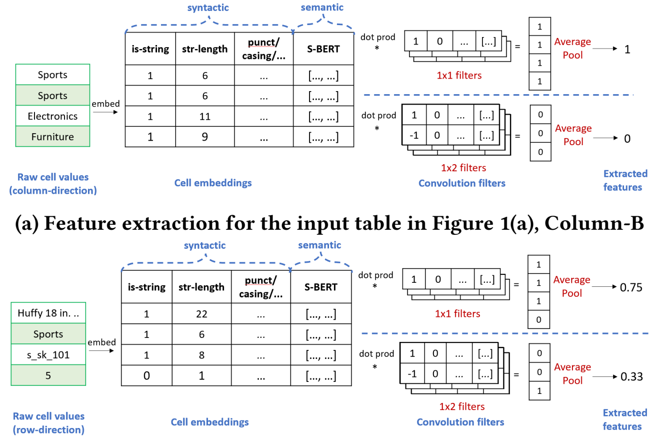
We should note that this is just one example 1x1 filter – there exists many such 1x1 filters (shown as stacked in the figure), all of which have learned weights that extract different aspects of syntactic/semantic information from input cells (string-length, semantic-meaning, etc.), thus forming a holistic representation of values in the column, to facilitate downstream comparison of "similar" columns (e.g., to identify repeating rows/columns), as mentioned above as signal (2) for our task.

The 1x2 filters, on the other hand, work to "compare" adjacent values in the same column, which intuitively test for homogeneity. For instance, assuming there is a simple 1x2 filter with only +1 and -1 weights in the first column, as shown in the figure. Performing a dot-product in this case "compares" the "`is-string`" type info for neighboring cells, using a sliding window for rows from the top to bottom, which results in $[0, 0, 0]$ (because the convolution computes $1*1+1*(-1) = 0$). This is again averaged to produce a feature-value 0, indicating no type difference, and thus good homogeneity, in the list of given values in the column-direction.

This is again only one example 1x2 filter – there are many other 1x2 filters with different learned-weights (stacked in the figure) that use different syntactic/semantic features to test for homogeneity between neighboring cells, which corresponds to the signal (1) we want to extract as mentioned earlier.

Recall that our CNN-inspired architecture uses convolution filters to scan line-by-line, in both row and column directions. So in the row-direction our filters work in a similar manner.

The same operations in row-direction is shown in Figure 8(b), which uses Row-2 of the table in Figure 1(a) as example. In this case we have a list of heterogeneous cell values "`Huffy 18 in.`", "`Sports`", "`s_sk_101`", "`5`", etc. In this case, performing a dot-product using the same 1x2 filter produces a feature-vector of $[0, 0, 1]$ (note that the last entry is 1 because the "`is-string`" value for the last two input cells are 1 and 0, leading to a convolution of $1*1+(-1)*0 = 1$). Average-pooling would then produce 0.33 here, indicating inconsistent types for the list of values in the row-direction (0 would



(a) Feature extraction for the input table in Figure 1(a), Column-B

(b) Feature extraction for the input table in Figure 1(a), Row-2

Figure 8: Example feature extraction using 1x1 and 1x2 filters

indicates homogeneity, with +1/-1 filter-weights). Other 1x2 filters would work in similar manners, to identify more signals of heterogeneity in the row-direction, all of which are important ingredients to identify latent patterns in the table and corresponding transformations.

These first-level of features-values from row/column-directions will then go through a second-level of 1x1 and 1x2 convolution filters, to compare and identify similar rows/columns (based on row/column representation from 1x1 filters), to ultimately reveal repeating rows and columns like the color-coded patterns show in Figure 1. These tabular features will pass down to the next output layers, for final classifications.

Output layers. Our output layers use two fully connected layers followed by softmax classification, as shown in Figure 7, which produces an output vector that encodes both the predicted operator-type, and its parameters. For example, since we consider 8 possible operator types in our DSL, we encode this as a 8-dimension one-hot vector. Similarly, we represent parameters of each operator as additional bits in the same output vector, resulting in an output vector of 270 dimensions, which in effect makes multiple predictions (operator-type and parameters) simultaneously, for a given T.

We apply standard softmax functions [45] on each prediction vector, so that the output of each prediction is normalized into a probability distribution.

4.3.2 *Training and inference.*

We now describe how we train this model shown in Figure 7, and at inference time, use it to synthesize transformations.

