Multimodal AutoML on Structured Tables with Text Fields

Xingjian Shi†
Jonas Mueller†
Nick Erickson
Mu Li
Alexander J. Smola
Amazon Web Services, CA, USA

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

Multimodal Data (Numeric & Categorical & Text)

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).

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 featureizing 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.

Multi-Modal Network: Options

(a) All-Text. Convert numeric and categorical values into additional text tokens.
(b) Fix-Early. Transformer operates on learned embeddings for each feature.
(c) Fix-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.

Multi-Modal Transformer Network

• Handling multiple text columns
• Multi-tower Network Architecture
• Easy to fit/deploy

Multimodal Benchmark Results

Interesting Findings

• Neural embedding of text followed by tabular modeling is often outperformed by: N-gram featureization 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)

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