Multimodal AutoML on Structured Tables with Text Fields

Xingjian Shi* XJSHI@AMAZON.COM
Jonas Mueller* JONASMUE@AMAZON.COM
Nick Erickson NEERICK@AMAZON.COM
Mu Li MLI@AMAZON.COM
Alexander J. Smola ALEX@SMOLA.ORG
Amazon Web Services, CA, USA

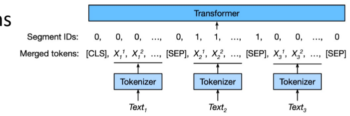
Code: https://github.com/awslabs/autogluon AutoGluon

Tutorial: https://auto.gluon.ai/stable/tutorials/tabular_prediction/tabular-multimodal-text-others.html

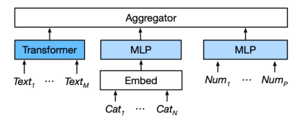
```
from autogluon.tabular import TabularPredictor
predictor = TabularPredictor(label='class').fit('train.csv', presets='best_quality', hyperparameters='multimodal')
predictions = predictor.predict('test.csv')
```

Multi-Modal Transformer Network

- Handling multiple text columns



- Multi-tower Network Architecture



- Easy to fit/deploy

```
from autogluon.text import TextPredictor
predictor = TextPredictor('label').fit(train_data)
```

Multimodal Benchmark Results

Method	prod	qaq	qaa	cloth	airbnb	ae	mercari	jigsaw	imdb	fake	kick	jc	wine	news	channel	avg. ↑	mrr ↑
Choosing Text-Net:	NLP Backbones and Finetuning Tricks																
RoBERTa	0.588	0.412	0.268	0.700	0.344	0.953	0.561	0.960	0.731	0.929	0.751	0.615	0.811	-0.000	0.301	0.595	0.07
ELECTRA	0.705	0.410	0.356	0.718	0.349	0.955	0.586	0.965	0.750	0.824	0.754	0.606	0.813	0.003	0.315	0.607	0.17
+ Exponential Decay $\tau = 0.8$	0.728	0.436	0.431	0.743	0.337	0.953	0.579	0.963	0.852	0.963	0.760	0.664	0.808	0.004	0.308	0.635	0.09
+ Average 3 ★	0.729	0.451	0.432	0.746	0.350	0.954	0.581	0.965	0.858	0.961	0.766	0.656	0.807	0.004	0.307	0.638	0.12
Choosing Multimodal-Net:	Fusion Strategy																
All-Text	0.907	0.454	0.419	0.746	0.366	0.957	0.599	**0.967**	0.840	0.967	**0.799**	0.645	0.810	0.013	0.480	0.665	0.19
Fuse-Early	**0.913**	0.441	0.418	0.745	0.377	0.953	0.596	**0.967**	0.843	0.960	0.770	0.653	0.806	0.013	0.474	0.662	0.24
Fuse-Late, Concat ★	0.907	0.449	**0.445**	0.747	0.395	0.958	0.603	0.966	0.857	0.961	0.773	0.639	0.812	0.015	0.481	0.667	0.17
Fuse-Late, Mean	0.912	**0.458**	0.431	0.748	0.399	0.955	0.602	**0.967**	0.869	0.963	0.773	0.625	0.807	0.015	0.478	0.667	0.09
Fuse-Late, Max	0.910	0.452	0.429	0.747	0.401	0.956	0.599	0.966	0.863	0.957	0.761	0.634	0.808	0.015	0.484	0.665	0.12
Choosing Aggregation:	Multimodal Model Ensembling																
Pre-Embedding	0.895	0.216	0.247	0.642	0.449	0.972	0.433	0.586	0.871	0.926	0.743	0.491	0.680	0.012	0.526	0.579	0.13
Text-Embedding	0.867	0.446	0.432	0.748	0.430	0.972	0.434	0.587	0.855	0.962	0.790	0.658	0.830	0.008	0.502	0.635	0.20
Multimodal-Embedding	0.907	0.439	0.437	0.749	0.438	0.974	0.432	0.587	0.847	0.967	0.794	**0.683**	0.829	0.007	0.517	0.640	0.18
Weighted-Ensemble	0.907	0.439	0.429	0.744	0.453	0.976	0.597	0.957	0.876	0.923	0.787	0.641	0.814	0.018	0.554	0.674	0.39
Stack-Ensemble ★	0.909	0.456	0.438	**0.751**	0.459	0.977	**0.605**	**0.967**	**0.878**	0.964	0.797	0.624	0.836	**0.020**	**0.556**	**0.683**	**0.59**
	Tabular AutoML + Feature Engineering Baselines																
AG-Weighted	0.891	0.046	0.076	-0.002	0.426	0.841	0.098	0.587	0.845	0.686	0.668	0.004	0.173	0.016	0.549	0.394	0.11
AG-Stack	0.891	0.046	0.077	0.001	0.435	0.841	0.098	0.587	0.844	0.697	0.670	0.003	0.175	0.017	0.550	0.395	0.10
AG-Weighted+ N-Gram	0.892	0.426	0.382	0.610	0.450	0.978	0.526	0.909	0.842	0.966	0.772	0.357	0.829	0.019	0.546	0.633	0.11
AG-Stack+ N-Gram	0.895	0.414	0.383	0.654	**0.466**	**0.979**	0.569	0.915	0.850	**0.968**	0.775	0.612	**0.842**	**0.020**	0.548	0.659	0.19
H2O AutoML	0.869	0.247	0.159	0.163	0.329	0.976	0.430	0.531	0.813	0.756	0.669	0.411	0.478	0.014	0.530	0.492	0.11
H2O AutoML + Word2Vec	0.859	0.244	0.285	0.624	0.347	0.973	0.534	0.847	0.827	0.943	0.755	0.443	0.778	0.013	0.524	0.600	0.16
H2O AutoML + Pre-Embedding	0.846	0.227	0.312	0.644	0.367	0.969	0.282	0.572	0.874	0.893	0.738	0.549	0.571	0.007	0.501	0.557	0.12

