ICE-T: Interactions-aware Cross-column Contrastive Embedding for Heterogeneous Tabular Datasets

Tomas Tokar[1, 2], **Scott Sanner**[2, 3]

[1]Wondeur AI
[2]University of Toronto
[3]Vector Institute for AI

Abstract

Finding high-quality representations of heterogeneous tabular datasets is crucial for their effective use in downstream machine learning tasks. Contrastive representation learning (CRL) methods have been previously shown to provide a straightforward way to learn such representations across various data domains. Current tabular CRL methods learn joint embeddings of data instances (tabular rows) by minimizing a contrastive loss between the original instance and its perturbations. Unlike existing tabular CRL methods, we propose leveraging frameworks established in multimodal representation learning, treating each tabular column as a distinct modality. A naive approach that applies a pairwise contrastive loss to tabular columns is not only prohibitively expensive as the number of columns increases, but as we demonstrate, it also fails to capture interactions between variables. Instead, we propose a novel method called ICE-T that learns each columnar embedding by contrasting it with aggregate embeddings of the complementary part of the table, thus capturing interactions and scaling linearly with the number of columns. Unlike existing tabular CRL methods, ICE-T allows for column-specific embeddings to be obtained independently of the rest of the table, enabling the inference of missing values and translation between columnar variables. We provide a comprehensive evaluation of ICE-T across diverse datasets, demonstrating that it generally surpasses the performance of the state-of-the-art alternatives.

Introduction

Heterogeneous tabular datasets constitute an extremely important, yet often overlooked, data class, which remains to be challenging for application of deep neural networks (Shwartz-Ziv and Armon 2022). As with other data types the key is to find good quality data representations that facilitate the downstream tasks (Bengio, Courville, and Vincent 2013). Contrastive representation learning (CRL) offers a very straightforward way for learning such representations, without the need of associated data labels (Le-Khac, Healy, and Smeaton 2020; Ericsson et al. 2022).

Multimodal CRL Recently, there has been increasing attention on a type of CRL methods known as multimodal CRL. Multimodal CRL has traditionally been applied to data comprising multiple distinct modalities, such as image-text or audio-video pairs, from which it derives its name (Deldari et al. 2022; Zong, Mac Aodha, and Hospedales 2023). Multimodal CRL offers a unique way of learning modality-specific embeddings that are coordinated with embeddings from other modalities while allowing inputs of one modality to be embedded independently of the others (Guo, Wang, and Wang 2019). Therefore, embeddings can be obtained even in the absence of some modalities and can be used to

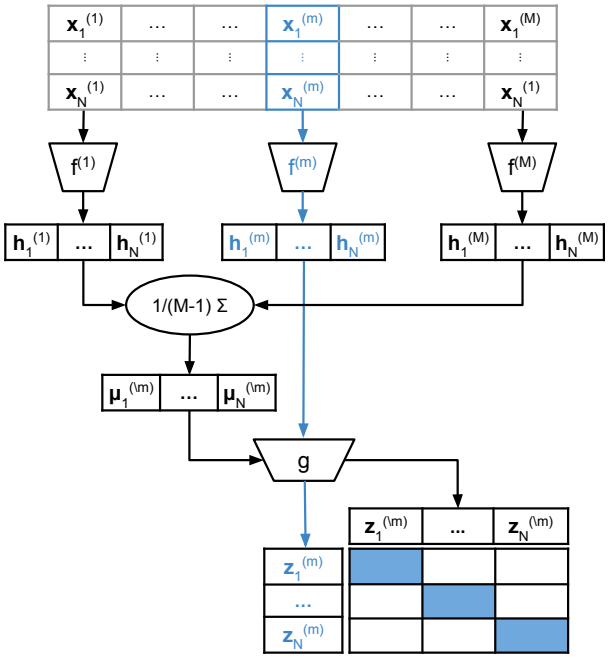

Figure 1: A schematic representation of ICE-T. Tabular entries are passed through variable-specific encoders $f^{(m)}$ to produce intermediate latent representations $\mathbf{h}^{(m)}$. For each variable m we compute the associated *anchor* $\mu^{(\setminus m)}$ as the average of the representations of the remaining variables. These resulting vectors are then projected via a shared neural subnetwork g to produce the final embeddings. The cosine similarities of these embeddings are used to compute the contrastive loss for each variable.

infer the values of the missing ones—a task often referred to as *modality translation* (Kaur, Pannu, and Malhi 2021). The idea of adopting these methods for learning representations of heterogeneous tabular data is thus very appealing.

Motivation Intuitively, heterogeneous tabular data can be treated as multimodal data, where each variable (tabular column) is considered a single modality. The central idea is to learn variable-specific mappings that project inputs into a common latent space so that the embeddings of inputs are similar if they belong to the same instance (tabular row) while dissimilar otherwise. A naive approach would involve computing similarities between all pairs of variable-specific embeddings. However, such a pairwise contrasting scales quadratically with the number of variables and becomes prohibitively costly as the number of variables increases. Moreover, pairwise contrastive learning fails to capture interactions between variables, which can corrupt the resulting embeddings (e.g, dataset depicted in Figure 2). This situation is common in tabular data, which typically contain interacting variables, and can render pairwise embedding approaches inapplicable (Borisov et al. 2022).

