Multimodal Attribute Extraction

Robert L. Logan IV
University of California
Irvine, CA
rlogan@uci.edu

Samuel Humeau
Diffbot
Mountain View, CA
sam@diffbot.com

Sameer Singh
University of California
Irvine, CA
sameer@uci.edu

1 Introduction

Given the large collections of unstructured and semi-structured data available on the web, there is a crucial need to enable quick and efficient access to the knowledge content within them. Traditionally, the field of information extraction has focused on extracting such knowledge from unstructured *text* documents, such as job postings, scientific papers, news articles, and emails. However, the content on the web increasingly contains more varied types of data, including semi-structured web pages, tables that do not adhere to any schema, photographs, videos, and audio. Given a query by a user, the appropriate information may appear in any of these different modes, and thus there's a crucial need for methods to construct knowledge bases from different types of data, and more importantly, combine the evidence in order to extract the correct answer.

Motivated by this goal, we introduce the task of *multimodal attribute extraction*. Provided contextual information about an entity, in the form of any of the modes described above, along with an attribute query, the goal is to extract the corresponding value for that attribute. While attribute extraction on the domain of text has been well-studied [4, 7, 16, 18, 20], to our knowledge this is the first time attribute extraction using a combination of multiple modes of data has been considered. This introduces additional challenges to the problem, since a multimodal attribute extractor needs to be able to return values provided any kind of evidence, whereas modern attribute extractors treat attribute extraction as a tagging problem and thus only work when attributes occur as a substring of text.

In order to support research on this task, we release the Multimodal Attribute Extraction (MAE) dataset[1], a large dataset containing mixed-media data for over 2.2 million commercial product items, collected from a large number of e-commerce sites using the Diffbot Product API.[2] The collection of items is diverse and includes categories such as electronic products, jewelry, clothing, vehicles, and real estate. For each item, we provide a textual product description, collection of images, and open-schema table of attribute-value pairs (see Figure 1 for an example). The provided attribute-value pairs only provide a very weak source of supervision; where the value might appear in the context is not known, and further, it is not even guaranteed that the value can be extracted from the provided evidence. In all, there are over 4 million images and 7.6 million attribute-value pairs. By releasing such a large dataset, we hope to drive progress on this task similar to how the Penn Treebank [14], SQuAD [19], and Imagenet [6] have driven progress on syntactic parsing, question answering, and object recognition, respectively.

To asses the difficulty of the task and the dataset, we first conduct a human evaluation study using Mechanical Turk that demonstrates that all available modes of information are useful for detecting values. We also train and provide results for a variety of machine learning models on the dataset. We observe that a simple *most-common value classifier*, which always predicts the most-common value for a given attribute, provides a very difficult baseline for more complicated models to beat (33% accuracy). In our current experiments, we are unable to train an image-only classifier that can outperform this simple model, despite using modern neural architectures such as VGG-16[21] and Google's Inception-v3[22]. However, we are able to obtain significantly better performance using a

---
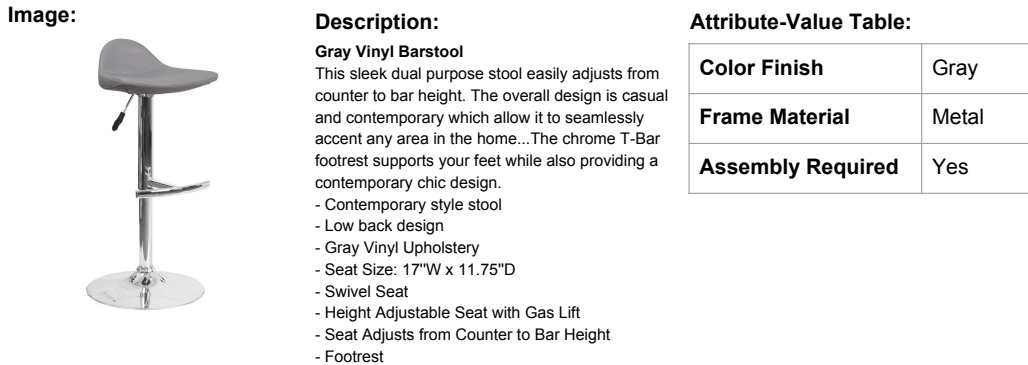
[1]The dataset is freely available at: https://rloganiv.github.io/mae.
[2]https://www.diffbot.com/products/automatic/product/