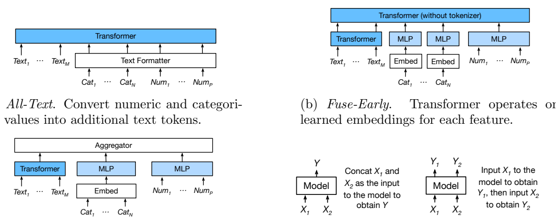
Table 3: Predictive performance of AutoML strategies over our multimodal benchmark. Column 'avg.' lists each method's average score (across datasets) and 'mrr' lists the mean reciprocal rank among all models evaluated in the benchmark. Each subsection encapsulates the variants compared at a design stage, with the final choice (best avg.) marked by ★.

Multimodal Data (Numeric & Categorical & Text)

name	desc	goal	country	currency	created_at	final_status
The Secret Order - The Game that gives back Gl...	Can you trust your friends? Solve the puzzle? ...	5000.0	GB	GBP	1424101105	0
Booker Family Foods. Home made, the way food s...	Community based, home-made-foods producer, to ...	2500.0	US	USD	1404617242	0
J.A.E.S.A : Next Generation Artificial Intelli...	A true next generation AI with the ability to ...	30000.0	CA	CAD	1399078600	1

Table 1: Example of data in our multimodal benchmark with text (*name, desc*), numeric (*goal, created_at*), and categorical (*country, currency*) columns. From these features, we predict if a Kickstarter project will be funded (*final_status*).

Multi-Modal Network: Options



(a) *All-Text.* Convert numeric and categorical values into additional text tokens.

(b) *Fuse-Early.* Transformer operates on learned embeddings for each feature.

(c) *Fuse-Late.* Separate branches encode each modality, aggregate via mean/max/concat.

(d) Notation used in these figures.

Figure 1: Fusion strategies in Multimodal-Net, dense output layers on top are not shown.

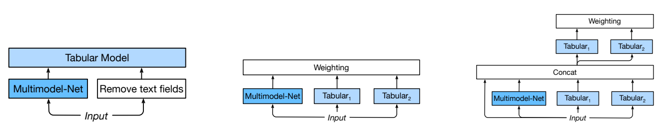
Interesting Findings

- Neural embedding of text followed by tabular modeling is often outperformed by: N-gram featurization or leveraging text neural nets for their *predictions* (stack ensembling) not *representations* (embeddings)
- In multimodal networks, fusing modalities in *early* layers (Transformers with cross-modality attention) is **not** necessarily superior to older multi-tower architectures that fuse representations in *late* layers
- End-to-end multimodal neural net is improved by *stack ensembling* this network with tabular models trained in separate stages (not end-to-end)

Modeling Multimodal Data with Text Fields

- Fit Text Neural Network on text features (Transformer)
- Fit classical Tabular Models on numeric+categorical features (GBDT, random forest, etc.)
- Option 1: Ensemble Tabular & Text Models
- Option 2: Fit Tabular Models after featurizing text into vector form (N-gram, word2vec, or Transformer embedding)
- Option 3: Adapt Text Network to additionally operate on numeric + categorical features

Aggregating Text & Tabular Models: Options



(a) *Embedding-as-Feature* (b) *Weighted-Ensemble* (c) *Stack-Ensemble*

Figure 2: Methods to combine *Multimodal-Net* and classical tabular models.

Tabular 1, 2, ... = Tabular Models (eg. Boosted Tree, Random Forest, etc.)

Rank in Tabular+Text ML Competition Leaderboards

- 1st in "MachineHack: Predict The Data Scientists Salary In India"
- 1st in "MachineHack: Product Sentiment Classification"
- 2nd in "MachineHack: Predict The Price Of Books"
- 2nd in "Kaggle: California House Prices"
- 2nd in "Kaggle: Mercari Price Suggestion"
  - 2380 teams with $100,000 prize money