Understanding Advertisements with BERT

Kanika Kalra, Bhargav Kurma, Silpa Sreelatha, Manasi Patwardhan, Shirish Karande
TCS Research, Pune, India
{kalra.kanika, bhargav.kurma, silpa.sreelatha,
manasi.patwardhan, shirish.karande} @tcs.com

Abstract

We consider a task based on CVPR 2018 challenge dataset on advertisement (Ad) understanding. The task involves detecting the viewer's interpretation of an Ad image captured as text. Recent results have shown that the embedded scene-text in the image holds a vital cue for this task. Motivated by this, we fine-tune the base BERT model for a sentence-pair classification task. Despite utilizing the scene-text as the only source of visual information, we could achieve a hit-or-miss accuracy of 84.95% on the challenge test data. To enable BERT to process other visual information, we append image captions to the scene-text. This achieves an accuracy of 89.69%, which is an improvement of 4.7%. This is the best reported result for this task.

1 Introduction

The advertisement understanding challenge dataset of CVPR 2018 collected textual inputs from a set of viewers to capture their interpretations of Ad images (Hussain et al., 2017). The task is to rank the given valid and negatively sampled invalid interpretations of an image. Initial approaches to the problem tried capturing the visual semantics with a combination of object proposal features and relationships of objects with common symbolism (Doshi and Hinthorn, 2018; Ye and Kovashka, 2018; Ahuja et al., 2018). Recently, Dey et al. (2019a,b) have obtained a significant improvement in performance by utilizing the text embedded in the image (termed as the scene-text) as another channel of information. These approaches do not evaluate the validity of an interpretation by using attention to associate the words and phrases of the interpretation to fragments of textual and visual cues in the image. For example, in the Ad of the car company in Figure 1, the words and phrases from a viewer's input 'I should buy this *car* because it would add some *excitement to my life*' can be associated with the object 'car' in the image and the phrase 'add spark to life' in the scene-text. To capture these mappings, we need a model that can simultaneously pay attention to the image and the interpretations at various levels of granularity.

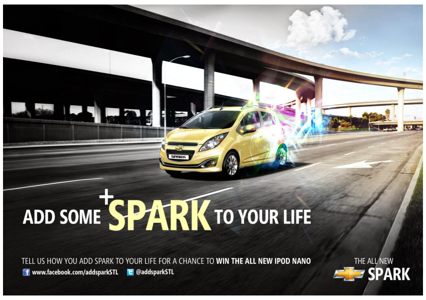
SPARK
ADD SOME +SPARK TO YOUR LIFE
TELL US HOW YOU ADD SPARK TO YOUR LIFE FOR A CHANCE TO WIN THE ALL NEW IPOD NANO
www.facebook.com/addsparkSTL
@addsparkSTL
THE ALL NEW
SPARK

Figure 1: Ads Dataset: Textual and Visual Cues

The recently proposed BERT pre-trained language model (Devlin et al., 2019) has provided excellent performance on several NLP tasks. The underlying attention-based transformer architecture (Vaswani et al., 2017) allows BERT to capture contextual representations. We leverage the pre-trained base BERT model to capture a contextual representation of the viewer's interpretation with respect to the visual and textual cues in the image.

One of the challenges we face is to provide information on visual cues to the BERT model. We overcome this challenge by extracting dense-cap captions(densecaps) (Johnson et al., 2016) to provide textual information about the image objects, their properties, and interactions. This is motivated by the approaches of Visual Question Answering (VQA) (Li et al., 2019; Hudson and Manning, 2019), question generation (Zhang et al., 2017), which talk about leveraging more abstract text or concept-level information instead of pixel-level information of an image.

We fine-tune BERT for the sentence-pair classifi-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7542–7547
July 5 - 10, 2020. ©2020 Association for Computational Linguistics

cation task, where the scene-text and the densecaps form the first sentence, and the viewer's interpretation forms the second sentence. With this approach, we achieve an accuracy of 89.69% and recall@3 of 2.411, which is by far the best reported result for this task.

2 Preliminaries

2.1 Dataset

The challenge dataset (Hussain et al., 2017) has 64,028 images. Every image has 3 to 5 interpretations in terms of Action-Reason Pairs (ARPs), which are the answers provided by a set of crowd-workers to the questions, viz. 'What should I do according to this ad?' and 'Why should I do it?' respectively. These form the valid set of ARPs. For every image, 10 to 12 ARPs are randomly negatively sampled, forming the invalid set of ARPs.