Image:

Description:

Gray Vinyl Barstool
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...The chrome T-Bar footrest supports your feet while also providing a contemporary chic design.
- Contemporary style stool
- Low back design
- Gray Vinyl Upholstery
- Seat Size: 17"W x 11.75"D
- Swivel Seat
- Height Adjustable Seat with Gas Lift
- Seat Adjusts from Counter to Bar Height
- Footrest

Attribute-Value Table:

Color Finish	Gray
Frame Material	Metal
Assembly Required	Yes

Figure 1: An example item with its descriptions: image, tabular attributes, and textual.

Table 1: MAE dataset statistics.

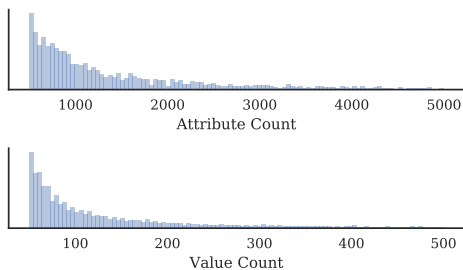|	
# products	2.2 m
# images	4.0 m
# attribute-value pairs	7.6 m
# unique attributes	2.1 k
# unique values	23.6 k

Figure 2: Histograms of attribute and value counts.

text-only classifier (59% accuracy). We hope to improve and obtain more accurate models in further research.

2 Multimodal Product Attribute Extraction

Since a multimodal attribute extractor needs to be able to return values for attributes which occur in images as well as text, we cannot treat the problem as a labeling problem as is done in the existing approaches to attribute extraction. We instead define the problem as following: Given a product i and a query attribute a, we need to extract a corresponding value v from the evidence provided for i, namely, a textual description of it (D_i) and a collection of images (I_i). For example, in Figure 1, we observe the image and the description of a product, and examples of some attributes and values of interest. For training, for a set of product items $\mathcal{I}$, we are given, for each item $i \in \mathcal{I}$, its textual description D_i and the images I_i, and a set A_i comprised of attribute-value pairs (i.e. $A_i = \{\langle a_i^j, v_i^j \rangle\}_j$). In general, the products at query time will not be in $\mathcal{I}$, and we do not assume any fixed ontology for products, attributes, or values. We evaluate the performance on this task as the accuracy of the predicted value with the observed value, however since there may be multiple *correct* values, we also include hits@k evaluation.

The MAE Dataset The MAE dataset is composed of mixed media data for 2.2 million product items, obtained by running the Diffbot Product API on over 20 million web pages from 1068 different commercial websites. As in the task definition, there is a textual description, set of product images, and open-schema table of product attributes for every item. The Diffbot API obtains this information using a machine learning based extractor which uses visual, textual and layout features of the fully rendered product webpage. For example, attribute-value pairs are automatically extracted from tables present on product webpages. Due to the automated nature of this collection process, there is some noise present in the dataset. For instance, the same attribute may be represented many different ways (e.g. *Length*, *length*, *len.*). We use regular-expression based preprocessing to normalize the most common attributes, however, we leave values unnormalized. We also remove any attribute-value pairs
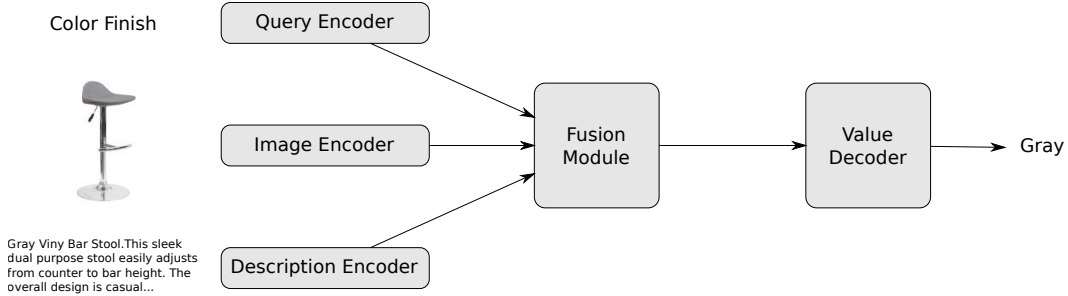
2

