AutoML using Metadata Language Embeddings

Iddo Drori, Lu Liu, Yi Nian, Sharath Koorathota, Jie S. Li, Antonio Khalil Moretti, Juliana Freire, Madeleine Udell

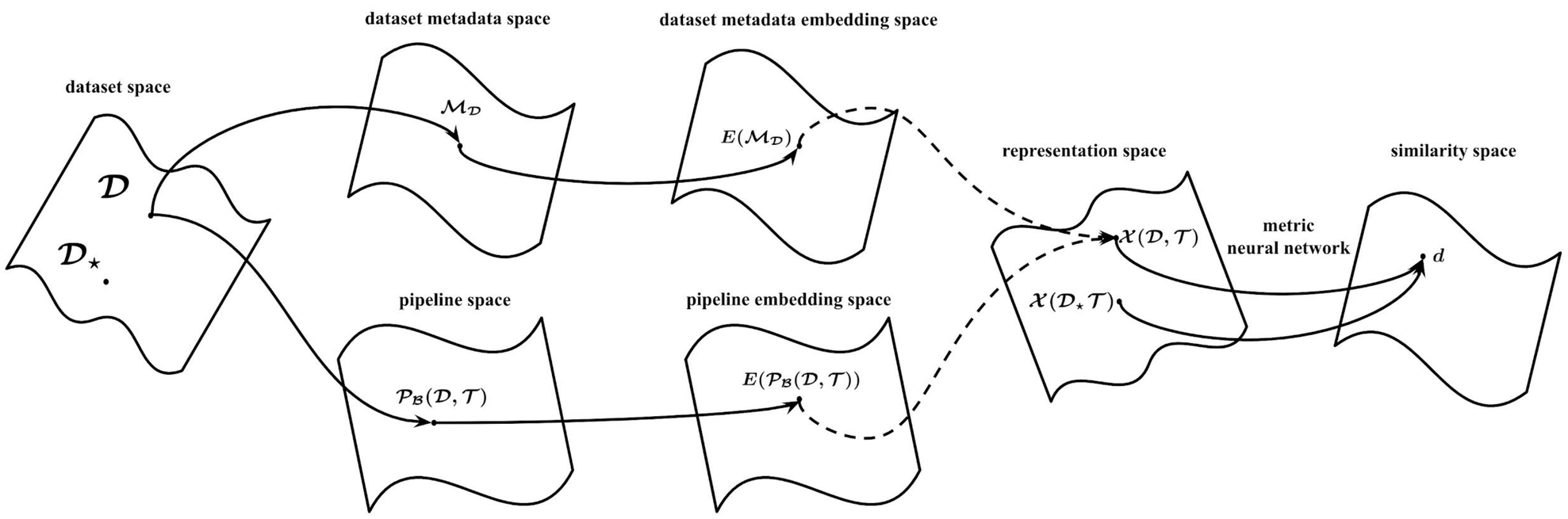
Columbia University in the City of New York · NYU · Cornell University

Neural Information Processing Systems Foundation

Data, Models, and Code: github.com/idrori/automl-embedding

AutoML Embeddings of Dataset Metadata and Pipelines

As a human choosing a supervised learning algorithm, it is natural to begin by reading a text description of the dataset and documentation for the algorithms you might use. We demonstrate that the same idea improves the performance of automated machine learning methods. We use language embeddings from modern NLP to improve state-of-the-art AutoML systems by augmenting their recommendations with vector embeddings of datasets and of algorithms.



Notation	Description
$\mathcal{D}$	Dataset
$\mathcal{M}_\mathcal{D}$	Metadata of dataset $\mathcal{D}$
$\mathcal{T}$	Machine learning task (classification, regression)
$\mathcal{P}$	Solution pipeline
$\mathcal{O},\mathcal{S},\mathcal{A},\mathcal{G},\mathcal{H}$	OBOE, AutoSklearn, AlphaD3M, TPOT, and human algorithm
$\mathcal{P}_\mathcal{B}(\mathcal{D},\mathcal{T})$ for $\mathcal{B}\in\{\mathcal{O},\mathcal{S},\mathcal{A},\mathcal{G},\mathcal{H}\}$	Solution pipeline on dataset $\mathcal{D}$ for task $\mathcal{T}$
$\mathcal{V}(\mathcal{P}_\mathcal{B},\mathcal{D},\mathcal{T})$	Evaluating performance of pipeline $\mathcal{P}_\mathcal{B}$ on $\mathcal{D}$ and $\mathcal{T}$
E	Pre-trained language embedding
$E(\mathcal{M}_\mathcal{D})$	Language embedding of dataset metadata
$d(E(\mathcal{M}_{\mathcal{D}_i}),E(\mathcal{M}_{\mathcal{D}_j}))$	Distance between dataset metadata embeddings
$\mathcal{D}_\star=\underset{\mathcal{D}_i}{\text{argmin}}\|E(\mathcal{M}_D),E(\mathcal{M}_{D_i})\|$	Nearest neighbor of $\mathcal{D}$ under distance of embeddings
$\mathcal{P}_\star=\mathcal{P}(\mathcal{D}_\star,\mathcal{T})$	Pipeline of most similar embedding
$\mathcal{V}(\mathcal{P}_\star,\mathcal{D},\mathcal{T})$	Direct pipeline transfer using dataset metadata embedding
$E(\mathcal{P}_\mathcal{B}(\mathcal{D},\mathcal{T}))$	Language embedding of solution pipeline
$\mathcal{X}(\mathcal{D},\mathcal{T})$	Representation of embeddings for dataset $\mathcal{D}$ and task $\mathcal{T}$
Interaction between embeddings	
$\mathcal{I}(\mathcal{D}_i,\mathcal{D}_j)=(\mathcal{X}(\mathcal{D}_i,T),\mathcal{X}(\mathcal{D}_j,\mathcal{T}))$	Neural network input: pair of representations $\mathcal{X}(\mathcal{D},\mathcal{T})$
$\mathcal{O}(\mathcal{D}_i,\mathcal{D}_j)=d(E(\mathcal{P}_\mathcal{H}(\mathcal{D}_i,\mathcal{T})),E(\mathcal{P}_\mathcal{H}(\mathcal{D}_j,\mathcal{T}))$	Network output: distance between human pipeline embeddings

Results

Dataset	OBOE	AutoSklearn	AlphaD3M	TPOT	Human	**Ours DE**	**Ours PE**
Seattle	**1.00**	0.66	**1.00**	**1.00**	0.92	**1.00**	**1.00**
Insurance	0.47	0.50	0.35	0.51	0.48	0.47	**0.52**
Forest	0.73	0.83	0.83	0.84	0.84	**0.85**	**0.85**
Credit	0.93	**0.94**	0.93	**0.94**	**0.94**	**0.94**	**0.94**
Titanic	0.70	0.80	0.70	0.77	0.86	**0.87**	**0.87**
HR	0.84	0.83	0.87	**0.90**	0.86	0.86	**0.90**
Kobe	0.61	0.60	**0.64**	0.62	0.62	0.61	0.62
Patients	0.64	0.71	**0.72**	0.68	0.68	0.68	0.69

Machine learning pipeline evaluations for AutoML systems and human pipelines. All AutoML pipelines are computed given a **minute** of computation. In comparison, ours DE refers to using only the dataset metadata embedding in under a **second** of computation for zero-shot AutoML. Ours PE refers to using the best single pipeline embedding in a minute of computation.