Proposed method To address this problem, we propose a novel CRL approach called **I**nteractions-aware Cross-column **C**ontrastive **E**mbedding for heterogeneous **T**abular datasets (ICE-T) (Figure 1). ICE-T contrasts column-specific embeddings with the embedded aggregates of the remaining columns in *linear time* and allows to account for interactions among variables, while preserving the advantages of multimodal CRL. ICE-T is a simple, versatile, and data-agnostic approach specifically designed for heterogeneous tabular data but can be easily applied to other domains. Overall, ICE-T offers superior or generally competitive performance compared to other methods.

Contributions (1) We offer a simple, easily reproducible example that illustrates the failure of multimodal CRL methods to capture interactions among variables (2) We introduce a novel method called ICE-T that addresses this limitation (3) We provide a comprehensive experimental comparison between ICE-T, five CRL methods and shallow learning benchmarks across a range of datasets, including frequently overlooked image/text–tabular data; measuring performance in four tasks (i) cross-modal translation, (ii) clustering, (iii) supervised learning and (iv) transfer learning.

Related Work

Previous efforts to apply CRL to heterogeneous tabular data utilized joint CRL, where tabular rows undergo random transformations to produce their perturbed analogs, and the latent representations are learned by contrasting the original data against their perturbed analogs. This approach is best exemplified by **Scarf** (Bahri et al. 2021) and **SubTab** (Ucar, Hajiramezanali, and Edwards 2021). In Scarf, rows are perturbed by replacing inputs from a random subset of variables by values drawn from the respective variable marginals. In SubTab rows are perturbed by randomly dividing variables to create overlapping subsets. In addition to contrastive embedding loss, the authors of SubTab propose using additional distance loss and reconstruction loss.

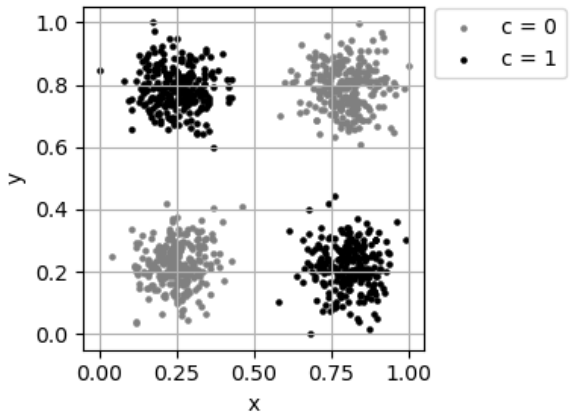

Figure 2: The XOR blobs can form a heterogeneous tabular dataset with three variables: x, y and c, where each variable depends on the interaction between the other two. As we demonstrate, this interaction poses a challenging problem for multimodal CRL methods.

Unlike these methods, we propose to borrow frameworks established in the domain of multimodal representation learning and to approach tabular data as multimodal datasets, where each tabular column is treated as a single modality. From this perspective we identified three methods that are closely related to our line of research:

CLIP (Radford et al. 2021) produces text and image embeddings that are similar if the input text and image are related, and dissimilar otherwise. It consists of modality-specific encoders, followed by linear projections into a common latent space, which are trained under the InfoNCE loss (Oord, Li, and Vinyals 2018), using cosine similarity, across a very large collection of image-text pairs. While the term "CLIP" usually refers to a specific model, it may denote a CRL method that trains modality-specific mappings only through pairwise cosine similarity across modalities (Table 1). Under this generalization, CLIP provides a versatile approach that can be applied to any type of modalities, as exemplified by (Alayrac et al. 2020; Wang et al. 2021; Guzhov et al. 2022).

GMC (Poklukar et al. 2022) was originally proposed for multimodal time-series data where modality-specific mappings are trained in coordination with joint representation learning. It consists of modality-specific encoders followed by deep neural projection heads to map inputs into a common latent space. Additionally, GMC projects inputs into the same latent space jointly through a designated neural subnetwork. The objective of GMC is to maximize the cosine similarity between the modality-specific embeddings and the joint embedding belonging to the same instance, while minimizing the similarity among embeddings belonging to different instances. The minimized term includes pairwise similarities between modality-specific embeddings, the similarity between joint and modality-specific embeddings, and the

