"A Distorted Skull Lies in the Bottom Center..." Identifying Paintings from Text Descriptions

Anupam Guha,[1] **Mohit Iyyer,**[1] **Jordan Boyd-Graber**[2]
[1]University of Maryland, Department of Computer Science and UMIACS
[2]University of Colorado, Department of Computer Science
aguha@cs.umd.edu, miyyer@umiacs.umd.edu,
Jordan.Boyd.Graber@colorado.edu

Abstract

Most question answering systems use symbolic or text information. We present a dataset for a task that requires understanding descriptions of visual themes and their layout: identifying paintings from their descriptions. We annotate paintings with contour data, align regions with entity mentions from an ontology, and associate image regions with text spans from descriptions. A simple embedding-based method applied to text-to-image coreferences achieves state-of-the-art results on our task when paired with bipartite matching. The task is made all the more difficult by scarcity of training data.

1 Knowledge from Images

Question answering is a standard NLP task that typically requires gathering information from knowledge sources such as raw text, ontologies, and databases. Recently, vision and language have been amalgamated into an exciting and difficult task: using images to ask or answer questions.

While humans can easily answer complex questions using knowledge gleaned from images, visual question answering (VQA) is difficult for computers. Humans excel at this task because they abstract key concepts away from the minutiae of visual representations, but computers often fail to synthesise prior knowledge with confusing visual representations.

We present a new instance of visual question answering: can a computer identify an artistic work given only a textual description? Our dataset contains images of paintings, tapestries, and sculptures covering centuries of artistic movements from dozens of countries. Since these images are of cultural importance, we have access to many redundant descriptions of the same works, allowing us to create a naturalistic but inexpensive dataset. Due to the complex and oblique nature of questions about paintings, their visual complexity, and the relatively small data size, prior approaches used for VQA over natural images are infeasible for our task.

We formalise the task in Section 3, where we also present a preliminary system (ARTMATCH) and compare with it a data-driven text baseline to illustrate the usefulness and versatility of our method (Section 4). Finally, in Section 5 we compare our task and system to previous work that combines NLP and vision.

2 Describing Art

University Challenge (UK) or quiz bowl (USA) has previously been studied for question answering using text-based methods (Boyd-Graber et al., 2012). However, some quiz bowl questions are inherently visual in that their answers are works of art.

Figure 1 shows an example of a painting description and associated annotations (to be described later) from a quiz bowl question. Identifying paintings from textual descriptions of their contents is difficult; for example, many disparate paintings feature two men (*Stag at Sharky's*, *The Sacrifice of Isaac*, and *Kindred Spirits*). Given their varied style, composition, and depiction, how do we teach computers to infer the *meaning* of a painting?

To capture the meaning, we rely on redundant descriptions of entities in paintings offered by multiple text spans in these questions. The man on the right in Figure 1 is variously described as a "Frenchman", "a diplomat", "a man in black", and "George de Selve".

Proceedings of 2016 NAACL Human-Computer Question Answering Workshop, pages 43–47,
San Diego, California, June 12-17, 2016.

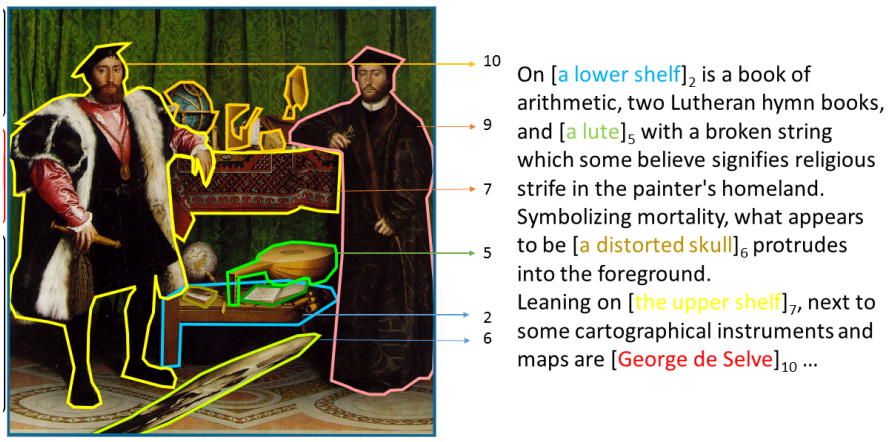

Figure 1: A painting (left) with image regions matched to coreference chains in a question (right). The question uses a variety of oblique mentions to make the trivia question more difficult; some entities (e.g., de Selve) are mentioned again later in the question.