Figure 3: Basic architecture of the multimodal attribute extraction model.

that satisfy any of the following frequency conditions: the attribute occurs less than 500 times, the value occurs less than 50 times, or the attribute's most common value makes up more than 80% of the attribute-value pairs. The data is split into a training, validation, and test set using an 80-10-10 split.

Mechanical Turk Evaluation Since the attributes and values have been extracted as they appear on the web sites, there is no guarantee that the attribute-value pairs appear in either the product images or textual descriptions. We perform a study using Amazon Mechanical Turk to determine the extent to which this issue affects the dataset, as well as collect a *gold* evaluation dataset of attribute-value pairs that are guaranteed to show up in the context information. Mechanical Turk workers are presented a product's images and textual description, and asked to determine whether they can predict the value for a given product attribute (from a list of choices) using the provided information, and if so, using which pieces of information. We use a majority vote to eliminate noise in these annotations. The (preliminary) results of this study suggest that only 42% of the attribute-value pairs can be found using contextual information. Of those, 35% could be found using the product's image and 70% could be found using the textual description. This suggests that while textual descriptions are the most useful mode for attribute extraction, there is still beneficial information contained in images.

3 Multimodal Fusion Model

In this section, we formulate a novel extraction model for the task that builds upon the architectures used recently in tasks such as image captioning, question answering, VQA, etc. The model is composed of three separate modules: (1) an encoding module that uses modern neural architectures to jointly embed the query, text, and images into a common latent space, (2) a fusion module that combines these embedded vectors using an attribute-specific attention mechanism to a single dense vector, and (3) a similarity-based value decoder which produces the final value prediction. We provide an overview of this architecture in Figure 3.

Encoding Module We assign a dense embedding for each attribute and values, i.e. attribute a is represented by a k-dimensional vector $\mathbf{c}_a$, and value v by $\mathbf{c}_v$, where the vectors are learned during training. For textual description D, we first tokenize the text using the Stanford tokenizer [13], followed by embedding all of the words using the Glove algorithm [17] on all of the descriptions in the training data. We use the CNN architecture of Kim [10], that consists of CNN layers, max-pooling, and a fully-connected layer, to combine these pretrained embeddings to a single dense vector for the description, $\mathbf{c}_D$. Embeddings of the images I are also produced using convolutional neural networks. Specifically, we obtain intermediate image representations using the output of the fc7 layer (after applying the ReLU non-linearity) of a pretrained 16-layer VGG model [21]. We then feed the output through a fully connected layer to obtain a k-dimensional embedding for each image. The final embedding $\mathbf{c}_I$ is produced by performing max-pooling over the image embeddings.

Fusion To fuse attribute embeddings $\mathbf{c}_a$ with the text and image embeddings, $\mathbf{c}_D$ and $\mathbf{c}_I$, we experiment with two different techniques. The first, called **Concat**, is to concatenate the three of them and then feed them through a fully-connected layer, in order to produce the fused encoding $\mathbf{c}$. The second approach, called **GMU** for gated multimodal unit [3], first fuses the attribute vector $\mathbf{c}_a$ with $\mathbf{c}_I$ and $\mathbf{c}_D$ independently using fully-connected layers, resulting in $\mathbf{c}_I^a$ and $\mathbf{c}_D^a$. We combine

Table 2: Baseline model results.

	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 Baseline - Concat	59.48	87.33	93.23	97.07
Multimodal Baseline - GMU	52.92	85.07	92.23	97.26

them by first creating gating vector $z = \sigma(\mathbf{W}_z[\mathbf{c}_D^a, \mathbf{c}_I^a])$, followed by, $\mathbf{c} = z \cdot \mathbf{c}_D^a + (1 - z) \cdot \mathbf{c}_I^a$. For *unimodal* baselines, the fusion module is replaced by a fully-connected layer.