Method	Batch loss
Scarf	$\mathcal{L} = \sum_i^B -\log \frac{\exp\left(s(\mathbf{z}_i^{(1:M)}, \tilde{\mathbf{z}}_i^{(1:M)})/\tau\right)}{\sum_j \exp\left(s(\mathbf{z}_i^{(1:M)}, \tilde{\mathbf{z}}_j^{(1:M)})/\tau\right)}$
SubTab	$\mathcal{L} = \sum_i^B \sum_{\mathcal{A},\mathcal{B}} \Big[-\log \frac{\exp\left(s(\mathbf{z}_i^{(\mathcal{A})}, \mathbf{z}_i^{(\mathcal{B})})/\tau\right)}{\sum_j \mathbb{1}_{j\neq i} \exp\left(s(\mathbf{z}_i^{(\mathcal{A})}, \mathbf{z}_j^{(\mathcal{B})})/\tau\right)} + (\mathbf{z}_i^{(\mathcal{A})} - \mathbf{z}_i^{(\mathcal{B})})^2 \Big]$
CLIP	$\mathcal{L} = \sum_i^B \sum_m^M \sum_{n\neq m}^M -\log \frac{\exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)})/\tau\right)}{\sum_j \exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(n)})/\tau\right)}$
GMC	$\mathcal{L} = \sum_i^B \sum_m^M -\log \frac{\exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(1:M)})/\tau\right)}{\sum_{j\neq i} \exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(1:M)})/\tau\right) + \exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(m)})/\tau\right) + \exp\left(s(\mathbf{z}_i^{(1:M)}, \mathbf{z}_i^{(1:M)})/\tau\right)}$
MCN	$\mathcal{L} = \sum_i^B \Big[\sum_m^M \sum_{n\neq m}^M -\log \frac{\exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)}) - \delta\right)}{\exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(n)}) - \delta\right) + \sum_{j\neq i} \exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(n)})\right)} - \sum_m^M \log \frac{\exp\left(s(\mathbf{z}_i^{(m)}, C_i') - \delta\right)}{\sum_k^K \exp\left(s(\mathbf{z}_i^{(m)}, C_k)\right)} \Big]$
ICE-T	$\mathcal{L} = \sum_i^B \sum_m^M -\log \frac{\exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_i^{(\setminus m)})/\tau\right)}{\sum_j \exp\left(s(\mathbf{z}_i^{(m)}, \mathbf{z}_j^{(\setminus m)})/\tau\right)}$

Table 1: Unified comparison of contrastive loss functions used by state-of-the-art methods vs. ICE-T (ours); s indicates cosine similarity, τ is learnable parameter $\mathbf{z}^{(1:M)}$ indicates joint instance embedding; $\mathbf{z}^{(\mathcal{A})}$ and $\mathbf{z}^{(\mathcal{B})}$ indicate partial instance embeddings; δ is a hyperparameter; C_i' indicates the centroid that is nearest to the i-th fused multimodal feature and C_k is the k-th centroid; g is a neural function.

similarity between pairs of joint embeddings.

MCN (Chen et al. 2021) combines multimodal CRL with latent space clustering. It encodes inputs via modality-specific encoders and associated linear projections. Subsequently, MCN calculates joint representations by aggregating modality-specific embeddings together via an arithmetic mean. These joint embeddings, referred to by the authors as *fused multimodal features*, are subjected to online k-means clustering (Cohen-Addad et al. 2021). Euclidean distance between the features and a set of k centroids is computed and each feature is assigned the nearest centroid. In addition to maximizing pairwise similarity among the modality-specific embeddings, MCN aims to maximize the similarity between these embeddings and the centroids that are nearest to the respective multimodal features; the authors also add the reconstruction loss to the overall loss of the model.

Algorithm 1: ICE-T – batch loss computation.

Inputs:
$\{x^{(1)}, \ldots, x^{(M)}\}$ {Data batch}
$\{f_{(1)}, \ldots, f^{(M)}\}$ {Mappings}
g, l {Projection, Loss function}
$\Sigma \leftarrow 0$ {Initialize sum}
for $m = 1$ **to** M **do**
  $h^{(m)} \leftarrow f^{(m)}(x^{(m)})$ {Intermediate representations}
  $\Sigma \leftarrow \Sigma + h^{(m)}$ {Update sum}
end for
$L \leftarrow 0$ {Initialize loss}
for $m = 1$ **to** M **do**
  $\mu^{(\setminus m)} \leftarrow (\Sigma - h^{(m)})/(M-1)$ {Calculate anchor}
  $z_1 \leftarrow g(h^{(m)})$ {Apply g}
  $z_2 \leftarrow g(\mu^{(\setminus m)})$
  $s \leftarrow s(z_1, z_2)$ {Get similarities}
  $L \leftarrow L + l(s)$ {Update loss}
end for

Interactions-aware Cross-column Contrastive Embedding

Preliminaries

Let $\mathbf{x}_i$ denote the i-th data instance, comprising M inputs from different variables: $\mathbf{x}_i = (\mathbf{x}_i^{(m)})_{m=1}^M$. Assume that each input $\mathbf{x}^{(m)}$ comes from a distinct variable-specific input space $\mathcal{X}^{(m)}$: $\mathbf{x}^{(m)} \in \mathcal{X}^{(m)}$. Furthermore, consider $\mathbf{x} \sim p(\mathbf{x})$ and $\mathbf{x}^{(m)} \sim p(\mathbf{x}^{(m)})$, where $p(\mathbf{x})$ denotes the joint data distribution, and $p(\mathbf{x}^{(m)})$ denotes the marginal distribution of given variable m; and let $p(\mathbf{x}^{(m)}|\mathbf{x}^{(\setminus m)})$ denote the conditional probability of $\mathbf{x}^{(m)}$ given the inputs $\mathbf{x}^{(\setminus m)} = (\mathbf{x}^{(n)})_{n\neq m}^M$.

