Lu, H., Liu, Q., Ichien, N., Yuille, A. L., & Holyoak, K. J. (2019). Seeing the meaning: Vision meets semantics in solving pictorial analogy problems. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Seeing the Meaning: Vision Meets Semantics in Solving Pictorial Analogy Problems

Hongjing Lu[1,2]
hongjing@ucla.edu

Qing Liu[3]
qingliu@jhu.edu

Nicholas Ichien[1]
ichien@ucla.edu

Alan L. Yuille[3]
alan.yuille@jhu.edu

Keith J. Holyoak[1]
holyoak@lifesci.ucla.edu

[1] Department of Psychology, [2] Department of Statistics
University of California, Los Angeles, Los Angeles, CA 90095 USA
[3] Department of Cognitive Science
Johns Hopkins University, Baltimore, MD 21218 USA

Abstract

We report a first effort to model the solution of meaningful four-term visual analogies, by combining a machine-vision model (ResNet50-A) that can classify pixel-level images into object categories, with a cognitive model (BART) that takes semantic representations of words as input and identifies semantic relations instantiated by a word pair. Each model achieves above-chance performance in selecting the best analogical option from a set of four. However, combining the visual and the semantic models increases analogical performance above the level achieved by either model alone. The contribution of vision to reasoning thus may extend beyond simply generating verbal representations from images. These findings provide a proof of concept that a comprehensive model can solve semantically-rich analogies from pixel-level inputs.

Keywords: analogy; relations; learning; machine vision; word embeddings

Introduction

In everyday life, humans continually perceive the world and interpret it in terms of meaningful objects and events. The representations extracted by perception are elaborated into semantic representations that can be communicated by language and further transformed by reasoning processes. The "holy grail" of cognitive science is to develop integrated theories that link perception to language and higher cognition. A natural testbed for developing such integrated theories is the task of reasoning by analogy from meaningful visual inputs. Here we report a first effort to develop a comprehensive model of the solution of visual analogies, by combining a model that can translate pixel-level inputs into verbal captions with a model that can translate semantic vectors for words into coherent patterns of semantic relations.

Figure 1 depicts an example of the analogies on which we focus. This problem is one of a set of 18 developed by Krawzcyk et al. (2008), some of which were adapted from an earlier set created by Goranson (2002), hence dubbed the Goranson Analogy Test (GAT). The upper row presents a pictorial problem in the form *A:B :: C:?*. The task is to select the best analogical completion from among a set of four options shown in the bottom row. For this example, the analogical solution based on matching relations is to choose the pie (wine is made from grapes, as pie is made from pumpkin). The three distractors include one that is semantically related to the *C* term but fails to match the *A:B* relation (witch), one that is visually similar but also fails to match *A:B* (basketball), and one that is simply unrelated (books). Critically, the analogical solution cannot in any obvious way be derived from visual information alone, because the core relation is semantic/functional rather than visual. For example, the fact that wine is made from grapes is not depicted in the visual input; rather, it must be retrieved from semantic memory. Thus, vision is necessary but not sufficient to reliably solve such semantically-rich picture analogies.

The GAT was originally developed as a tool to evaluate the impact of neuropsychological disorders. Krawczyk et al. (2008) found that frontal and temporal patients were impaired to varying degrees, notably showing an elevated tendency to choose the semantic or perceptual distractors. Age-matched controls (approximately age 60) achieved about 98% accuracy even in the presence of similar distractors.

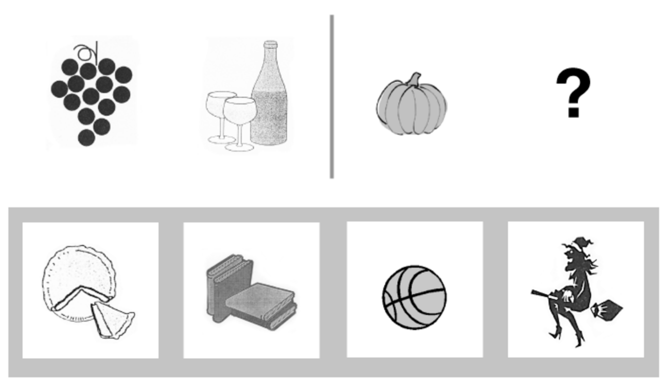

Figure 1. Example of a 4-term pictorial analogy with four alternatives (from Krawzcyk et al., 2008).