Loss Function We use a variant of the contrastive loss function introduced by Chopra et al. [5]. Let $\mathbf{c}$ denote the embedding produced by the fusion layer. Our goal is to produce an embedding which is close to the value embedding $\mathbf{c}_v^+$ (e.g. the one from the training example), and distant from other value embeddings $\mathbf{c}^-$. In order to measure closeness we use cosine similarity, denoted by g, followed by a variant of squared hinge loss:

$$\mathcal{L}^+(\mathbf{c}, \mathbf{c}_v^+) = (1 - g(\mathbf{c}, \mathbf{c}_v^+))^2; \quad \mathcal{L}^-(\mathbf{c}, \mathbf{c}_v^-) = \begin{cases} g(\mathbf{c}, \mathbf{c}_v^-)^2 & \text{if } g(\mathbf{c}, \mathbf{c}_v^-) \geq 0 \\ 0 & \text{else} \end{cases}$$

$$\mathcal{L}(\mathbf{c}, \mathbf{c}_v^+, \mathbf{c}_v^-) = \mathcal{L}^+(\mathbf{c}, \mathbf{c}_v^+) + \mathcal{L}^-(\mathbf{c}, \mathbf{c}_v^-)$$

(1)

where a negative value is sampled for each training example from the empirical distribution of value counts displayed in Figure 2. To obtain a value prediction given context, we identify the value with embedding $\mathbf{c}_v$ closest to the context embedding $\mathbf{c}$, according to cosine similarity g.

4   Experiments

We evaluate on a subset of the MAE dataset consisting of the 100 most common attributes, covering roughly 50% of the examples in the overall MAE dataset. To determine the relative effectiveness of the different modes of information, we train image and text only versions of the model described above. Following the suggestions in Zhang and Wallace [25] we use a 600 unit single layer in our text convolutions, and a 5 word window size. We apply dropout to the output of both the image and text CNNs before feeding the output through fully connected layers to obtain the image and text embeddings. Employing a coarse grid search, we found models performed best using a large embedding dimension of $k = 1024$. Lastly, we explore multimodal models using both the *Concat* and the *GMU* strategies. To evaluate models we use the hits@k metric on the values.

The results of our experiments are summarized in Table 2. We include a simple most-common value model that always predicts the most-common value for a given attribute. Observe that the performance of the image baseline model is almost identical to the most-common value model. Similarly, the performance of the multimodal models is similar to the text baseline model. Thus our models so far have been unable to effectively incorporate information from the image data. These results show that the task is sufficiently challenging that even a complex neural model cannot solve the task, and thus is a ripe area for future research.

Model predictions for the example shown in Figure 1 are given in Table 3, along with their similarity scores. Observe that the predictions made by the current image baseline model are almost identical to the most-common value model. This suggests that our current image baseline model is essentially ignoring all of the image related information and instead learning to predict common values.

5   Related Work

Our work is related to, and builds upon, a number of existing approaches.

The introduction of large curated datasets has driven progress in many fields of machine learning. Notable examples include: The Penn Treebank [14] for syntactic parsing models, Imagenet [6] for object recognition, Flickr30k [24] and MS COCO [12] for image captioning, SQuAD [19] for

Table 3: Top 5 predictions on the data in Figure 1 when querying for *color finish*.

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

question answering and VQA [2] for visual question answering. Despite the interest in related tasks, there is currently no publicly available dataset for attribute extraction, let alone multimodal attribute extraction. This creates a high barrier to entry as anyone interested in attribute extraction must go through the expensive and time-consuming process of acquiring a dataset. Furthermore, there is no way to compare the effectiveness of different techniques. Our dataset aims to address this concern.

Recently, there has been renewed interest in multimodal machine learning problems. Vinyals et al. [23] demonstrate an effective image captioning system that uses a CNN to encode an image which is used as the input to an LSTM [8] decoder, producing the output caption. This encoder-decoder architecture forms the basis for successful approaches to other multimodal problems such as visual question answering [1]. Another body of work focuses on the problem of unifying information from different modes of information. Kiela and Bottou [9] propose to concatenate together the output of a text-based distributional model (such as word2vec [15]) with an encoding produced from a CNN applied to images of the word. Lazaridou et al. [11] demonstrate an alternative approach to concatenation, where instead the a word embedding is learned that minimizes a joint loss function involving context-prediction and image reconstruction losses. Another alternative to concatenation is the gated multimodal unit (GMU) proposed in [3]. We investigate the performance of different techniques for combining image and text data for product attribute extraction in section 4.