Rationale

The central idea of ICE-T is to take intermediate hidden representations $\mathbf{h}^{(m)}$ obtained by variable-specific encoders $f^{(m)}$ and fuse the intermediate representations of the remaining variables into a single vector $\mu^{(\setminus m)}$, which we refer to as the *anchor*. The anchors are computed via the arithmetic mean: $\mu^{(\setminus m)} = 1/(M-1)\sum_{n\neq m} \mathbf{h}^{(n)}$, where M is the total number of variables. Integrating by averaging, instead of learning a joint representation via a fixed architecture (i) encourages additive embedding representations and (ii) makes the method robust against missing inputs and hence more versatile.

The two hidden vectors $\mathbf{h}^{(m)}$ and $\mu^{(\setminus m)}$ are passed through the identical neural subnetwork g to obtain final latent representations $\mathbf{z}^{(m)} = g(\mathbf{h}^{(m)})$ and $\mathbf{z}^{(\setminus m)} = g(\mu^{(\setminus m)})$, which are then compared by cosine similarity s. We will use $s^{(m)}(i,j)$ to denote similarity between the pair of latent vectors belonging to i-th and j-th instance: $s^{(m)}(i,j) :=$

Method	Embedding form		
	$\mathbf{z}^{(m)}$	$\mathbf{z}^{(1:M)}$	$\mathbf{z}^{(\mathcal{A})}$
Scarf		✓	
SubTab		✓	
CLIP	✓	!	!
GMC	✓	✓	!
MCN	✓	✓	!
ICE-T	✓	✓	✓

Table 2: ICE-T provides variable-specific $\mathbf{z}^{(m)}$, joint instance $\mathbf{z}^{(1:M)}$ and also partial instance embeddings $\mathbf{z}^{(\mathcal{A})}$ for any subset $\mathcal{A}$ of input variables generically, whereas some methods allow *post hoc* embeddings aggregation ("!").

$s(g(\mathbf{h}_i^{(m)}), g(\mu_j^{(\setminus m)}))$. The aim is to maximize the similarity between the embedding vectors and the matching anchors $s^{(m)}(i,i)$, while minimizing the non-matching ones $s^{(m)}(i,j)$.

For this purpose, we adopt the InfoNCE loss (Oord, Li, and Vinyals 2018), so the loss incurred from the i-th instance on variable m is given by:

$$l_i^{(m)} = -\log \frac{\exp(s^{(m)}(i,i)/\tau)}{\sum_j \exp(s^{(m)}(i,j)/\tau)} \tag{1}$$

Here τ is the "temperature", a trainable parameter controlling the sensitivity of the loss across the range of similarity values. The total batch loss is then the sum of losses incurred across all instances and modalities: $L = \sum_{(i,m)} l_i^{(m)}$ (cf. Algorithm 1).

Probabilistic Interpretation

We seek variable-specific neural mappings $f^{(m)} : \mathcal{X}^{(m)} \to \mathbb{R}^q$ and an additional neural mapping $g : \mathbb{R}^q \to \mathbb{R}^d$, such that for any $\mathbf{x}_i^{(m)} \in \mathcal{X}^{(m)}$, for $\forall m \in \{1, \dots M\}$, produce embeddings: $\mathbf{z}_i^{(m)} = g(f^{(m)}(\mathbf{x}_i^{(m)}))$; such that

$$p(\mathbf{x}^{(m)}|\mathbf{x}^{(\setminus m)}) \propto s(\mathbf{z}^{(m)}, \mathbf{z}^{(\setminus m)}) \tag{2}$$

where $s : \mathbb{R}^{M \times d} \to \mathbb{R}$ is a similarity function that quantifies the similarity between the variable-specific embedding $\mathbf{z}^{(m)}$ and the embedding of the remaining variables $\mathbf{z}^{(\setminus m)} = g\big(1/(M-1)\sum_{n \neq m} f^{(n)}(\mathbf{z}^{(n)})\big)$. Equation 2 thus implies that we want to maximize the similarity between the embeddings of inputs that are likely to be associated and minimize it otherwise.

Benefits

Embedding versatility Once trained, ICE-T can provide variable-specific embeddings $\mathbf{z}^{(m)}$, joint complete instance embeddings $\mathbf{z}^{(1:M)}$, and joint partial instance embeddings $\mathbf{z}^{(\mathcal{A})}$ of a values from any subset $\mathcal{A} \in \mathcal{P}(M)$, where $\mathcal{P}(M)$ is the power set of the variables it was trained on. This versatility is thanks to the use of arithmetic aggregation of the intermediate representation, instead of a fixed neural architecture (a used by, for example, in (Poklukar et al. 2022)). This gives ICE-T an advantage over many other methods, which require *post hoc* aggregation of the variable-specific embeddings (cf. Table 2).

Modality translation The important advantage of most multimodal CRL methods, including ICE-T, is support for modality translation. In the context of heterogeneous tabular datasets this can be viewed as a form of missing values imputation, where the value of one variable is estimated based on the values of the other modalities of the given instance. This can be formulated as the assignment:

$$\hat{\mathbf{x}}^{(m)} = \underset{\mathbf{x}^{(m)} \in \mathcal{X}^{(m)}}{\operatorname{argmax}} \; p(\mathbf{x}^{(m)}|\mathbf{x}^{(\mathcal{A})}) \tag{3}$$

where $\mathcal{A}$ is a subset of evidence variables so that $\mathcal{A} \in \mathcal{P}(M)$ and $m \notin \mathcal{A}$. Assuming $p(\mathbf{x}^{(m)}|\mathbf{x}^{(\mathcal{A})}) \propto s(\mathbf{z}^{(m)}, \mathbf{z}^{(\mathcal{A})})$, the above assignment converts to:

$$\mathbf{x}^{(m)} = \underset{\mathbf{x}^{(m)} \in \mathcal{X}^{(m)}}{\operatorname{argmax}} \; s\big(\mathbf{z}^{(m)}, \mathbf{z}^{(\mathcal{A})}\big) \tag{4}$$

which can be solved provided the input space $\mathcal{X}^{(m)}$ is well defined and can be sampled efficiently.