We can use this redundancy to learn the meaning of the pixels within the red contour.

In text, this is the problem of coreference resolution (Radford, 2004) as the multiple text spans refer to the same "real world" entity. Trivia questions have complex descriptive coreferent groupings (Guha et al., 2015). Thus, to annotate our dataset we map coreference groups in question text to regions in paintings using LabelMe (Russell et al., 2008), providing a direct mapping of text spans to groups of pixels in the images and their spatial properties.

Our dataset contains 128 paintings,[1] where each painting is the answer to a single quiz bowl question. First, we assign each object in a painting to a single class from an ontology with eight coarse and fifty two fine (level two) classes. This ontology is three levels deep and follows the hyponymy structure of ImageNet (Deng et al., 2009).[2] Then, we map each coreference group from the question text to an image contour from the painting (see Table 1). As the questions come from a game, the mentions are often oblique, making them hard to answer with text alone. For instance, a description of *Rain, Steam, and Speed* will avoid explicitly mentioning the painting's central object by name (a "locomotive") in favor of describing it in a roundabout way (e.g., a "conveyance").

[1] Annotated data and code available after blind review.

[2] Ontology provided as supplementary material.

Number of ...	dataset
Unique Paintings	128
Objects with contours	1,436
Coreferring text groups	1,104
Object gross classes	8
Object fine classes	52

Table 1: Statistics of our new question answering dataset

3 Identifying Paintings from Text

Given one of the questions in our dataset, our goal is to provide the name of the painting that it describes. Because our focus is not on building better feature extractors for paintings, we assume that we have gold visual annotations (e.g., object contours, classes, and locations).[3] The task is challenging due to the size of our dataset (only 128 annotated question/painting pairs), which prevents the training of most machine learning models, as well as high visual complexity and vagueness in question text.

3.1 A Text-Only Baseline

Our baseline model is "blind" in that it does not use any visual features to solve the task. We use the deep averaging network (Iyyer et al., 2015, DAN),

[3] When applied to paintings, state-of-the-art semantic segmentation models (Zheng et al., 2015) trained on natural images are only reliable for four coarse object classes (vs 60 coarse and fine ones), which leads to near-random accuracies for ARTMATCH.

which takes as input a textual description of a painting and learns a 128-label classifier over an average of embeddings from words in the question. Since this model does not do any visual mapping, we collect unannotated questions about our 128 paintings to form a respectably-sized training set of 503 questions. While the DAN has access to more data than our non-blind model, we hope that we can improve over the baseline using visual information.

3.2 Answer Questions Using Annotated Paintings

Our method, which we call ARTMATCH, assumes that some of the groups of coreferent text in a question describe visual objects in the associated painting. If we have a unified vector representation of visual object classes from painting regions and textual coreferent groups, a bipartite mapping can match them.

3.2.1 Matching Mentions to Images

To identify the painting described by a question, we first convert every question to a list of objects obtained from coreference chains (e.g., *a lute*, *a distorted skull*). On the painting side, we have a list of annotated visual objects. These two lists form the nodes of a bipartite graph on which we perform a maximum cardinality match (Hopcroft and Karp, 1973), where edge weights represent match strength. We consider the painting with the most matched edges as the answer; in case of a tie, the painting with the highest cumulative edge weight wins.

This process requires that our visual object classes are in the same vector space as objects found in textual coreference chains. For one chain, we compute a vector representation by averaging the embeddings of its words.[4] Also, for each visual object class, we obtain a set of synonyms and hyponyms and compute an averaged word vector over this set. Similarly, we produce averaged vectors over location and number attributes (e.g., the *single* attribute is represented by a vector average over {*single, one, a, an*}). Since distance between word embeddings measures semantic similarity, we assign an object class and attributes that have the highest cosine similarity to that chain's vector representation, as shown in Figure 2.