The challenge has provided 51223 images for training and 12805 images for testing. The dataset providers have taken care to ensure that there is no information leakage between these partitions by constraining the negative sampling to be from within each partition. The challenge contributors such as VSE++, ADVISE, and Cyberagent, have reported results on the test set (ref. Table 1).

Other prior works (Ye and Kovashka, 2018; Ahuja et al., 2018; Dey et al., 2019a,b) may have random partitions of the training images to obtain a validation (VAL) set, and report the results on some split of such a VAL set. A random (80-20) train-val split of the training images causes approximately 98% of the val split ARPs to overlap with the train ARPs. Unless they have taken care to partition the images first, and conduct negative sampling from only within each partition, this can lead to a possible information leakage. However, such sampling amounts to changing the training data split provided by the challenge, making it hard for the community to replicate the results. To provide a comprehensive comparison, we provide results on both the test set and the VAL split by considering a 5-fold split of the provided training images.

The challenge dataset also provides the annotations for advertisement strategies, sentiments, topics, symbolism, etc. In this work, we do not utilize these annotations. However, one can derive a potential benefit by including these annotations as additional channels of visual information. For example, previous works have included the 'symbol' annotations provided, as an additional stream. These annotations are image regions depicting symbol objects. A symbol object signifies an abstract concept. For instance, blood represents danger; muscle represents strength, etc.

2.2 Ranking Metrics

The task is to rank the validity of the ARPs concerning an image. We have considered various metrics to measure the quality of the ranking: **Accuracy:** Percentage of images having any one of the valid ARPs with rank one. **Rank:** Rank of the highest-ranked valid ARP, averaged over all images. **Rank Average:** Average of ranks of all valid ARPs of an image, further averaged over all images. **Recall@3:** Number of valid ARPs ranked in the top-3, averaged over all images.

3 Related Work

Hussain et al. (2017) introduced the CVPR challenge Ads dataset and established the baseline by modeling the task as VQA. In their proposed approach, a two-layer LSTM encodes the questions, and the last hidden layer output of VGGNet encodes the image. They convert the ARPs to a one-word answer by considering the word with the highest TF-IDF score, and the model predicts the word using a softmax layer. Symvise (Doshi and Hinthorn, 2018) uses an extension of the top-down, bottom-up attention approach (Anderson et al., 2018) by adding a symbol stream using the 'symbol' annotations provided by the dataset.

ADVISE (Ye and Kovashka, 2018) is the first paper that claims to take 'knowledge' into account for the given task and adapts (Hussain et al., 2017) for the ranking task. They use two branches, viz. (i) The main branch, which uses attention mechanism to represent an image as a weighted combination of object regions, (ii) The knowledge branch, which provides 'symbol' distribution for the image by making use of densecaps (Johnson et al., 2016) to map image to the 'symbol' labels. The embeddings received from both the branches are added to get the image embedding. They use triplet loss to learn an embedding space that keeps images closer to the valid ARPs.

Ahuja et al. (2018) proposes a weakly supervised learning algorithm that uses a multi-hop co-attention mechanism to iteratively refine the attention map that associates image proposals with symbol labels, thereby aggregating information from both modalities. They use max-margin loss to get

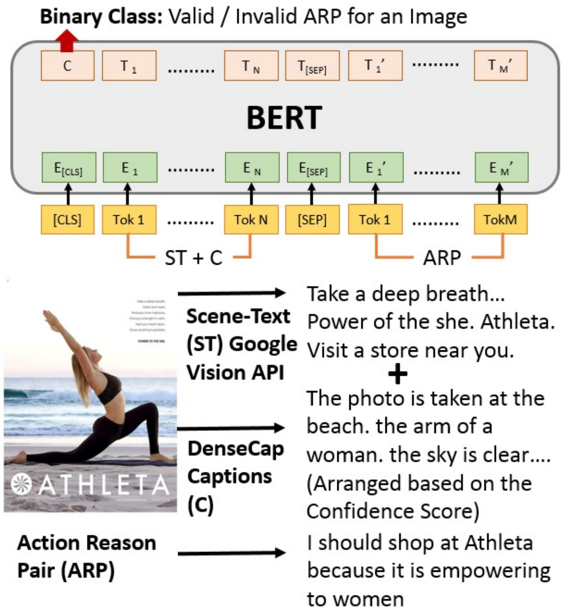
Binary Class: Valid / Invalid ARP for an Image
C
T1
TN
T[SEP]
T1′
TM′
BERT
E[CLS]
E1
EN
E[SEP]
E1′
EM′
[CLS]
Tok 1
Tok N
[SEP]
Tok 1
TokM
ST + C
ARP
Scene-Text (ST) Google Vision API
Take a deep breath... Power of the she. Athleta. Visit a store near you.
+
The photo is taken at the beach. the arm of a woman. the sky is clear.... (Arranged based on the Confidence Score)
DenseCap Captions (C)
ATHLETA
Action Reason Pair (ARP)
I should shop at Athleta because it is empowering to women

Figure 2: BERT Sentence-Pair Classification

the image-symbol embedding closer to the valid ARPs. Dey et al. (2019a) is the first approach that has considered scene-text as one of the inputs, along with the visual features. Their algorithm and training is similar to Ye and Kovashka (2018).