Linear scaling Methods that rely on pairwise contrasting do not scale well with the large number of variables due to the quadratic increase in the number of pairwise contrastive losses. Unlike these, ICE-T *scales linearly with the number of variables*, as can be seen in Algorithm 1, where we can compute the sum of all embeddings in one linear pass over modalities and then calculate the loss in the second linear pass by contrasting the modality-specific embeddings against their respective anchors computed in a leave-one-out manner. This provides an important computational benefit to our method.

Data-agnostic ICE-T can be readily adapted to accommodate various data types, including image or text variables, by selecting appropriate neural architectures to implement mappings $f^{(m)}$. This flexibility allows it to be applied to idiomatic tabular datasets containing images, text, or other modalities.

Experimental Design

Data

We used a collection of 44 real-world tabular datasets from the benchmark introduced in (Grinsztajn, Oyallon, and Varoquaux 2022)[1]. Moreover, to demonstrate applicability of ours and other methods on tables containing images and text, we used 2 additional image-tabular datasets and 2 text-tabular datasets; thus, in total, we used 48 datasets (cf. Table 3). Each dataset includes one categorical (classification), or numerical (regression) target variable (response).

To minimize the effects of data preparation, we restricted ourselves only to minimal data processing involving ordinal encoding of the categorical and rescaling/log-transforming some numerical variables (cf. Supplementary Materials). We

[1]We excluded the largest dataset (delays_zurich_airport) due to computational restrictions.

Name	Samples	Variables			
		Num	Cat	Txt	Img
Classification					
albert	58252	0	32	0	0
bank_marketing	10578	7	1	0	0
bioresponse	3434	419	1	0	0
california	20634	8	1	0	0
clothing	23486	2	5	2	0
compas	4966	3	9	0	0
covertype	566602	10	1	0	0
credit	16714	10	1	0	0
defaults	13272	20	2	0	0
diabetes	71090	7	1	0	0
electricity	38474	7	2	0	0
eye_movements	7608	20	4	0	0
heloc	10000	22	1	0	0
higgs	940160	24	1	0	0
jannis	57580	54	1	0	0
kickstarter	108128	3	4	3	0
miniboone	72998	50	1	0	0
road_safety	111762	14	17	0	0
skin_cancer	13354	1	4	0	1
streetview	11054	2	1	0	1
telescope	13376	10	1	0	0
Regression					
abalone	4177	8	1	0	0
ailerons	13750	34	0	0	0
airlines	1000000	4	2	0	0
allstate	188318	15	110	0	0
bike_sharing	17379	5	2	0	0
brazilian_houses	10692	8	4	0	0
cpu	8192	22	0	0	0
diamonds	53940	8	2	0	0
elevators	16599	17	0	0	0
house	22784	17	0	0	0
house_sales	21613	14	3	0	0
houses	20640	9	0	0	0
medical_charges	163065	4	0	0	0
mercedes	4209	1	359	0	0
miami_housing	13932	14	0	0	0
nyc_taxi	581835	4	13	0	0
pol	15000	27	0	0	0
seattle_crime	52031	1	4	0	0
sgemm	241600	4	6	0	0
soil	8641	4	1	0	0
sulfur	10081	7	0	0	0
superconduct	21263	80	0	0	0
supreme	4052	3	5	0	0
topo	8885	253	3	0	0
ukair	394299	3	4	0	0
wine_quality	6497	12	0	0	0
yprop	8885	43	0	0	0

Table 3: The 48 datasets used in this work and their respective number of samples and the number of variables by type.

randomly split the data into training, validation and testing portions, allocating 10% of the data for validation and another 10% for testing. For each dataset we performed multiple experimental replicates (5 for tabular and 3 for img/text-tabular data), each time using new random split and then averaged the obtained results.

Models and Training

We compared ICE-T against the five state-of-the-art CRL methods described in the previous section: (i) Scarf (ii) SubTab with contrastive and distance loss (iii) CLIP (iv) GMC and (v) MCN with contrastive and clustering loss. Note, since CLIP and MCN scale quadratically with the number of variables, we decided to exclude them from experiments on datasets with more than 25 variables, which would otherwise result in excessive computation runtimes.

For each *method* we trained multiple *models* using different configurations of hyperparameters controlling the model size and its training. Models were trained on the training sets of the data for up to 100 epochs using early stopping with patience set to 5 epochs and validation loss as the criterion.