Here we focus on the most fundamental question: how can such pictorial analogy problems be solved at all? On the face of it, the process begins with the human visual system operating on pixel-level inputs of the images in the problem to extract a verbal description and/or semantic categorization of the objects. Reasoning processes must use these object descriptions to determine the relation(s) linking paired objects. Based on these relational representations, the reasoner must then assess the degree of relational match between *A:B* and the alternative completions for *C*, finally choosing the option that provides the best match.

Despite decades of progress in developing computational models of visual perception, language processing, and analogical reasoning, no model has tackled the full range of processing required to solve meaningful visual analogies such as the GAT problems. Recent advances in machine vision have led to very significant progress in the recognition of objects from pixel-level representations (Krizhevsky, Sutskever & Hinton, 2012; Semonvan & Zisserman, 2015), including the automatic generation of verbal captions (Farhadi et. al., 2010; Mao et. al., 2016; Krishna et. al., 2016). However, artificial-intelligence (AI) models have been less successful in transforming visual inputs into semantic representations of *relations between objects*. AI models of visual analogy have generally focused on problems that can be solved on the basis of simple visual features, such as color and shape (Reed et. al., 2015; Sadeghi, Zitnick & Farhadi, 2015). In cognitive science, most analogy models have simply assumed high-level representations of complex propositions (usually hand-coded), without dealing with the problem of how these representations could be generated by perceptual processes. Lovett and Forbus (2017) describe a model that applies analogical reasoning to solve Ravens Progressive Matrices problems, which are a form of visual analogies based on transformations of geometrical shapes. However, the inputs provided to the model are high-level perceptual descriptions, rather than a matrix of pixels; and the Ravens test is entirely formal, devoid of any links to semantic knowledge. With important exceptions (e.g., Doumas, Hummel, & Sandhofer, 2008), analogy models have generally set aside the basic problem of how semantic relations could be learned from non-relational inputs.

Here we describe two computational models that together provide an approximate account of the entire process that may underlie solution of GAT problems. One model, ResNet50-A, aims to solve the picture analogies using purely visual information, while also generating verbal captions. The other, BART, aims to solve the same analogies based solely on verbal descriptions of the images. We further show that the analogy assessment derived by ResNet50-A using just visual information not only provides potential verbal inputs to BART, but also adds independent visual information that increases solution accuracy. We will first describe the operation of each of the two models, and then the results obtained by using them both separately and jointly.

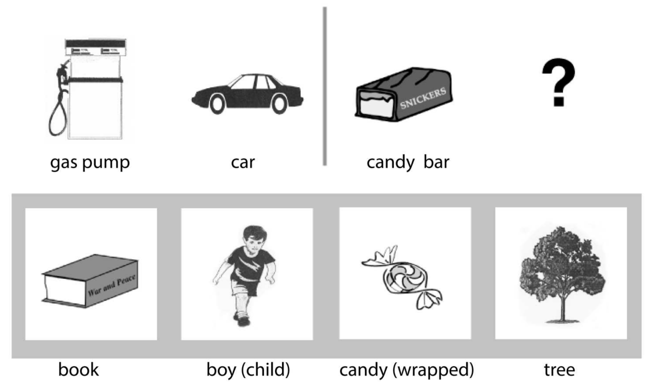

Figure 2. Example of a 4-term pictorial analogy with four alternatives, and corresponding descriptions verbally presented to patients (from Krawzcyk et al., 2008).