We draw the following learnings from the literature: (i) scene-text carries a strong signal (Dey et al., 2019b), (ii) densecaps can be used to embed external knowledge (Ye and Kovashka, 2018), (iii) capturing associations between modalities using co-attention mechanism is effective for the given task (Ahuja et al., 2018). Thus, in this paper, we leverage the pre-trained language model BERT (Devlin et al., 2019), which allows to learn contextual representations that capture associations between words and phrases of an ARP, and image inputs, using self-attention mechanism.

4 Proposed Approach

To abstract concepts from the pixel stream, we extract densecaps [1] (Johnson et al., 2016) of the image. We use Google Vision API[2] to extract scene-text from the image. We append the densecaps to the extracted scene-text to form a composite textual signal. This text is paired with an ARP to form sentence pairs, that are served as inputs to BERT, as shown in Figure 2.

The average token length of sentence-pairs is 147. For the samples for which the sentence-pair token length goes beyond the maximum allowed length (512 tokens) of the base BERT model, we truncate the length of the composite textual signal of the image. To avoid a significant information loss due to the truncation, we arrange the densecaps in decreasing order of their confidence score.

BERT (Devlin et al., 2019) has been pre-trained to use the [CLS] pin output for sentence-pair classification. Hence, we use the [CLS] pin output and fine-tune (BERT FT ST+C) (ref. Table 1) for the binary classification task to determine the validity of a candidate ARP with reference to the textual and visual cues of the image. We collect and rank the softmax outputs of all the ARPs concerning an image, to obtain their relative validity.

4.1 Ablation Studies

We fine-tune BERT with only scene-text - ARP pairs as input (BERT FT ST) and only densecaps - ARP pairs as input (BERT FT C) to understand the contribution of the different inputs. To understand the role of BERT's pretraining, we use BERT purely as a feature extractor (BERT FE ST+C) by training only a dense classifier layer over the [CLS] pin output. For training all of the above models, we use the batch size of 6, a learning rate of 2e-5, and 3 epochs.

Most of the prior work has considered an information retrieval setting in which the learned embedding of an image is matched with the learned embedding of an ARP. To compare specifically with such a setting, we have performed a sentence-pair matching task by using BERT in a siamese setting (Reimers and Gurevych, 2019). We extract sentence representations by mean-pooling the word vectors and use mean-squared-error loss over cosine similarity of the sentence vectors. We fine-tune siamese BERT (SBERT FT ST+C) as well as use it as a feature extractor (SBERT FE ST+C). We use a batch size of 16, a learning rate of 2e-5, and 4 epochs for its training.

There have been recent proposals for transformer-based cross-modal encoders such as LXMERT (Tan and Bansal, 2019) and ViLBERT (Lu et al., 2019), showing promising performance on VQA. To evaluate the efficacy of these models on the Ads dataset, we fine-tune them for a binary classification task that determines the validity of an ARP with reference to the object proposals obtained from an Ad image. We retain

[1] We used the April, 2019 version of the code from https://github.com/jcjohnson/densecap