Evaluation

Once trained, the models were applied to the training, validation and testing portion of the data to produce respective embeddings. We then evaluated the quality of the embedding vectors with respect to three common downstream tasks (i) *imputation* which we approach as cross-modal translation, (ii) *clustering*, and (iii) *supervised learning* – performed across test sets. Finally, we evaluated models in the role of pre-trained encoders, i.e. with respect to their support for (iv) *transfer-learning*. Note, for each method we report the best model performance as achieved in the given task.

Imputation (modality translation) The task aims to estimate values of one variable (query) in the dataset given the value of the remaining variables (evidence). We iterated over test set data columns, permuting and subsequently estimating the values of a single "query" column, by selecting the value whose embedding maximized the similarities to the embeddings of the remaining variables. In the case of ICE-T, the estimate was done as described by Equation 4, whereas for the remaining models, the estimate was made by maximizing sum of the pairwise similarities. The estimates were then compared to the original values to evaluate the quality of recovery (i.e., imputation). We used mean squared error (MSE) to evaluate the quality of imputation of the numerical variables and balanced accuracy score (Brodersen et al. 2010) for the categorical ones. To obtain a unified score across all variables, the results were first scaled to the $[0, 1]$ interval, so that higher values indicate better performance, and then averaged into an *imputation score*. Note that Scarf and SubTab do not support modality translation and were excluded from this evaluation.

Clustering We evaluated how well the obtained embedding vectors support downstream data clustering. For each dataset, a regular k-means clustering model was trained on the training set embeddings. The quality of clusters was evaluated by the *silhouette score* (Shahapure and Nicholas 2020). The hyperparameter k controlling the number of clusters was selected to maximize the silhouette score on the validation set. The test set silhouette score was used as the

evaluation criterion. We also performed k-means clustering using the raw data, instead of embedding vectors, to serve as a shallow benchmark (not applied to txt/img-tabular data).

Supervised learning Similarly to clustering, we evaluated how well the obtained embeddings support downstream supervised learning. For each dataset, we trained a KNN classifier, or regressor, using the training set embeddings and the associated targets. Predictive performance was evaluated by the balanced accuracy score for classification tasks or MSE for regression tasks. Hyperparameter k was selected to maximize predictive performance on validation set. The resulting test set performance was then used as the evaluation criterion. We performed KNN using the raw data as a shallow benchmark (not applied to txt/img-tabular data).

Transfer learning Pre-trained models were used as an encoders, on top of which we added neural prediction head, implemented as a single hidden layer, ReLU-activated MLP, forming a neural predictor. The predictor was trained on the training set for up to 100 epochs using early stopping with patience set to 5 epochs and validation loss as the criterion. Once trained, we evaluated the predictive performance of the resulting model on the test set. The predictions were evaluated by the balanced accuracy score for classification tasks and MSE for regression tasks. As an additional control, we also employed a vanilla MLP predictor with a single hidden layer, i.e., without any pre-trained components. For benchmarking, we used the prediction performance achieved by XGBoost (Chen and Guestrin 2016) with default parametrization (not applied to txt/img-tabular data).

Average relative score The task-specific performance of ICE-T and other methods, as achieved on a given dataset, were scaled to the $[0, 1]$ interval so that the higher value of this *relative score* indicates better performance. To quantify the overall task-specific performance, the resulting values were subsequently averaged across datasets into an *average relative score* (cf. Supplementary Materials).

Results

Synthetic Data Experiment

Consider the dataset illustrated in Figure 2, referred to as Gaussian XOR "blobs", or noisy XOR (Duch 2007). It is evident that the conditional class probability cannot be factorized into a product of conditionals: $p(c|x, y) \neq p(c|x)p(c|y)$. Similar non-factorizability holds for the two coordinates: $p(x|c, y)$ and $p(y|c, x)$. Hence, we hypothesize that learning mappings f by pairwise similarity comparison across the three variables will result in poor embeddings, which will be manifested by a failure to predict any of the three variables from the embeddings of the remaining two.

To validate this hypothesis, we generated training and testing sets consisting of 1,000 and 100 noisy XOR points, respectively; and tested whether the embeddings obtained by ICE-T support the imputation task (modality translation) better than the embeddings from other multimodal contrastive methods. The obtained results show that ICE-T indeed greatly outperforms the other methods, confirming our hypothesis (cf. Table 4).

Method	ACC ↑ c	MSE ↓ x	MSE ↓ y	Imputation score
CLIP	0.750	0.271	0.106	0.562
GMC	0.683	0.206	0.122	0.565
MCN	0.733	0.179	0.148	0.589
ICE-T	**0.900**	**0.008**	**0.006**	**0.982**

Table 4: The performance of ICE-T compared to other multimodal CRL methods in the synthetic data experiment (cf. Figure 2). The embedding vectors produced by ICE-T allow to predict each of the three variables using the remaining two better than those from other methods, demonstrating its ability to capture cross-columnar interactions in tabular data.

	Imputation	Clustering	Supervised learning	Transfer learning
Scarf		**0.674**	0.434	0.474
SubTab		0.344	0.522	0.332
CLIP	0.561	0.345	0.571	0.544
GMC	0.536	0.553	0.495	0.544
MCN	0.378	0.306	0.636	0.533
ICE-T	**0.604**	0.617	**0.687**	**0.824**

Table 5: The average relative score (higher is better, cf. Evaluation) in the four tasks as achieved across the 48 real-world datasets. ICE-T gives the best performance in imputation, supervised and transfer learning; and is second in clustering. Note, Scarf and SubTab do not support modality transfer and could not be evaluated for imputation.

