Multimodal Attribute Extraction

Robert L. Logan IV (UC Irvine) **Samuel Humeau** (Diffbot) **Sameer Singh** (UC Irvine)

Abstract

- Existing information extraction methods focus on extracting knowledge from text, however knowledge is also expressed through modes such as tables and images.
- We introduce the first benchmark evaluation and dataset for the task of **multi-modal attribute extraction** (MAE).

Problem Statement:

Given a collection of items I such that for each $i \in I$ there is a textual description T_i, a set of images P_i and a set of attribute-value pairs $\left\{\left(a_i^{(j)}, v_i^{(j)}\right)\right\}$, the task is to predict the value $v_i^{(j)}$ given attribute $a_i^{(j)}$.

Evaluation Metric:

Let $\hat{v}_{i,k}^{(j)}$ denote the top k predictions for value $v_i^{(j)}$ then:

$$Hits@k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(v_i^{(j)} \in \hat{v}_{i,k}^{(j)}\right)$$

MAE Dataset

The MAE dataset is composed of images, descriptions, and attribute-value tables for products collected using the Diffbot Product API.

The MAE dataset can be freely downloaded at: https://rloganiv.github.io/mae/

# Products	2.2 m
# Images	4.0 m
# Attribute-Value Pairs	7.6 m
# Unique Attributes	2.1 k
# Unique Values	23.6 k

MAE Dataset - Example

Title: Gray Vinyl Barstool

Image:

Attribute-Value Table:

Color Finish	Gray
Frame Material	Metal
Assembly Required	Yes

Text:

This sleek dual purpose stool easily adjusts from counter to bar height. The overall design is casual and contemporary which allow it to seamlessly accent any area in the home…

Example Model Outputs:

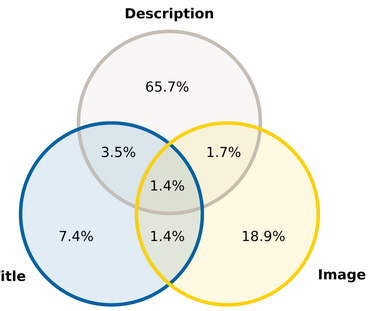
Most Common Value	White	Black	Steel	Chrome	Gray
	-	-	-	-	-
Image Baseline	Gray	Silver	Grey	White	Beige
	0.84	0.63	0.60	0.60	0.58
Text Baseline	White	Black	Blue	Gray	Brown
	0.81	0.70	0.63	0.62	0.59
Multimodal - Concat	Gray	Red	Green	Grey	Blue
	0.84	0.71	0.71	0.71	0.70
Multimodal - GMU	Gray	Blue	Brown	Green	Red
	0.85	0.71	0.69	0.68	0.67

Mechanical Turk Study

- Determine how often attributes can be inferred.
- Measure the usefulness of the different modes of information.
- Collect a 'gold' set of evaluation data.

Results:

- 46% of queries answerable.
- 85% human accuracy.
- 2,238 'gold' attribute-value pairs.



Proposed Model

Encoder/Decoder:

- Attributes / Values: Linear Embeddings
- Descriptions: Convolution on Word Embeddings
- Images: Inception-v3[1]

Fusion:

- Concatenation
- Gated Multimodal Unit[2]

Objective:

- Contrastive Loss Function

$$d(v, c) = \frac{v \cdot h}{|v||c|}$$

$$\mathcal{L} = max(0, 1 - d^2(v^+, c) + d^2(v^-, c))$$

Benchmarks

Performance of baseline models on top-100 attributes.

Model	Hits@1	Hits@5	Hits@10	Hits@20
Most Common Value	38.81 %	77.26 %	87.96 %	95.96 %
Image Baseline	38.07 %	76.11 %	86.99 %	95.00 %
Text Baseline	58.41 %	**87.49 %**	**93.94 %**	**98.00 %**
Multimodal - Concat	**59.48 %**	87.33 %	93.23 %	97.07 %
Multimodal - GMU	52.92 %	85.07 %	92.23 %	97.26 %

References

1. Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.
2. Arevalo, John, et al. "Gated Multimodal Units for Information Fusion." *arXiv preprint arXiv:1702.01992* (2017).

For any questions, email: **rlogan@uci.edu**