[2] https://cloud.google.com/vision/docs/ocr

Method	Image	TEST Data				VAL Data**			
	Input	Accu	Rank	Rank	Recall	Accu	Rank	Rank	Recall
		-racy		Avg	@3	-racy		Avg	@3
VSE++	O	62%†	-	-	-	66.6%‡	-	3.858‡	-
Symvise*	O	57.11%	1.998	4.227	1.601	59.73%	1.931	4.049	1.683
LXMERT	O	50.00%	2.262	5.000	1.410	53.22%	2.159	4.860	1.470
VilBERT	O	61.76%	1.860	4.19	1.710	64.13%	1.760	4.028	1.790
ADVISE	O + K	**69%**†	-	-	-	**72.84%**‡	-	**3.552**‡	-
cyberagent †	ST + O	82%	-	-	-	-	-	-	
VS (v1)	ST + O	-	-	-	-	88.70%	-	-	-
VS (v1)*	ST + O	**86.84%**	**1.264**	**3.072**	**2.259**	**89.28%**	**1.213**	**2.889**	**2.356**
VS (v3)	ST + O	-	-	-	-	**90.90%**	-	**3.090**	-
SBERT FE	ST + C	37.31 %	2.870	6.515	1.024	37.59 %	2.847	6.472	1.025
BERT FE	ST + C	81.94%	1.496	3.854	2.078	84.10%	1.423	3.744	2.141
SBERT FT	ST + C	84.54%	1.334	3.123	2.310	87.87%	1.269	2.993	2.413
BERT FT	C	60.09 %	2.175	4.489	1.667	62.81%	2.012	4.284	1.743
BERT FT	ST	84.95%	1.884	3.622	2.271	87.53%	1.774	3.502	2.353
BERT FT	ST + C	**89.69%**	**1.230**	**2.982**	**2.411**	**91.56%**	**1.189**	**2.830**	**2.487**

Table 1: Results on CVPR 2018 Challenge Data (FE: Feature Extractor, FT: Fine-Tuned, ST: Scene-Text, C: Densecap Captions, O: Object-Proposals, K: Knowledge) Symvise (Doshi and Hinthorn, 2018), VS(v1):Visual Semantics version 1 (Dey et al., 2019a), VS(v3): Visual Semantics version 3 (Dey et al., 2019b), LXMERT (Tan and Bansal, 2019), VilBERT (Lu et al., 2019), BERT (Devlin et al., 2019), SBERT: Siamese BERT (Reimers and Gurevych, 2019), * Our implementation , ** Results on their respective VAL splits, our results are on 5-fold train-val split, † Results from challenge leaderboard (https://evalai.cloudcv.org/web/challenges/challenge-page/86/evaluation), ‡ Results from ADVISE github page (https://github.com/yekeren/ADVISE-Image_ads_understanding) - April 2020.

the hyper-parameters provided in LXMERT and ViLBERT, except for a reduced learning rate of 4e-7.

5 Results and Analysis

In this section, we compare the performance of the models as presented in Table 1, draw empirical observations, and attempt to provide a rationale for the performances observed. We also provide qualitative insights for some failure cases by manually inspecting the data.

We first make a broad observation that the performance of all the techniques on the test data is inferior as compared to the VAL data. Information leakage can be one of the reasons for observing better performance on the VAL data. Hence, we limit most of the discussion to the test set, but one can observe that the comparative performance of the models is similar on the VAL set. VS(v3) has been published simultaneously to our work; hence we were unable to create results for the test data for this model. Nevertheless, we observe that (BERT FT ST+C) could give better performance on VAL Data**.

5.1 Performance Analysis

Our proposed (BERT FT ST+C) model achieves the best performance on all the metrics amongst the considered models. We observe that just using scene-text (BERT FT ST) gives an accuracy of 84.95%, which is within 1.89% of VS(v1)*. Furthermore, the performance of BERT with just densecaps as input (BERT FT C) is competitive with other models that use just the visual cues as input. We compare (BERT FT C) and (BERT FT ST), and observe that the contribution of scene-text in the accuracy is higher, compared to densecaps. This validates the primary observation of Dey et al. (2019a).

In Table 2, we compare the BERT models with different inputs in terms of the number of misses of one model that are converted to hits by another. This represents the potential advantage that a model can get by adding or removing an information channel. We observe that for the misses of the (BERT FT ST+C), (BERT FT ST) was able to make correct inference for 2.31% of the images, whereas (BERT FT C) could infer correctly for 4.02%. This leads us to the conclude that, for some images, scene-text and densecaps do not combine well, blocking cor-

	ST+C	ST	C
ST+C	0	7.76%	34.33%
ST	2.31%	0	33.36%
C	4.02%	8.50%	0

Table 2: Cell-(i, j): % of test set images that were misses by model j, converted to hits by model i

rect inference. This is further validated, when we observe that for the misses of the (BERT FT ST) model, (BERT FT ST+C) was able to make correct inference for 7.76% of the images which is 51.58% of the misses of (BERT FT ST), whereas (BERT FT C) could infer correctly for 8.50% which is 56.51% of the misses. The performance of (BERT FT C) is inferior to VilBERT, and ADVISE that directly operate on object proposals, implying a loss of information. Comparing VilBERT and (BERT FT C), we observe that VilBERT could give $\sim$ 18.5% unique hits. However, after the addition of scene-text (BERT FT ST+C), the unique hits of VilBERT have dropped to $\sim$ 4.8%. This shows that adding an object proposal stream to (BERT FT ST+C) could contribute only a low additional advantage. We make a similar comparison of VS(v1)* with (BERT FT ST+C) and observed that only 5% of the images get converted to hits by VS(v1)*. Note that this number is in the same range as 4.02% obtained for (BERT FT C).