GAT Dataset

The GAT dataset includes 18 picture analogies, each consisted of 7 images: the three images in the question, *A*, *B*, and *C*, and the four images for alternative D terms. All images are line drawings or clip art images. Each image was captured in the size of 140x140 pixels. The GAT dataset included a total of 126 images that fall into 118 distinct object categories. A verbal caption describing each image was used by Krawczyk et al. (2008) in their neuropsychological study; these captions were adopted as canonical verbal descriptions of each image for the semantic model, BART. Figure 2 shows a second example, along with approximations of the corresponding verbal descriptions used by Krawczyk et al. (2008). Note that in the neuropsychological study, the accompanying labels were presented orally by the experimenter, rather than in written form.

ResNet: From Pixels to Object Classification

Background

Deep convolutional neural networks (Krizhevsky, Sutskever, & Hinton, 2012; Simonyan & Zisserman, 2015) have led to a series of breakthroughs for a broad range of computer vision tasks. The network depth is of crucial importance. Recent work with deeper networks has exposed a degradation problem: as network depth increases, accuracy reaches a plateau, and then degrades rapidly as network depth increases further. ResNet (He, Zhang, Ren, & Sun, 2016) addresses the degradation problem by introducing a framework termed *deep residual learning*. ResNet fits a residual mapping, realized by a feedforward neural network with identity shortcut connections. Using this method, ResNet can be efficiently trained with as many as 1000 layers. Because of its compelling performance levels, ResNet has quickly emerged as one of the leading architectures for a wide range of tasks in computer vision. Here we adopt ResNet50 (the basic architecture with 50 layers) as a state-of-the-art approach to identifying and captioning the objects in GAT analogies. We then augment the model to create ResNet50-A (where the “A” stands for “Analogy”) by adding a decision procedure to generate potential analogical solutions based solely on visual information in the images.

Training Dataset

The GAT images are line drawings (as are most images used in picture analogy tests that have been developed for psychological research or cognitive assessments). Machine vision models are typically trained on photo-realistic images, and require additional training with line drawings in order to classify them. In order to provide suitable training for ResNet50, we created a database of clip art images that were similar to GAT images, but not identical to them. This dataset, termed the ClipArt dataset, includes the 118 object categories used in the GAT visual analogy problems. To create the ClipArt dataset, we queried Google Image Search using the "Search by image" function, uploading the corresponding GAT image and entering a phrase formed by concatenating the category label and the words "clip art". (For some categories, we visually checked the result and decided to replace "clip art" by "drawing", "sketch", or "cartoon".) We downloaded 200 images for each category and manually removed those that were duplicates or clearly wrong. Each category in the resulting ClipArt dataset was represented by 70-166 images. The images were then processed into gray scale and padded with zero on short edges to fit a 1:1 aspect ratio.

For each category, we randomly selected 50 images for training, and held the rest images for test, resulting in a total of 5900 training images and 5501 test images. Figure 3 juxtaposes a GAT image (left) with a ClipArt image (right) from the same category. To ensure that the model was able to generalize its visual recognition performance, the GAT dataset was only used to guide construction of the ClipArt dataset; the GAT images themselves were not used to train ResNet50.

Training

We implemented ResNet50 using Pytorch on a single TitanX GPU. The training task was image classification by minimizing the cross-entropy loss. The model was pretrained on the ImageNet dataset, and then fine-tuned on our ClipArt dataset for 200 epochs. Batch size was set equal to 120 and learning rate started at 0.01, followed by cosine annealing. For optimization, SGD optimizer was used with momentum = 0.9, weight decay = 0.0001. To prevent overfitting, small random image transformations (e.g., rotation, translation, scaling) were added to the input images. The model achieved a high performance level on the ClipArt test set, achieving 0.883 for top-1 accuracy (i.e., the correct object category label being identified as the first choice of the model), and 0.973 for top-5 accuracy (i.e., the correct object category label being identified as one of the top five choices of the model). When tested on the GAT images for the visual analogy problems, the model achieved 0.833 for top-1 accuracy and 0.984 for top-5 accuracy.