Real-world Data Experiments

For compactness, the results obtained across the 48 real-world datasets were conveyed as win-loss matrices (cf. Figure 3) and as a boxplots depicting the relative scores (cf. Figure 4) (the raw numbers are provided in Supplementary Tables 2–5). The overall performance across the datasets was quantified by average relative score (cf. Table 5). Based on the obtained results we summarize our findings as follows:

- Compared to other methods, embedding vectors produced by ICE-T give better support for imputation and supervised learning, but not clustering, for which Scarf provides a better alternative. This is likely because Scarf, unlike other methods, learns to exert embedding jointly by contrasting entire data instances, which may result in better global alignment of the embedding vectors.
- ICE-T serves in transfer learning better than other methods and transferring pre-trained ICE-T to neural predictor improves performance beyond that of XGBoost (benchmark). Interestingly, in similar experiments performed on independent tabular benchmark, Scarf surpassed XGBoost on approx. 60%, and the MLP used as a negative control on 45% of datasets (Bahri et al. 2021), strongly corroborating our results (cf. Figure 3, right).

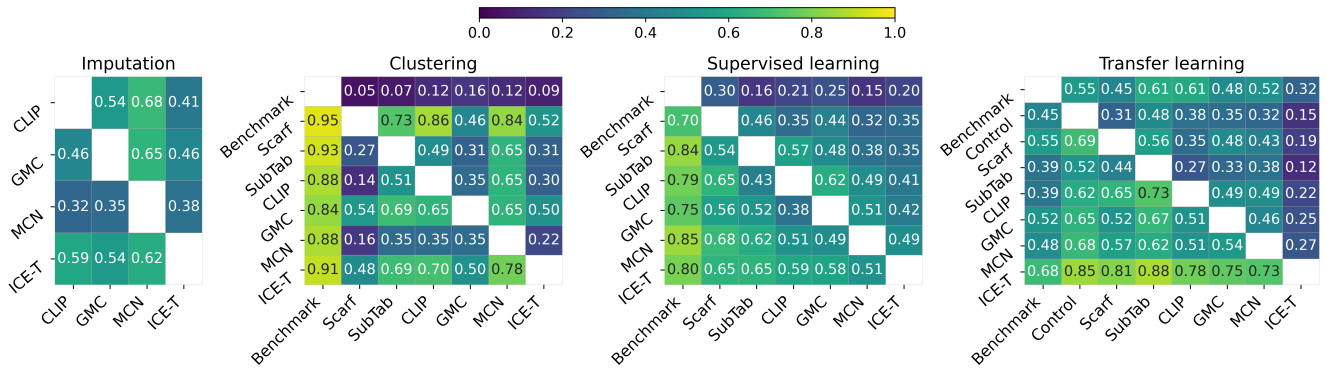

Figure 3: Heatmaps of win-loss matrices, where each value indicates the fraction of datasets where the method in the given row surpassed the method in the given column. Note that ICE-T (bottom row) surpassed other methods on the majority of the datasets (values > 0.5) in support of imputation, supervised and transfer learning; and performed reasonably well in clustering.