To our knowledge, we are the first to study the problem of attribute extraction from multimodal data. However the problem of attribute extraction from text is well studied. Ghani et al. [7] treat attribute extraction of retail products as a form of named entity recognition. They predefine a list of attributes to extract and train a Naïve Bayes model on a manually labeled seed dataset to extract the corresponding values. Putthividhya and Hu [18] build on this work by bootstrapping to expand the seed list, and evaluate more complicated models such as HMMs, MaxEnt, SVMs, and CRFs. To mitigate the introduction noisy labels when using semi-supervised techniques, More [16] incorporates crowdsourcing to manually accept or reject the newly introduced labels. One major drawback of these approaches is that they require manually labelled seed data to construct the knowledge base of attribute-value pairs, which can be quite expensive for a large number of attributes. Bing et al. [4] address this problem by using an unsupervised, LDA-based approach to generate word classes from reviews, followed by aligning them to the product description. Shinzato and Sekine [20] propose to extract attribute-value pairs from structured data on product pages, such as HTML tables, and lists, to construct the KB. This is essentially the approach used to construct the knowledge base of attribute-value pairs used in our work, which is automatically performed by Diffbot's Product API.

6 Conclusions and Future Work

In order to kick start research on multimodal information extraction problems, we introduce the multimodal attribute extraction dataset, an attribute extraction dataset derived from a large number of e-commerce websites. MAE features images, textual descriptions, and attribute-value pairs for a diverse set of products. Preliminary data from an Amazon Mechanical Turk study demonstrates that both modes of information are beneficial to attribute extraction. We measure the performance of a collection of baseline models, and observe that reasonably high accuracy can be obtained using only text. However, we are unable to train off-the-shelf methods to effectively leverage image data.

There are a number of exciting avenues for future research. We are interested in performing a more comprehensive crowdsourcing study to identify the ways in which different evidence forms are useful, and in order to create *clean* evaluation data. As this dataset brings up interesting challenges in multimodal machine learning, we will explore a variety of novel architectures that are able to combine the different forms of evidence effectively to accurately extract the attribute values. Finally, we are also interested in exploring other aspects of knowledge base construction that may benefit from multimodal reasoning, such as relational prediction, entity linking, and disambiguation.

Acknowledgments

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and vqa. *arXiv preprint arXiv:1707.07998*, 2017.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[3] J. Arevalo, T. Solorio, M. Montes-y Gómez, and F. A. González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[4] L. Bing, T.-L. Wong, and W. Lam. Unsupervised extraction of popular product attributes from web sites. In *Asia Information Retrieval Symposium*, pages 437–446. Springer, 2012.

[5] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[7] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano. Text mining for product attribute extraction. *SIGKDD Explor. Newsl.*, 8(1):41–48, June 2006. ISSN 1931-0145. doi: 10.1145/1147234. 1147241. URL http://doi.acm.org/10.1145/1147234.1147241.

[8] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[9] D. Kiela and L. Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Empirical Methods for Natural Language Processing (EMNLP)*, 2014.

[10] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[11] A. Lazaridou, N. T. Pham, and M. Baroni. Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*, 2015.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[13] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

[14] M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.

[15] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[16] A. More. Attribute extraction from product titles in ecommerce. *arXiv preprint arXiv:1608.04670*, 2016.

[17] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

[18] D. P. Putthividhya and J. Hu. Bootstrapped named entity recognition for product attribute extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567. Association for Computational Linguistics, 2011.

[19] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[20] K. Shinzato and S. Sekine. Unsupervised extraction of attributes and their values from product description. In *International Joint Conference on Natural Language Processing (IJCNLP)*, 2013.

[21] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[23] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[24] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[25] Y. Zhang and B. Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.