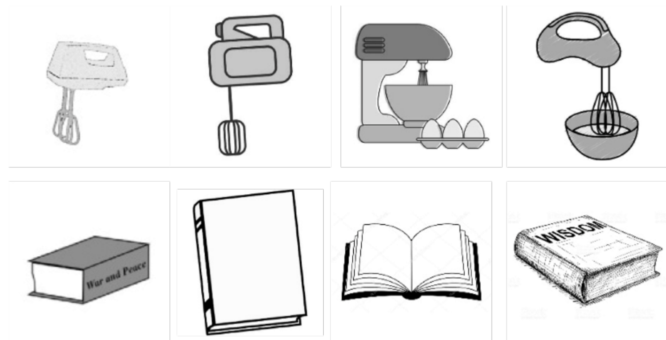

Figure 3. Example images. In each row, the first image is from the GAT dataset, while the remaining images are from the ClipArt dataset. Top row: images with label "electric mixer"; bottom row: images with label "book".

Analogical Inference

We extended ResNet50 to form ResNet50-A by adding a simple computation to derive analogy predictions from the model. We input each GAT image into the neural network and extracted the penultimate feature vector (the vector immediately prior to the output layer). This vector of length 2048 was used as the representation of the image. Mathematically, this transformation can be written as: $\boldsymbol{f} = F(\boldsymbol{I};\theta)$, where $\boldsymbol{I}$ is the input image, F is the function specified by the neural network and parametrized by θ, and $\boldsymbol{f} \in R^{2048}$ is the resulting feature vector. Thus, for each analogy question, we transfer images $\boldsymbol{I}_A$, $\boldsymbol{I}_B$, $\boldsymbol{I}_C$, $\boldsymbol{I}_{D1}$, $\boldsymbol{I}_{D2}$, $\boldsymbol{I}_{D3}$, $\boldsymbol{I}_{D4}$ into feature vectors $\boldsymbol{f}_A$, $\boldsymbol{f}_B$, $\boldsymbol{f}_C$, $\boldsymbol{f}_{D1}$, $\boldsymbol{f}_{D2}$, $\boldsymbol{f}_{D3}$, $\boldsymbol{f}_{D4}$, respectively.

A decision for an analogy problem in ResNet50-A is derived by selecting the best $D \in \{D_1, D_2, D_3, D_4\}$ such that the relation from *A* to *B* holds for *C* to *D*. To measure how similar the projection from $\boldsymbol{f}_A$ to $\boldsymbol{f}_B$ is to the projection from $\boldsymbol{f}_C$ to $\boldsymbol{f}_D$, we adopted a generic formulation based on cosine distances of the difference vectors. The same approach has been used in the Word2vec model (Zhila et al., 2013). The preferred answer $\widehat{D}$ is defined as the *D* image that generates minimum cosine distance between difference vectors:

$$\widehat{D} = \arg\min_{D \in \{D_1, D_2, D_3, D_4\}} \cos\left(\boldsymbol{f}_B - \boldsymbol{f}_A, \boldsymbol{f}_D - \boldsymbol{f}_C\right)$$

Note that this procedure for solving a visual analogy is more sophisticated than simply choosing the $\widehat{D}$ most similar to *C*, since the selection focuses on matching the visual *relation* between the *A:B* and *C:D* image pairs.

For the GAT problems, this purely visual model achieved 44% accuracy in selecting the correct *D* term. Its other choices were distributed across the three distractors (11%, 17% and 28% probabilities of choosing semantic distractors, visual distractors, and unrelated distractors, respectively). Since chance accuracy would be 25%, the purely visual analogy model achieved analogical accuracy well above chance (although well short of the level achieved by neurotypical human adults).