[4]We use publicly available 300-dimensional word2vec embeddings trained on Google News (Mikolov et al., 2013).

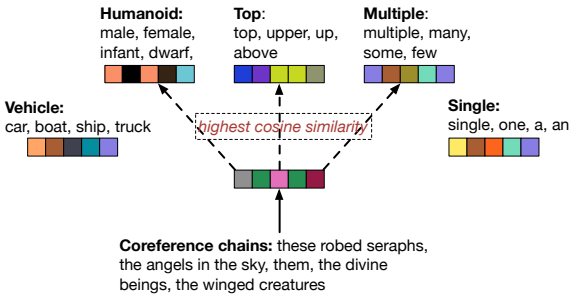

Figure 2: Using word2vec representations from coreferent groupings in a description to deduce object class and attributes by cosine similarity.

We easily combine ARTMATCH's matching with DAN by modifying the weight of a bipartite match by multiplying that weight with the probability of that question-answer pair being correct as given by DAN.

4 Performance and Analysis

We investigate the performance both locally (matching specific objects) and globally (identifying the correct image) before doing an error analysis.

First, we examine the **individual matching on objects** and measure accuracy on three different tasks:

1. Does ARTMATCH properly match coarse and fine visual *object classes* to question text (e.g., is "an angel" mentioned in the question matched to an image region depicting an angel)?[5]
2. Are the matches is in the correct *locations* (e.g., is the "angel" in the top-left corner?)?
3. Is the *number* of objects correctly matched (e.g., are there two or three angels)?

Table 2 shows the results of these experiments. Additionally, for the highest-scoring paintings (the answers output by ARTMATCH), 13.2% of objects are exactly matched with location and number; without considering those attributes, 20.4% of fine-grained classes are matched.

Next, we look at the main task of **identifying the correct painting**. As Table 3 shows, ARTMATCH nearly matches the blind DAN baseline with just the coarse and fine object classes. Spatial location and number attributes boost ARTMATCH above the baseline, and combining both systems pushes accuracy

[5]Coarse class metrics are provided for all objects while fine class performance is only for those that have fine annotations.

Feature	P	R	F1
Coarse object class	0.72	0.38	0.45
Fine object class	0.72	0.60	0.60
Object location	0.32	0.25	0.24
Object number	0.96	0.81	0.88

Table 2: Individual metrics of classes and features detected by word embeddings from coreference chains describing objects

Method	accuracy
DAN	59.4%
ARTMATCH: fine objects	42.0%
ARTMATCH: all objects	58.6%
ARTMATCH: objects+attributes	61.7%
ARTMATCH+DAN	**65.7%**

Table 3: Our system vs the blind baseline. DAN is trained on 503 questions but has no visual information. ARTMATCH has visual features from paintings but no training data. Combining both leads to a significant increase in performance.

by four absolute points, indicating that the models are learning complementary information.

Having established that our simple method of incorporating visual information can achieve significant gains in accuracy, we now proceed to analyse instances in which our system does well and the DAN does not (and vice-versa).

4.1 Error Analysis

There are 34 questions for which the DAN fails but ARTMATCH succeeds. For many of these questions, the DAN fails because it overfits to common clues. Given a test question about *Melencholia I*, the DAN answers *Madonna with the Long Neck*, as the training questions about both paintings repeatedly mention a female figure and cherubs. However, the question also mentions geometric figures, the spatial locations of which enable ARTMATCH to answer correctly.