Training time: Loss Function. Given a training input table T, its ground truth operator O and corresponding parameters $P = (p_1, p_2, ...)$, let $\hat{O}$ and $\hat{P} = (\hat{p}_1, \hat{p}_2, ...)$ be the model predicted probability distributions of O and P respectively. The training loss on T can be computed as the sum of loss on all predictions (both the operator-type, and parameters relevant to this operator):

$$Loss(T) = L(O, \hat{O}) + \sum_{p_i \in P, \hat{p_i} \in \hat{P}} L(p_i, \hat{p_i}) \quad (1)$$

Here $L(y, \hat{y})$ denotes the cross-entropy loss [45] commonly used in classification – let y be a n-dimensional ground truth one-hot vector, and $\hat{y}$ a model predicted vector, $L(y, \hat{y})$ is defined as:

$$L(y, \hat{y}) = -\sum_{i=1}^{n} y_i log(\hat{y}_i) \quad (2)$$

Given large amounts of training data $\mathbf{T}$ (generated from our self-supervision in Section 4.2), we train our Auto-Tables model by minimizing the overall training loss $\sum_{T \in \mathbf{T}} Loss(T)$ using gradient descent until convergence. We will refer to this trained model as H.

Inference time: Synthesizing transformations. At inference time, given an input T, our model H produces a probability for any candidate step O_P that is instantiated with operator O and parameters $P = (p_1, p_2, \ldots)$, denoted by $Pr(O_P|T)$, as:

$$Pr(O_P|T) = Pr(O) \cdot \prod_{p_i \in P} Pr(p_i) \quad (3)$$

Using the predicted probabilities, finding the most likely transformation step O_P^* given T is then simply:

$$O_P^* = \arg\max_{O,P} Pr(O_P|T) \quad (4)$$

This gives us the most likely one-step transformation given T. As we showed in Figure 4, certain tables may require multiple transformation steps for our task.

To synthesize multi-step transformations, intuitively we can invoke our predictions step-by-step until no suitable transformation can be found. Specifically, given an input table T, at step (1) we can find the most likely transformation O_P^1 for T using Equation (4), such that we can apply O_P^1 on T to produce an output table $O_P^1(T)$. We then iterate, and at step (2) we feed $O_P^1(T)$ as the new input table into our model, to predict the most likely $O_P^2(T)$, and produce an output table $O_P^2(O_P^1(T))$. This iterates until at the k-th step, a "none" transformation is predicted (recall that "none" is a no-op operator in our DSL in Table 1, to indicate that the input table is already relational and requires no transformations). The resulting $M = (O_P^1, O_P^2, \ldots)$ then becomes the multi-step transformations we synthesize for the original T.

The procedure above is an intuitive sketch of multi-step synthesis, though it considers only the top-1 choice at each step. In general we need to consider top-k choices at each step, to find the most likely multi-step transformations overall. We perform the general search procedure of the most likely top-k steps using beam search [45], as outlined in Algorithm 2.

We start in Algorithm 2 with an empty pipeline M and the original input table T. At each iteration, we invoke model H on top-k output tables from the last iteration, to obtain the top k candidate operators for each (Line 6). We perform the predicted transformations and expand each M with one additional predicted step to get M_{next} (Line 8), whose probability can be computed as the product of the probability of its operators (Line 9). If a predicted operator is "none", we reach a terminal state and save it as a candidate pipeline (Line 10-11). Otherwise, we keep the current pipeline in the beam for further search (Line 13). At the end of each iteration, we rank all partial pipelines by probabilities, and keep only the top k pipelines with the highest probability (Line 14). We terminate the search after a total of L steps (Line 3), and return the top-k with the highest probabilities as output (Line 15-16).

Algorithm 2: Multi-step pipeline synthesis by top-k search

```
input  : Auto-Tables model H, input table T
output : Top-k predicted pipelines by probabilities: M_1, M_2 ... M_k
1  Cands = [], M ← [], M.prob = 1 // initialize
2  B_cur ← [(T, M)]
3  for i = 1, 2, ... L do
4      B_next ← []
5      foreach (T, M) in B_cur do
6          Ô_p1, Ô_p2., .Ô_pk ← H(T) // top k predictions
7          for j = 1, 2, ...k do
8              T_next ← Ô_pj(T), M_next ← M.append(Ô_pj)
9              M_next.prob ← M.prob × Ô_pj.prob
10             if Ô_pj = none then
11                 Cands.append(M_next)
12             else
13                 B_next.append((T_next, M_next))
14     sort B_next by M.prob, B_cur ← B_next[: k]
15 Sort Cands by M.prob
16 return Cands[: k]
```

We demonstrate Algorithm 2 using the following example.

Example 3. We revisit Example 1. Given the input table T shown on the left of Figure 4, we invoke our trained model H to predict likely transformations, where the top-2 is: (1) O_1: "transpose" with probability 0.5, which leads to an output table $O_1(T)$ (shown in the middle of Figure 4), (2) O_2: "stack" (with parameters: start-idx = Col-B, end-idx=Col-E) which also has a probability 0.5, that will lead to an output table $O_2(T)$. We keep both 1-step candidates $\{O_1, O_2\}$, and continue our search of possible second steps.