BART: From Verbal Semantics to Relations

The BART model (*Bayesian Analogy with Relational Transformations*) takes as inputs semantic vectors representing word meanings and uses supervised learning to acquire representations of semantic relations. The model was originally applied to learning comparatives (e.g., *larger*, *smarter*; Lu, Chen & Holyoak, 2012), but has recently been generalized to acquire an extremely wide range of semantic relations (e.g., *synonym*, *antonym*, *cause-effect*; Lu, Wu & Holyoak, 2019). For the present project, the inputs to the BART model were word embedding for individual words, each embedding consisting of 300-dimension vectors with continuous-valued features. The word embeddings were obtained by training a deep-learning model, Word2vec (Mikolov et al., 2013; Le & Mikolov, 2014) on a large text corpus (Google News). BART takes as inputs word pairs instantiating a relation, where each pair is represented by the concatenation of the Word2vec vector for each individual word. For example, a vector formed by concatenating the individual vectors for *love* and *hate* would constitute a positive example of the *antonymy* relation. The same word pair might also serve as a negative example of the *category:instance* relation.

Training Dataset

For the present project, we trained BART by combining two datasets of semantic relations. First, the SemEval-2012 Task 2 dataset (Jurgens et al., 2012) was used to teach BART the representations for 79 abstract semantic relations. This dataset is based on a taxonomy of semantic relations and includes 10 general types (e.g., *class inclusion*, *similar*, *contrast*, *cause-purpose*). The dataset includes 3215 word pairs, with 35~48 pairs for each of the 79 relations. The second dataset, developed by Popov, Hristova, and Royce (2017), includes some specific and concrete relations (e.g., the relation *constitution* with examples *brick:house*, *thread:cloth*; the relation *cover* with examples such as *house:roof*; and the relation *boundary* with examples such as *wall:room*). This dataset includes 58 specific relations drawn from ten general categories of relations. Two relations with inadequate numbers of examples were removed. The remaining 56 relations included 12~25 word pairs as examples for each relation.

Training

The BART model consists of a three-stage process to learn a broad range of semantic relations (Lu, Wu & Holyoak, 2019). In its first stage, BART exploits the heuristic that features playing similar functional roles will tend to occupy similar ranks in an ordering of differences between paired words. BART uses the difference ranking operations to generate augmented feature by partially align important features. In the second stage, BART selects a subset of important features. In the third stage, BART adopts Bayesian learning and uses the selected features of word pairs $\mathbf{f}_s$ in training examples to estimate weights distributions $\mathbf{w}$ for representing a particular relation R by applying Bayes rule as:

$$P(\mathbf{w}|\mathbf{f}_s, R) \propto P(R|\mathbf{f}_s, \mathbf{w})P(\mathbf{w}). \qquad (1)$$

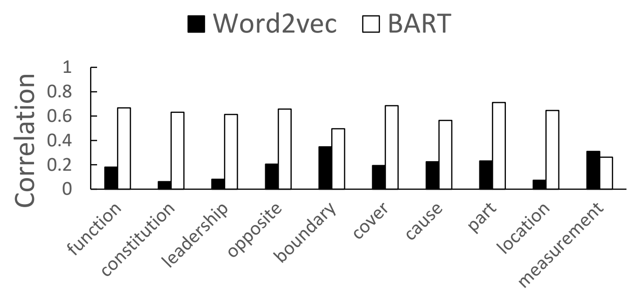

Figure 4. Model predictions of human data for relation typicality in Popov et al. (2017) dataset: Correlations between human generation frequencies and model predictions for 10 relation types for BART (after training with 10 positive examples of each relation) and for the baseline Word2vec model.

After learning, BART calculates the probability of a word pair instantiating a relation. An important aspect of both the Jurgens et al. (2012) and the Popov et al. (2017) norms is that in each set, the word pairs instantiating each relation form a typicality ordering established by human judgments. As reported in Lu et. al. (2019), BART achieved high rank-order correlations between human typicality ratings and predicted probabilities derived from the model for the abstract relations in the Jurgens et al. dataset. Across all 79 individual relations, the model's mean Spearman correlation with the human ordering was .81 (range from .65 to .91). The performance of BART considerably exceeded the mean correlation of .34 achieved using Word2vec itself as a baseline.