Conversely, there are thirty-one questions where ARTMATCH fails but DAN succeeds. Some of these questions contain text constructs such as the painter's name that are repeated in both training and test questions, which makes it easy for the DAN to solve (e.g., "Identify this most famous work of Claude Monet"). In other cases, ARTMATCH answers incorrectly because of spurious matches due to substantial visual similarity between various objects in paintings. For example, in a question about *The Holy Trinity* by Masaccio, "St. John" is assigned the close but incorrect class of "statue" while "Jesus" is correctly identified as a person. Further confused with spatial similarities between the paintings, ARTMATCH's answer is *Supper at Emmaus*, which has Jesus but no St. John. In other cases, peripheral similarity leads to the central mismatch being overlooked, motivating an an attention mechanism to focus on "significant" entities for future work (Mnih et al., 2014).

5 Related Work

Our work is specifically related to previous work on visual question answering and more generally to multimodal applications of vision and language.

Visual QA has previously focused on content questions (Antol et al., 2015; Ren et al., 2015; Andreas et al., 2015), while we focus on identity questions. Relatedly, Zhu et al. (2015) find semantic links between images and text via an attention model.

We use coreference to connect text and image regions, similar to Kong et al. (2014). However, not all text is "visual" (Dodge et al., 2012) and not all image regions can be described textually (Berg et al., 2012). While we focus on meaning, structure of text (Elsner et al., 2014) can also be inferred from images. Socher et al. (2014) match sentences to images; however, our dataset is unique in that the text is intentionally oblique (rather than direct descriptions) and our images—paintings—are more varied visually.

Aside from QA, images have been successfully used to generate captions (Karpathy and Fei-Fei, 2014; Mao et al., 2014; Vinyals et al., 2014; Xu et al., 2015; Chen and Zitnick, 2014). While we use vision to aid an NLP task, others have gone in the opposite direction, inducing correspondences between words and video clips (Yu and Siskind, 2013), words and action models (Ramanathan et al., 2013), and language and perception (Matuszek et al., 2012).

6 Conclusion and Future Work

The major contribution of this work is to extend question answering to a complex visual setting by presenting an annotated dataset and a simple system that manages to exceed the performance of a strong text-only baseline QA system. The next challenge is to scale up this dataset to enable end-to-end training pipelines for answering questions using raw images.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2015. Deep compositional question answering with neural module networks. *arXiv preprint arXiv:1511.02799*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *International Conference on Computer Vision*.

Alexander C Berg, Tamara L Berg, Hal Daume III, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, et al. 2012. Understanding and predicting importance in images. In *Computer Vision and Pattern Recognition*.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Empirical Methods in Natural Language Processing*.

Xinlei Chen and C Lawrence Zitnick. 2014. Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*.

Jia Deng, Wei Dong, Richard Socher, Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Computer Vision and Pattern Recognition*.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C Berg, et al. 2012. Detecting visual text. In *Proceedings of the Association for Computational Linguistics*.

Micha Elsner, Hannah Rohde, and Alasdair DF Clarke. 2014. Information structure prediction for visual-world referring expressions. *Proceedings of the European Chapter of the Association for Computational Linguistics*.

Anupam Guha, Mohit Iyyer, Danny Bouman, and Jordan Boyd-Graber. 2015. Removing the training wheels: A coreference dataset that entertains humans and challenges computers. In *North American Association for Computational Linguistics*.

John E Hopcroft and Richard M Karp. 1973. An nˆ5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on computing*, 2(4):225–231.

Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep unordered composition rivals syntactic methods for text classification. In *Association for Computational Linguistics*.

Andrej Karpathy and Li Fei-Fei. 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Chen Kong, Dahua Lin, Mayank Bansal, Raquel Urtasun, and Sanja Fidler. 2014. What are you talking about? text-to-image coreference. In *Computer Vision and Pattern Recognition*.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Proceedings of Advances in Neural Information Processing Systems*.

Andrew Radford. 2004. *English syntax: An introduction*. Cambridge University Press.

Vignesh Ramanathan, Percy Liang, and Li Fei-Fei. 2013. Video event understanding using natural language descriptions. In *International Conference on Computer Vision*.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Proceedings of Advances in Neural Information Processing Systems*.

Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. 2008. Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *Proceedings of the Association for Computational Linguistics*.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. 2015. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2015. Visual7w: Grounded question answering in images. *arXiv preprint arXiv:1511.03416*.