For the second step, if we follow the path of O_1 we will operate on $O_1(T)$ as the new input table, for which the top-2 predicted steps is: (1) O_3 "stack" (start-idx = Col-C, end-idx=Col-E), with probability 0.8, and (2) O_4 "none" with probability 0.1. Alternatively, if we follow the path of O_2 we would have $O_2(T)$ as the new input, for which we also generate its top-2. This leads to a total of $2 \times 2 = 4$ possible 2-step transformations, from which we pick the top-2 with the highest probabilities, to continue our search with 3-steps, etc.

We rank all resulting multi-step transformations by probabilities. This returns $\{O_1, O_3\}$ as the most likely (with probability 0.5*0.8 = 0.4), which is indeed the desired transformation as discussed in Example 1.

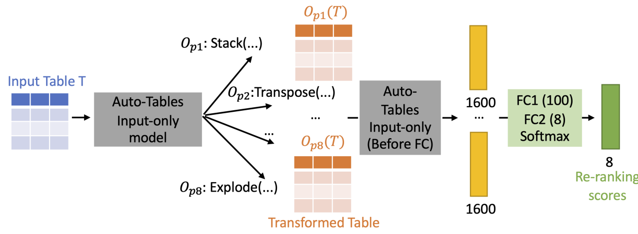


Figure 9: Input/output re-ranking: model architecture.

4.4 Input/output Re-ranking

So far, our synthesis model is "input-only", as it only uses the characteristics of the input table T to predict transformations M. However, sometimes this is not enough, as the characteristics of the output table, $M(T)$ would also provide useful signals. We illustrate this using the following example.

EXAMPLE 4. In Example 3, based only on the input T in Figure 4, our model predicts both O_1 "transpose" and O_2 "stack" as possible choices (both with probability=0.5). "Stack" was incorrectly ranked high, because from T alone "stack" looks plausible, as T has a large number of homogeneous columns (Col-B to E), which fits the typical pattern for "stack" as shown in Figure 1(a).

We can better predict whether O_1 or O_2 is more suitable, if we apply both programs on T and inspect the resulting output $O_1(T)$ and $O_2(T)$. It can be verified that for $O_1(T)$ values in the same columns are homogeneous, whereas $O_2(T)$ (using "stack") leads to a table where values such as "ES", "MS" (from "GroupID") become intermixed with integers in the same columns, which is not homogeneous and not ideal, and is something that our tabular model can detect and penalize. Inspecting the output $O_1(T)$ and $O_2(T)$ thus allows us to correctly re-rank O_1 as a more likely transformation than O_2, which is difficult when a model looks at T alone.

This motivates us to develop an "input/output-based" re-ranking model as shown in Figure 9. After the input-only synthesis model (Section 4.3) produces top-k likely operators $\{O_{pi}, i \in [k]\}$ (e.g., we consider top-8 operators for re-ranking in our experiments), the re-ranking model will look at all output transformed tables $\{O_{pi}(T), i \in [k]\}$ and aims to generate a re-ranking score for each of them indicating which operator is more suitable based on the output transformed tables. To do so, similar to the input-only model, we need to first convert each transformed table into a feature vector using table embedding, dimension reduction and feature extraction layers. Since the input-only model has been trained well at this time, we directly reuse the architecture and weights of these layer from the pre-trained input-model [5]. We then concatenate the feature vectors of all transformed tables and use fully connected layers followed by a softmax function to produce a k-dimension vector as re-ranking scores. For training, we consider the re-ranking as a classification task to predict which of the k transformed tables is the ground truth. Thus, the training loss can be computed using cross-entropy loss. We train the re-ranking model using the same training data generated from self-supervision in Section 4.2

[5]We remove the fully-connected output layers from the input model, which are specific to predicting synthesis outcomes and not relating to extracting table features.

5 EXPERIMENTS

We perform extensive evaluation on the performance of different algorithms, using real test data. The results show that our method significantly outperforms the baseline methods in terms of both quality and efficiency. Our labeled benchmark data is available on GitHub[6] for future research.

5.1 Experimental Setup

Benchmarks. To study the performance of our method in real-world scenarios, we compile an ATBench benchmark using real cases from three sources: (1) online user forums, (2) Jupyter notebooks, and (3) real spreadsheet-tables and web-tables.

Forums. Both technical and non-technical users ask questions on forums, regarding how to restructure their tables. As Figure 3 shows, users often provide sample input/output tables to demonstrate their needs. We sample 23 such questions from StackOverflow and Excel user forums as test cases. (We feed Auto-Tables with user-provided input tables, and evaluate whether the correct transformation can be synthesized to produce the desired output table given by users).