For the Popov et al. (2017) dataset, which includes more specific/concrete relations, BART was trained with just 10 word pairs as positive examples of each relation. As shown in Figure 4, BART achieved higher correlations with human typicality as indexed by generation frequencies (mean $r = .59$) than did the Word2vec model (mean $r = .19$).

Analogical Inference

To solve 4-term verbal analogy problems, BART forms a distributed representation of the specific relation between each word pair in a problem. BART uses its pool of learned relations to create a more refined representation of the relation(s) between two paired words. The posterior probabilities calculated for all known relations form a relation vector, with each element indicating how likely a word pair instantiates a specific relation. Hence, the result of this operation is to create a distributed representation of the relation(s) between two words, with the original semantic features being projected into a transformed space that can be used to assess relation probabilities.

For analogical reasoning, BART had available 79 relations derived by training on the Jurgens et al. (2012) norms, plus 56 relations derived by training on the Popov et al. (2017) norms.

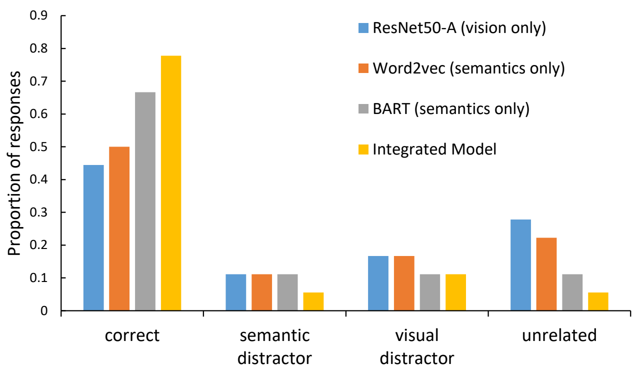

Figure 5. Proportion of responses for GAT problems for which the model's selection was the analogical option (correct), the semantic distractor, the visual distractor, and the unrelated option. Besides ResNet50-A and BART, we also report results obtained using Word2vec, and the integrated model (i.e., ResNet50-A combined with BART).

Of the latter, six relations showed weak correlations with human typicality ratings, indicating BART had failed to learn them adequately from the small number of available examples. Further examinations of the training sets for these six relations revealed that a substantial number of word pairs either included ambiguities or were otherwise questionable as instances of the relation. Accordingly, these six relations were dropped, leaving 50 relations from the Popov et al. set to be included in the relational representations, for a total of 129 learned relations.

Because BART creates relations structured by distinct roles, the model can generate the converse of any learned relation in a rule-based fashion (without additional training). For example, having learned the relation *category:instance*, BART can directly generate the converse relation *instance:category*. By applying converse formation to all trained relations, BART doubled its pool of relations, so that a total set of 258 semantic relations were available to solve GAT analogy problems.

To apply the BART model to GAT problems, the input was the verbal captions for images provided in the study by Krawczyk et al. (2008). Considered as a comprehensive model, this makes the link between ResNet50 and BART only approximate: although ResNet50 achieves high accuracy in generating the target captions, its performance is still less than perfect.

We were also faced with the problem that for many GAT images the optimal caption is a multi-word phrase (e.g., *gas pump*, *woman sewing*). To obtain semantic vectors for phrases that were not included in the Word2vec dictionary, we sometimes substituted one-word near-synonyms for which a vector was available. When that was not feasible, we used a simple averaging method, forming a vector for a phrase by averaging the vectors for its content words (cf. Kintsch, 2001).

For any pair of semantic vectors, BART uses its learned weights to calculate the posterior probability that the pair instantiates each relation in the repertoire of the model. The vector of length 258 formed by these posterior probabilities provides a distributed representation of the specific relation between the two expressions in the pair. Similarly to the procedure we followed to enable ResNet50-A to solve visual analogies, BART's preferred answer $\widehat{D}$ is that which minimizes the cosine distance between the *A:B* relation and the relation formed by *C* paired with each available option.