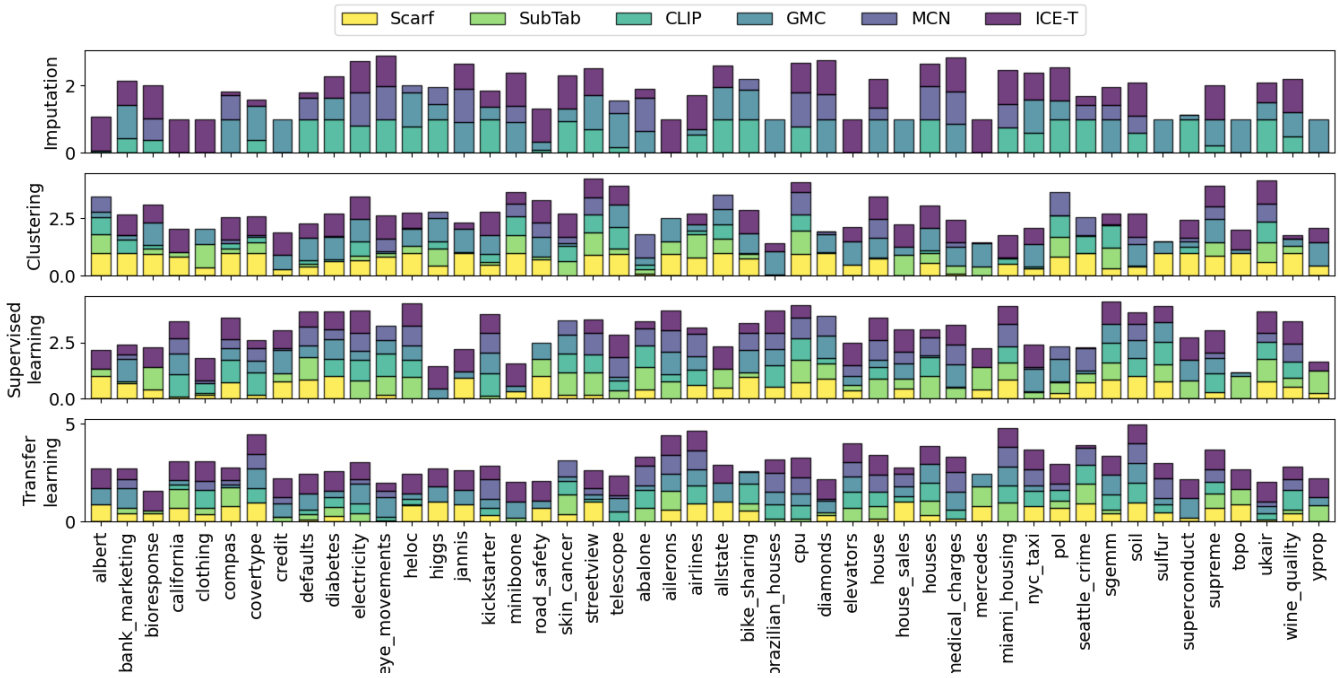

Figure 4: For each task, we scaled the results obtained on the given dataset to $[0, 1]$ interval, so that 1 is assigned to the best performing method, and 0 to the least performing method. The resulting quantities are depicted as segments in a stacked bar plot. A larger segment indicates better performance relative to other methods. As can be seen, ICE-T either outperforms other methods (the largest segment in the given bar), or generally provides good performance, across the majority of the datasets.

Discussion

We propose to approach heterogeneous tabular data as a type of multimodal data, where each variable, i.e., each tabular column, is treated as single modality. Multimodal CRL on tabular data has the potential to provide important advantages over existing methods.

However, unlike multimodal data, most tabular data are characterized by a large number of variables. This significantly penalizes methods using pairwise variable contrasting, which may become prohibitively expensive with growing number of variables. Moreover, in tabular data, interactions among variables are likely to occur.

This motivated ICE-T, whose *key insight is that its loss contrasts each modality with the embedding of a mean of intermediate representation of other modalities*, allowing it to perform in *linear time* and, as we demonstrated, to *capture variable interactions*. Based on our results, we conclude that ICE-T provides better support in most downstream tasks.

Supplementary materials and the code are available at https://github.com/tomastokar/ICET

Acknowledgements

This work was supported by Mitacs Industrial Upskilling Award (IT29286).

References

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33: 25–37.

Bahri, D.; Jiang, H.; Tay, Y.; and Metzler, D. 2021. Scarf: Self-Supervised Contrastive Learning using Random Feature Corruption. In *International Conference on Learning Representations*.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828.

Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; and Kasneci, G. 2022. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Brodersen, K. H.; Ong, C. S.; Stephan, K. E.; and Buhmann, J. M. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, 3121–3124. IEEE.

Chen, B.; Rouditchenko, A.; Duarte, K.; Kuehne, H.; Thomas, S.; Boggust, A.; Panda, R.; Kingsbury, B.; Feris, R.; Harwath, D.; et al. 2021. Multimodal clustering networks for self-supervised learning from unlabeled videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8012–8021.

Chen, T.; and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cohen-Addad, V.; Guedj, B.; Kanade, V.; and Rom, G. 2021. Online k-means clustering. In *International Conference on Artificial Intelligence and Statistics*, 1126–1134. PMLR.

Deldari, S.; Xue, H.; Saeed, A.; He, J.; Smith, D. V.; and Salim, F. D. 2022. Beyond just vision: A review on self-supervised representation learning on multimodal and temporal data. *arXiv preprint arXiv:2206.02353*.

Duch, W. 2007. Learning data structures with inherent complex logic: neurocognitive perspective. *Man-Machine Systems and Cybernetics (CIMMACS'07), Tenerife, Canary Islands, Spain*, 294–303.

Ericsson, L.; Gouk, H.; Loy, C. C.; and Hospedales, T. M. 2022. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3): 42–62.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35: 507–520.

Guo, W.; Wang, J.; and Wang, S. 2019. Deep multimodal representation learning: A survey. *Ieee Access*, 7: 63373–63394.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 976–980. IEEE.

Kaur, P.; Pannu, H. S.; and Malhi, A. K. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39: 100336.

Le-Khac, P. H.; Healy, G.; and Smeaton, A. F. 2020. Contrastive representation learning: A framework and review. *Ieee Access*, 8: 193907–193934.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Poklukar, P.; Vasco, M.; Yin, H.; Melo, F. S.; Paiva, A.; and Kragic, D. 2022. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, 17782–17800. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Shahapure, K. R.; and Nicholas, C. 2020. Cluster quality analysis using silhouette score. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, 747–748. IEEE.

Shwartz-Ziv, R.; and Armon, A. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90.

Ucar, T.; Hajiramezanali, E.; and Edwards, L. 2021. Subtab: Subsetting features of tabular data for self-supervised representation learning. *Advances in Neural Information Processing Systems*, 34: 18853–18865.

Wang, L.; Luc, P.; Recasens, A.; Alayrac, J.-B.; and Oord, A. v. d. 2021. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*.

Zong, Y.; Mac Aodha, O.; and Hospedales, T. 2023. Self-supervised multimodal learning: A survey. *arXiv preprint arXiv:2304.01008*.