Notebooks. Data scientists frequently restructure tables using Python Pandas, often inside Jupyter Notebooks. We sample 79 table-restructuring steps extracted from the Jupyter Notebooks crawled in [55, 56] as our test cases. We use the transformations programmed by data scientists as the ground truth.

Excel+Web. A large fraction of tables "in the wild" require transformations before they are fit for querying, as shown in Figure 1 and 2. We sample 56 real web-tables and 86 spreadsheet-tables (crawled from a search engine) that require such transformations, and manually write the desired transformations as the ground truth.

Combining these sources, we get a total of 244 test cases as our ATBench (of which 26 cases require multi-step transformations). Each test case consists of an input table T, ground-truth transformations M_g[7], and an output table $M_g(T)$ that is relational. Detailed statistics of the benchmark can be found in our technical report [1].

Evaluation Metrics. We evaluate the quality and efficiency of different algorithms in synthesizing transformations.

Quality. Given an input table T, an algorithm A may generate top-k transformations $(\hat{M}_1, \hat{M}_2, ...\hat{M}_k)$, ranked by probabilities, for users to inspect and pick. We evaluate the success rate of synthesis using the standard $Hit@k$ metric [49], defined as:

$$Hit@k(T) = \sum_{i=1}^{k} \mathbf{1}(\hat{M}_i(T) = M_g(T))$$

which looks for exact matches between the top-k ranked predictions $(\hat{M}_i(T), 1 \leq i \leq k)$ and the ground-truth $M_g(T)$. The overall $Hit@k$ on the entire benchmark, is then simply the average across all test cases T. We report $Hit@k$ up to $k = 3$.

Efficiency. We report the latency of synthesis using wall-clock time. All experiments are conducted on a Linux VM with 24 vCPU cores, and 4 Tesla P100 GPUs.

[6]https://github.com/LiPengCS/Auto-Tables-Benchmark

[7]It should be noted that for some test cases, there may be more than one transformation sequence that can produce the desired output. We enumerate all such sequences in our ground-truth, and mark an algorithm as correct as long as it can synthesize one ground-truth sequence.

Table 2: Quality comparison using Hit@k, on 244 test cases

Method	No-example methods				By-example methods			
	Auto-Tables	TaBERT	TURL	GPT-3.5-fs	FF	FR	SQ	SC
Hit @ 1	**0.570**	0.193	0.029	0.196	0.283	0.336	0	0
Hit @ 2	**0.697**	0.455	0.071	-	-	-	0	0
Hit @ 3	**0.75**	0.545	0.109	-	-	-	0	0
Upper-bound	-	-	-	-	0.471	0.545	0.369	0.369

Table 3: Synthesis latency per test case

Method	Auto-Tables	Foofah (excl. 110 timeout cases)	FlashRelate (excl. 91 timeout cases)
50 %tile	**0.127s**	0.287s + human effort	3.4s + human effort
90 %tile	**0.511s**	22.891s + human effort	57.16s + human effort
95 %tile	**0.685s**	39.188s + human effort	348.6s + human effort
Average	**0.224s**	5.996s + human effort	59.194s + human effort

Table 4: Ablation Studies of Auto-Tables

Method	Full	No Re-rank	No Re-rank &				
			No Aug	No Bert	No Syn	1x1 Only	5x5
Hit@1	**0.570**	0.508	0.463	0.467	0.504	0.471	0.480
Hit@2	**0.697**	0.652	0.582	0.627	0.648	0.607	0.594
Hit@3	**0.75**	0.730	0.656	0.693	0.676	0.652	0.660

Table 5: Sensitivity to different semantic embeddings.

Embedding methods	sentenceBERT	fastText	GloVe	No Semantic
Hit@1	0.508	0.529	0.525	0.467
Hit@2	0.652	0.656	0.676	0.627
Hit@3	0.730	0.734	0.734	0.734
Avg. latency per-case w/ this embedding	0.299s	0.052s	0.050s	0.026s

Methods Compared. We compare with the following methods.