For the GAT problems, the BART model achieved 67% accuracy in choosing the correct *D* term; other choice probabilities were 11%, 11% and 1% to choose semantic distractors, visual distractors, and unrelated distractors, respectively (see Figure 5). To provide a baseline semantic model, the performance of Word2vec (Mikolov et al., 2013), which does not learn specific semantic relations, can be compared with the performance of BART. The Word2Vec model achieved 50% accuracy in choosing the correct *D* term; other choice probabilities were 11%, 17% and 22% to choose semantic distractors, visual distractors, and unrelated distractors, respectively.

Integration of Visual and Semantic Models

Finally, we examined the performance of an integrated model of solving pictorial analogies, formed by combining the measure of relational similarity obtained from the vision model (ResNet50-A) with the comparable measure obtained from the semantic model (BART). Two free parameters were introduced to create the integrated model.

We first transformed the vectors used by each model to put them on a common scale. The relational similarity measure from the visual model is based on difference vectors of visual features derived from the penultimate layer of ResNet50-A. These difference vectors take values in the range of -8 to 8. In contrast, the BART model forms relation vectors using posterior probabilities within the range [0 1]. To place the two vectors on a similar scale, we introduced a nonlinear transformation with an exponential function for the visual difference features v as $\exp(\alpha v)$ with a scale parameter, set at $\alpha = 2$. Cosine distances based on these transformed visual difference vectors were used to compute relational distance using the visual module:

$$D_v = \cos\left(\exp\left(\boldsymbol{\alpha}(\boldsymbol{f}_B - \boldsymbol{f}_A)\right), \exp\left(\boldsymbol{\alpha}(\boldsymbol{f}_D - \boldsymbol{f}_C)\right)\right).$$

The relational similarity measure derived from the semantic module, D_s, was calculated by directly using BART as described in the preceding section. The final relational distance measure was a weighted average of the measures from the visual and semantic modules, $D = \lambda D_v + (1 - \lambda)D_s$, with the weight set as $\lambda = .3$.

Figure 5 presents a summary of the results for solving GAT analogy problems based on the visual-only model (ResNet50-A), two semantic models (Word2Vec and BART), and the integrated model based relational distance measures from both visual (ResNet50-A) and semantic (BART) models. The integrated model achieved the highest accuracy (78%) in solving GAT analogy problems; other choice probabilities were 6%, 11% and 6% to choose semantic distractors, visual distractors, and unrelated distractors, respectively.

We explored the space of parameter values, and found that performance of the integrated model was quite robust. In general, the basic results were the same for a broad range of parameter values for α, as long as the value of λ was less than .5, so that the final decision was primarily driven by the semantic module, based on BART.

Discussion

The present paper provides a proof-of-concept that vision, language, and reasoning can be integrated to create a comprehensive computational model of how humans or machines might solve meaningful visual analogies. Here our focus has been on a vision module (ResNet50-A) that can generate verbal captions for line drawings, combined with a semantic module (BART) that takes word embeddings based on verbal captions and generates representations of semantic relations. Each model includes a decision procedure for assessing the similarity of relations between objects/words and selecting the best analogical completion from among a set of alternatives. The vision module alone achieves above-chance analogical performance on the GAT problems (picture analogies in *A:B :: C:?* format); the semantic module alone is more successful; and an integration of the two modules (biased to emphasize semantics, but also influenced directly by vision) is yet more successful, achieving 78% accuracy.

Perhaps the most surprising finding from our computational experiments is that the vision module alone was able to achieve above-chance accuracy in selecting the analogical completion, even though the critical relation is semantic/functional. Despite some shortcomings of visual deep learning models (Baker, Lu, Erlikhman & Kellman, 2018), the features in the later layers may capture parallels involving visual context (e.g., the fact that airplanes and eagles both cooccur with sky in many natural images, analogous to the fact that ships and fish both cooccur with water in natural images). Apparently, for some GAT problems, the similarity of the visual difference between the *A:B* pair to that between the *C:D* options is at least weakly correlated with the semantic relations that define the analogical answer. Moreover, the visual module continues to add useful information on top of that provided by the semantic module. Thus, vision may play two important functions in solving picture analogies: generating verbal captions that in turn feed the semantic module, and directly providing visual correlates of semantic relations.