We observe that, BERT without any fine-tuning (BERT FE ST+C) has achieved an accuracy of 81.94% by itself. Fine-tuning BERT (BERT FT ST+C) results in an improvement of only 7.75%. This shows that BERT's pre-training has played a significant role in achieving this accuracy. However, the performance of matching BERT features (SBERT FE ST+C), which does not use attention between the ARP and the composite textual signal of the image, achieves only 37.31% in comparison to (BERT FE ST+C). This substantiates our argument that using attention to associate words and phrases in the ARPs to textual and visual cues in the image helps the task. Nevertheless, after fine-tuning, (SBERT FT ST+C) achieves an accuracy of 84.54%, which, though inferior to 89.69% (BERT FT ST+C), is within 2.3% of VS(v1)*.

5.2 Does BERT have to do any work ?

We wanted to evaluate the indirect inference BERT has to conduct. Towards this, we analyze the syntax matches of densecaps and scene-text with the ARPs concerning an image on the test data. Two sentences are said to have a syntax match if there is atleast one word common between them. We remove non-alphanumeric characters and additionally perform stemming on ARPs and densecaps. We perform POS tagging on densecaps and ARPs and consider only Nouns, Pronouns, Adjectives, and Adverbs POS Tags for syntax match analysis. We observe that 14.12%, 56.73%, and 62.46% of samples show syntax matches between valid ARPs and inputs of (BERT FT C), (BERT FT ST) and (BERT FT ST+C), respectively. Meanwhile, the corresponding numbers for the invalid ARPs are 6.32%, 10.58%, and 15.93%. This establishes that syntax matches are a major discriminating factor. However, a comparison with Table 1 shows that the performance of these models cannot be entirely attributed to syntax matches.

5.3 Failure of Neural Extractors

We manually inspect 900 randomly sampled images from the test dataset and made the following qualitative observations on the errors/limitations of the scene-text extractor and densecaps. We observe that for 82.6% of the images, at least some scene-text was not detected. We also notice that spelling errors were substantial. The causes for these could be the usage of a non-standard font, poor resolution, curvy or rotated text, non-English language, or overlapping with an object. We observe several spurious and false-positive dense captions. In future, the captions could be more helpful if they capture (i) additional object classes, e.g., cigarettes, ice-cream, etc., (ii) semantic attributes such as age or emotions, (ii) object parts or fine-granular classification, e.g., ketchup bottle or perfume, (iii) object interactions, (iv) scene or situation depicted in the image such as office, fight, romance, etc.

6 Conclusion and Future work

The scene-text holds vital information and can be used to achieve good accuracy on this task. Syntax matches play a vital role in achieving the accuracy, but are not entirely the reason behind it. Although the conversion of visual cues to captions cause a loss of information, the addition of scene-text mitigates most of the loss. Using attention to associate the ARPs with the textual and visual cues is helping the task. Better emotion, scene, scene-text, object detection and captions might lead to further improvement of performance.

References

Karuna Ahuja, Karan Sikka, Anirban Roy, and Ajay Divakaran. 2018. Understanding visual ads by aligning symbols and objects using co-attention. In *CVPR Workshop: Towards Automatic Understanding of Visual Advertisements (ADS)*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Arka Ujjal Dey, Suman Kumar Ghosh, and Ernest Valveny. 2019a. Beyond visual semantics: Exploring the role of scene text in image understanding. *arXiv preprint arXiv:1905.10622v1*.

Arka Ujjal Dey, Suman Kumar Ghosh, and Ernest Valveny. 2019b. Beyond visual semantics: Exploring the role of scene text in image understanding. *arXiv preprint arXiv:1905.10622v3*.

Rohan Doshi and William Hinthorn. 2018. Symbolic vqa on visual advertisements with symvise networks.

Drew Hudson and Christopher D Manning. 2019. Learning by abstraction: The neural state machine. In *Advances in Neural Information Processing Systems*, pages 5901–5914.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1705–1715.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. 2019. Visual question answering as reading comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6319–6328.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3980–3990, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Keren Ye and Adriana Kovashka. 2018. ADVISE: Symbolism and external knowledge for decoding advertisements. In *Proceedings of the European Conference on Computer Vision (ECCV 2018)*, pages 837–855.

Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. Automatic generation of grounded visual questions. *International Journal of Computational Intelligence and Applications*.