- *Auto-Tables*. This is our approach and is the only method that does not require users to provide input/output examples (unlike other existing methods). In order to train Auto-Tables, we generate 1.4M (input-table, transformation) pairs evenly distributed across 8 operators, following the self-supervision procedure (Section 4.2), using 15K base relational tables crawled from public sources [8]. We take a fixed size of input with the first 100 rows and 50 columns at the top-left corner of each table and use zero-padding for tables with less rows or columns. We implement our method using PyTorch [47], trained using Adam optimizer, with a learning rate of 0.001 for 50 epochs, using a batch size of 256.
- *Foofah (FF)* [38] synthesizes transformations based on input/output examples. We use 100 cells from the top-right of the ground-truth output table for Foofah to synthesize programs, which simulate the scenario where a user types in 100 output cells (a generous setting as it is unlikely that users are willing to provide so many examples in practice). We test Foofah using the authors original implementation [14], and we time-out each case after 30 minutes.
- *Flash-Relate (FR)* [28] is another approach to synthesize table-level transformations, which however would require input/output examples. We used an open-source re-implementation of FlashRelate [25] (since the original system is not publicly available), and we provide it with 100 example output cells from the ground-truth. We use a similar time-out of 30 minutes for each test case.
- *SQLSynthesizer (SQ)* [58] is a SQL-by-example algorithm that synthesizes SQL queries based on input/output examples. We use the authors implementation [26], provide it with 100 example output cells, and also set a time-out of 30 minutes.
- *Scythe (SC)* [54] is another SQL-by-example method. We used authors implementation [23] and provide it with 100 example output cells, like previous methods.
- *TaBERT* [57] is a table representation approach developed in the NLP literature, and pre-trained using table-content and captions for NL-to-SQL tasks. To test the effectiveness of TaBERT in our transformation task, we replace the table representation in Auto-Tables (i.e., output of the feature extraction layer in Figure 7) with TaBERT's representation, and train the following fully connected layers using the same training data as ours.
- *TURL* [32] is another table representation approach for data integration tasks. Similar to *TaBERT*, we test the effectiveness of TURL by replacing Auto-Tables representation with TURL's.
- *GPT* [29] is a family of large language models pre-trained on text and code, which can follow instructions to perform a variety of tasks. While we do not expect GPT to perform well on Auto-Tables tasks, we perform a comparison nevertheless, using GPT-3.5[9] as a baseline. We perform few-shot in-context learning, using a description of the operators, together with pairs of (input-table, desired-operator) in the prompt to demonstrate the task. We provide one example demonstration per operator, for a total of 7 examples (which fit in the context allowed by GPT-3.5). We denote this method as GPT-3.5-fs (few-shot).[10]

5.2 Experiment Results

Quality Comparison. Table 2 shows the comparison between Auto-Tables and baselines, evaluated on our benchmark with 244 test cases. We group all methods into two classes: (1) "No-example methods" that do not require users to provide any input/output examples, which include our Auto-Tables, and variants of Auto-Tables that use TaBERT and TURL for table representations, respectively; and (2) "By-example methods" that include Foofah (FF), FlashRelate (FR), SQLSynthesizer (SQ), and Scythe (SC), all of which are provided with 100 ground truth example cells.

As we can see, Auto-Tables significantly outperforms all other methods, successfully transforming 75% of test cases in its top-3, *without needing users to provide any examples*, despite the challenging nature of our tasks. Recall that in our task, even for a single-step transformation, there are thousands of possible operators+parameters to choose from (e.g., a table with 50 columns that requires "`stack`" will have 50x50 = 2,500 possible parameters of start_idx and end_idx) and for two-step transformations, the search space is in the millions (e.g., for "`stack`" alone it is $2500^2 \approx 6M$), which is clearly non-trivial.

Compared to other no-example methods, Auto-Tables outperforms TaBERT and TURL respectively by 37.7 and 54.1 percentage point on Hit@1, 20.5 and 64.1 percentage point on Hit@3. This shows the strong benefits for using our proposed table representation and model architecture, which are specifically designed for the table transformation task (Section 4.3).

Compared to by-example methods, the improvement of Auto-Tables is similarly strong. Considering the fact that these baselines use 100 output example cells (which users need to manually type), whereas our method uses 0 examples, we argue that Auto-Tables is clearly a better fit for the table-restructuring task at hand. Since

[8]We use the dataset from [44], which has thousands of relational Power-BI models crawled from public sources. We sample 15K fact and dimension tables from these models as our "base" relational tables. Since our training data is collected via Power-BI data models, they are completely separate from our test data (Web and Excel tables).

[9]We used the "gpt-3.5-turbo" API endpoint, accessed in July 2023.

[10]Note that GPT-3.5-fs is still a no-example method, as we use general-purpose examples to demonstrate each operator in our few-shot examples, which are fixed and do not vary based on different input tables.