The present project is only a first step toward the "holy grail" of a unified model connecting perception to thinking. The performance of the integrative model falls short of the high accuracy level achieved by healthy human adults not under time pressure (Krawcyzk et al., 2008). A number of incremental improvements are worth pursuing. ResNet50 might benefit from additional training on line drawings. Its accuracy in captioning might also be improved by making use of contextual information (e.g., the presence of a pumpkin as the *C* term in Figure 1 might aid in recognizing the pie). If the captioning accuracy of the visual module could be improved, its output could be directly passed to BART (rather than allowing BART direct access to optimal captions). Furthermore, future investigations need to explore how to combine visual and semantic knowledge to solve generative tasks in analogical reasoning (Chen, Lu & Holyoak, 2017).

Deeper developments would include adopting more sophisticated techniques for translating multi-word captions into semantic vectors, and eventually dealing with structured text descriptions of analogical scenes (Richland, Morrison, & Holyoak, 2006). Perhaps most intriguing is the possibility of creating hybridized visuosemantic representations that would allow perception to meld with meaning.

Acknowledgments

We thank Qi Xie and Amberly Tam for helping develop the ClipArt dataset. This research was funded by NSF grant BSC-1827374 to KH and HL, and BCS-1827427 to AY.

References

Baker, N., Lu, H, Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*(12), e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Chen, D., Lu, H., & Holyoak, K. J. (2017). Generative inferences based on learned relations. *Cognitive Science*. *41*, 1062-1092.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1-43.

Goranson, T. E. (2002). On diagnosing Alzheimer's disease: Assessing abstract thinking and reasoning. *Dissertation Abstracts International: Section B: Sciences and Engineering*, *62*, 4785.

Farhadi, A., Hejrati, M., Sadeghi, M. A., Young, P., Rashtchian, C., Hockenmaier, J., & Forsyth, D. (2010). Every picture tells a story: Generating sentences from images. *Proceedings of the 11th European Conference on Computer Vision*, 15–29.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Jurgens, D. A., Mohammad, S. M., Turney, P. D., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, 356-364.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*, 173-202.

Krawczyk, D. C., Morrison, R. G., Viskontas, I., Holyoak, K. J., Chow, T. W., Mendez, M. F., Miller, B. L., & Knowlton, B. J. (2008). Distraction during relational reasoning: The role of prefrontal cortex in interference control. *Neuropsychologia*, *46*(7), 2020-2032.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y, Li, L.-J., Shamma, D. A., Bernstein, S., & Li, F.-F. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, *123*(1), 32-73.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of the 31st International Conference on Machine Learning*, *32*(2), 1188-1196.

Lovett, A., & Forbus, K. (2017). Modeling visual problem solving as analogical reasoning. *Psychological Review*, *124*(1), 60-90.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

Lu, H., Wu, Y. N., & Holyoak, K. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA, 116*, 4176-4181.

Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. *Computer Vision and Pattern Recognition*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Popov, V., Hristova, P., & Royce, A. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722-745.

Reed, S. E., Zhang, Y., Zhang, Y., & Lee, H. (2015). Deep visual analogy-making. *Advances in Neural Information Processing Systems*, *28*, 1252-1260.

Richland, L. E., Morrison, R. G., & Holyoak, K. J. (2006). Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, *94*(3), 249-271.

Sadeghi, F., Zitnick, C. L., & Farhadi, A. (2015). VISALOGY: Answering visual analogy questions. *Advances in Neural Information Processing*, *28*, 1882-1890.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *Proceedings of the International Conference on Learning Representations 2015*, 1-14.

Zhila, A., Yih, W., Meek, C., Zweig, G., & Mikolov, T. (2013). Combining heterogeneous methods for measuring relational similarity. *Proceedings of NAACL-HLT*, 1000-1009.