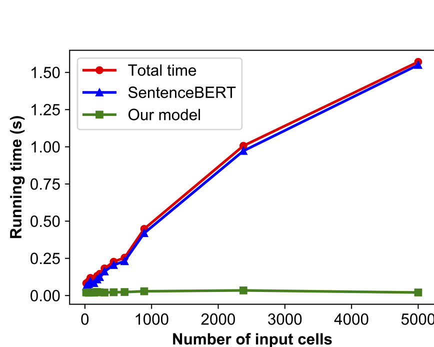


Figure 10: Auto-Tables latency analysis

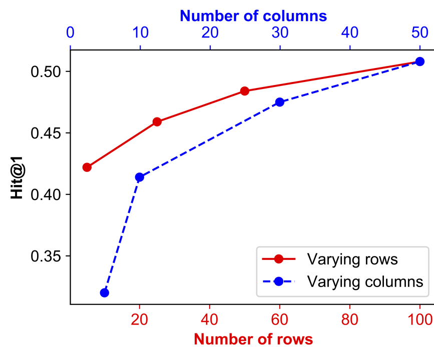


Figure 11: Vary input size

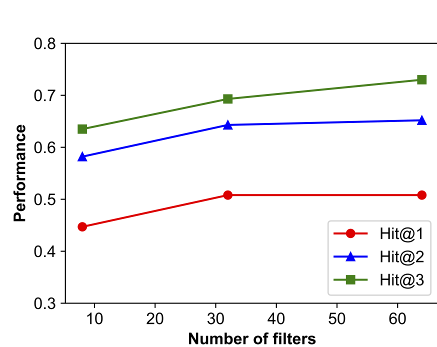


Figure 12: Vary number of filters

some of these methods (FF and FR) only return top-1 programs, we also report in the last row their "upper-bound" coverage, based on their DSL (assuming all transformations supported in their DSL can be successfully synthesized).

Additional quality results. We report additional results on quality, such as a breakdown by benchmark sources, and Hit@K in the presence of input tables that are already relational (for which Auto-Tables should correctly detect and not over-trigger, by performing no transformations), in our technical report [1].

Running Time. Table 3 compares the average and 50/90/95-th percentile latency, of all methods to synthesize one test case. Auto-Tables is interactive with sub-second latency on almost all cases, whose average is 0.224. Foofah and FlashRelate take considerably longer to synthesize, even after we exclude cases that time-out after 30 minutes. This is also not counting the time that users would have to spend typing in output examples for these by-example methods, which we believe make Auto-Tables substantially more user-friendly for our transformation task.

Figure 10 shows the average latency of Auto-Tables, on cases with different number of non-empty input cells. As we can see, the latency grows linearly as the number of cells increases, but since we only need to use at most the top-left 100 rows and 50 columns to correctly synthesize a program, this is always bounded by a couple of seconds at most. Furthermore, we notice that the running time is dominated by SentenceBERT embedding, which accounts for 91.5% of the latency. In comparison, the actual inference time of Auto-Tables (the green line) is very small and almost constant.

Ablation Study We perform ablation studies to understand the benefit of Auto-Tables components, which is shown in Table 4.

Contribution of Input/Output Re-Ranking. To study the contribution of our re-ranking model (Section 4.4), we compare the performance of Auto-Tables with and without re-ranking. Table 4 shows that our "Full" method (with re-ranking) produces substantially better Hit@1 and Hit@2 compared to "No Re-rank".

Contribution of Data Augmentation. To study the benefits of data augmentation in training data generation (Section 4.2), we disable augmentation when generating training data (i.e., using only the base relational tables). Table 4 shows this result under "No Aug", which suggests that our Hit@k drop substantially, underscoring the importance of data augmentation.

Contribution of Embeddings. Recall that we use both syntactic embedding and semantic embedding (sentenceBERT) to represent each cell (Section 4.3). To understand their contributions, we remove each embedding in turn, and the results are shown under "No Bert" and "No Syntactic" in Table 4. Both results show a substantial drop in performance, confirming their importance (semantic embedding with sentenceBERT is likely more important, as removing it leads to a more significant drop).

Contribution of 1D Filters. Recall that we use convolution filters of size 1x1 and 1x2 to extract features from rows and columns (Section 4.3). To understand the effectiveness of this design, we evaluate our method with alternative filters. First, we replace all the 1x2 filters with 1x1 filters. The result is labeled "1x1 Only" and shows a significant drop. Second, we replace all filters with filters of size 5x5 that is common in computer vision tasks [42, 51], which leads to another substantial drop. Both results confirm the effectiveness of our model design that is tailored to table tasks.

Sensitivity analysis We perform sensitivity analysis to understand the effect of different settings in Auto-Tables.

Varying Input Size. In Auto-Tables, we feed the top 100 rows and left-most 50 columns from the input table T into the model, which is typically enough to correctly predict the right transformations. To understand its effect on model performance, in Figure 11, we vary the number of rows/columns used here and show the input-only model performance. As we can see, when we increase the number of rows/columns that the model uses, the resulting quality improves until it plateaus at about 30 columns and 50 rows.

Varying Number of Filters. Figure 12 shows the quality of Auto-Tables input-only model with different numbers of convolution filters (the total number of 1x1 and 1x2 filters for rows/columns before AvgPool in the feature extraction layer in Figure 7). As we can see, using 32 filters is substantially better than 4 filters, as it can extract more features. However, the improvement beyond 32 filters is not significant, suggesting diminishing returns beyond a certain level of model capacity.

Additional results. We report additional results such as sensitivity to different embeddings, error analysis, and accuracy of parameter predictions, in [1] in the interest of space.

6 CONCLUSIONS AND FUTURE WORK

We propose a new problem of synthesizing transformations to relationalize tables. By leveraging visual characteristics of input tables using compute-vision-inspired algorithms, we obviate the need for users to provide input/output examples, which is a substantial departure from prior work. Future directions include extending the functionality to a broader set of operators, and exploring the applicability of this technique on other classes of transformations.

ACKNOWLEDGMENTS

We thank Dr. Kexin Rong and Dr. Xu Chu for their generous support and valuable feedback, as well as three anonymous VLDB reviewers for their helpful comments on our manuscript.

REFERENCES

[1] [n.d.]. Auto-Tables: full version. https://arxiv.org/abs/2307.14565.
[2] [n.d.]. Example Excel forum question: Hard to query without transformations (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/power-query-data-cleaning-unpivot-transpose-etc/m-p/2400300.
[3] [n.d.]. Example Excel forum question: Hard to query without transformations (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/unpivot-grouped-data/m-p/3686239.
[4] [n.d.]. Example Excel forum question: Hard to query without transformations (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/unpivot-monthly-data/m-p/1867836.
[5] [n.d.]. Example Excel forum question: Table analysis provides unexpected results (Retrieved in 02/2023). https://answers.microsoft.com/en-us/msoffice/forum/all/excel-ideas-feature/c9574cf9-dccc-4356-95d3-07d268e39d82.
[6] [n.d.]. Example Excel forum question to relationalize tables: Data restructuring using Excel (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/data-restructuring-using-excel/m-p/287547.
[7] [n.d.]. Example Excel forum question to relationalize tables: Pivot chart 4 columns (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/pivot-chart-4-columns-set-responses-to-4-questions/m-p/2329880.
[8] [n.d.]. Example Excel forum question to relationalize tables: Pivot table issue. (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/pivot-table-issue/m-p/3015448.
[9] [n.d.]. Example Excel forum question to relationalize tables: Transpose data for analysis (Retrieved in 02/2023). https://techcommunity.microsoft.com/t5/excel/transposing-data-for-better-analysis/m-p/1297106.
[10] [n.d.]. Example StackOverflow forum question to relationalize tables: Melt index (Retrieved in 02/2023). https://stackoverflow.com/questions/53917303/pandas-melt-with-multi-index-data-set-and-resetting-index-why-is-this-working.
[11] [n.d.]. Example StackOverflow forum question to relationalize tables: Melt multiple columns (Retrieved in 02/2023). https://stackoverflow.com/questions/51519101/simultaneously-melt-multiple-columns-in-python-pandas.
[12] [n.d.]. Example StackOverflow forum question to relationalize tables: Melt with multiple value vars (Retrieved in 02/2023). https://stackoverflow.com/questions/45066873/pandas-melt-with-multiple-value-vars.
[13] [n.d.]. Example StackOverflow forum question to relationalize tables: Reshape wide-to-long in Pandas (Retrieved in 02/2023). https://stackoverflow.com/questions/36537945/reshape-wide-to-long-in-pandas.
[14] [n.d.]. Foofah code on GitHub. https://github.com/umich-dbgroup/foofah.
[15] [n.d.]. Pandas API in Python. https://pandas.pydata.org/.
[16] [n.d.]. Pandas operator: Explode. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.explode.html.
[17] [n.d.]. Pandas operator: FFill. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.ffill.html.
[18] [n.d.]. Pandas operator: Melt. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.melt.html.
[19] [n.d.]. Pandas operator: Pivot. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.pivot.html.
[20] [n.d.]. Pandas operator: Stack. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.stack.html.
[21] [n.d.]. Pandas operator: Transpose. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.transpose.html.
[22] [n.d.]. Pandas operator: Wide-to-long. (Retrieved in 02/2023). https://pandas.pydata.org/docs/reference/api/pandas.wide_to_long.html.
[23] [n.d.]. PATSQL code on GitHub. https://github.com/NAIST-SE/PATSQL.
[24] [n.d.]. R operator: pivot-longer, which is similar to Wide-to-long. (Retrieved in 02/2023). https://tidyr.tidyverse.org/reference/pivot_longer.html.
[25] [n.d.]. Reimplementation of FlashRelate code on GitHub. https://github.com/BEE-Synth/Bee/tree/291a824622e36fccfa43461e85be3f836e3f4eff/Eval/Benchmarks/Spreadsheet/flashrelate-01.
[26] [n.d.]. Scythe code on GitHub. https://github.com/Mestway/Scythe.
[27] [n.d.]. Trifacta: Standardize Using Patterns. (Retrieved in 07/2023). https://docs.trifacta.com/display/DP/Standardize+Using+Patterns.
[28] Daniel W Barowy, Sumit Gulwani, Ted Hart, and Benjamin Zorn. 2015. FlashRelate: extracting relational data from semi-structured spreadsheets using examples. *ACM SIGPLAN Notices* 50, 6 (2015), 218–228.
[29] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
[30] Edgar F Codd. 1990. *The relational model for database management: version 2.* Addison-Wesley Longman Publishing Co., Inc.
[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition.* Ieee, 248–255.
[32] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. Turl: Table understanding through representation learning. *ACM SIGMOD Record* 51, 1 (2022), 33–40.
[33] Yihan Gao, Silu Huang, and Aditya Parameswaran. 2018. Navigating the data lake with datamaran: Automatically extracting structure from log datasets. In *Proceedings of the 2018 International Conference on Management of Data.* 943–958.
[34] Sumit Gulwani, William R Harris, and Rishabh Singh. 2012. Spreadsheet data manipulation using examples. *Commun. ACM* 55, 8 (2012), 97–105.
[35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.
[36] Yeye He, Xu Chu, Kris Ganjam, Yudian Zheng, Vivek Narasayya, and Surajit Chaudhuri. 2018. Transform-data-by-example (TDE) an extensible search engine for data transformations. *Proceedings of the VLDB Endowment* 11, 10 (2018), 1165–1177.
[37] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349* (2020).
[38] Zhongjun Jin, Michael R Anderson, Michael Cafarella, and HV Jagadish. 2017. Foofah: Transforming data by example. In *Proceedings of the 2017 ACM International Conference on Management of Data.* 683–698.
[39] Zhongjun Jin, Yeye He, and Surajit Chauduri. 2020. Auto-transform: learning-to-transform by patterns. *Proceedings of the VLDB Endowment* 13, 12 (2020), 2368–2381.
[40] William Kent. 1983. A simple guide to five normal forms in relational database theory. *Commun. ACM* 26, 2 (1983), 120–125.
[41] Martin Koehler, Edward Abel, Alex Bogatu, Cristina Civili, Lacramioara Mazilu, Nikolaos Konstantinou, Alvaro AA Fernandes, John Keane, Leonid Libkin, and Norman W Paton. 2019. Incorporating data context to cost-effectively automate end-to-end data wrangling. *IEEE Transactions on Big Data* 7, 1 (2019), 169–186.
[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *CACM* 60, 6 (2017), 84–90.
[43] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. 2021. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems* (2021).
[44] Yiming Lin, Yeye He, and Surajit Chaudhuri. 2023. Auto-BI: Automatically Build BI-Models Leveraging Local Join Prediction and Global Schema Graph. *Proceedings of the VLDB Endowment* (2023).
[45] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective.* MIT press.
[46] Arash Dargahi Nobari and Davood Rafiei. 2022. Efficiently transforming tables for joinability. In *2022 IEEE 38th International Conference on Data Engineering (ICDE).* IEEE, 1649–1661.
[47] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).
[48] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://arxiv.org/abs/1908.10084
[49] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval.* Vol. 39. Cambridge University Press Cambridge.
[50] Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data* 6, 1 (2019), 1–48.
[51] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[52] Keita Takenouchi, Takashi Ishio, Joji Okada, and Yuji Sakata. 2020. PATSQL: efficient synthesis of SQL queries from example tables with quick inference of projected columns. *arXiv preprint arXiv:2010.05807* (2020).
[53] Quoc Trung Tran, Chee-Yong Chan, and Srinivasan Parthasarathy. 2009. Query by output. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data.* 535–548.
[54] Chenglong Wang, Alvin Cheung, and Rastislav Bodik. 2017. Synthesizing highly expressive SQL queries from input-output examples. In *SIGPLAN.* 452–466.
[55] Cong Yan and Yeye He. 2020. Auto-suggest: Learning-to-recommend data preparation steps using data science notebooks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.* 1539–1554.
[56] Junwen Yang, Yeye He, and Surajit Chaudhuri. 2021. Auto-pipeline: synthesizing complex data pipelines by-target using reinforcement learning and search. *Proceedings of the VLDB Endowment* (2021).
[57] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. *arXiv preprint arXiv:2005.08314* (2020).
[58] Sai Zhang and Yuyin Sun. 2013. Automatically synthesizing sql queries from input-output examples. In *2013 28th IEEE/ACM International Conference on Automated Software Engineering (ASE).* IEEE, 224–234.
[59] Erkang Zhu, Yeye He, and Surajit Chaudhuri. 2017. Auto-join: Joining tables by leveraging transformations. *Proceedings of the VLDB Endowment* 10, 10 (2017), 1034–1045.