2021 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-4115-5/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICIP42928.2021.9506539

A DIAGNOSTIC STUDY OF VISUAL QUESTION ANSWERING WITH ANALOGICAL REASONING

Ziqi Huang[1], Hongyuan Zhu[2], Ying Sun[2], Dongkyu Choi[3], Cheston Tan[2], Joo-Hwee Lim[1,2]*

[1]Nanyang Technological University, [2]I2R, A*STAR, [3]IHPC, A*STAR

ABSTRACT

The deep learning community has made rapid progress in low-level visual perception tasks such as object localization, detection and segmentation. However, for tasks such as Visual Question Answering (VQA) and visual language grounding that require high-level reasoning abilities, huge gaps still exist between artificial systems and human intelligence. In this work, we perform a diagnostic study on recent popular VQA in terms of analogical reasoning. We term it as *Analogical VQA*, where a system needs to reason on a group of images to find analogical relations among them in order to correctly answer a natural language question. To study the task in depth, we propose an initial diagnostic synthetic dataset *CLEVR-Analogy*, which tests a range of analogical reasoning abilities (e.g. reasoning on object attributes, spatial relationships, existence, and arithmetic analogies). We benchmark various recent state-of-the-art methods on our dataset and compare the results against human performance, and discover that existing systems fall shorts when facing analogical reasoning involving spatial relationships. The dataset and code will be publicly available to facilitate future research.

Index Terms— analogical reasoning, visual reasoning, Visual Question Answering (VQA), synthetic dataset, benchmark

1. INTRODUCTION

Analogical reasoning is a crucial component of cognition and intelligence [1]. For example, in the US-based Scholastic Aptitude Test (SAT), to answer the question *"Hand is to palm as foot is to what?"*, subjects should reason analogically in lexical terms: *palm* is the inner surface of *hand*, then the underside of *foot* will be *sole*. As the analogy can be implicit and ambiguous, this problem is challenging and not well solved.

Although deep learning has recently achieved significant progress in computer vision and natural language processing, especially in Visual Question Answering (VQA), to what extent existing methods can perform analogical visual reasoning is still questionable and under-explored. Specifically, given several image pairs, where there are certain common rules applied from the first to the second image every pair, we would like to test if existing methods can identify and apply the rules to a new image to answer the natural language questions regarding the changes (see Fig. 1 as an example).

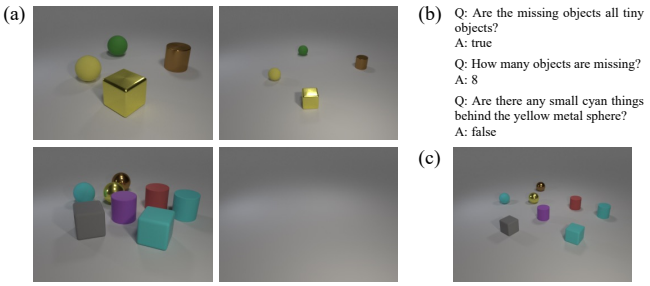

Fig. 1: A sample Analogical VQA problem from CLEVR-Analogy. (a) a group of images with analogical relations. In each row, from the image on the left to the right, every object changes from large size to small size, with other attributes unchanged. The bottom-right image is intentionally designed to be incomplete, requiring the solver to recover its content using analogical reasoning. (b) questions asked on the incomplete image in (a), and the ground-truth answers. (c) the ground-truth content of the incomplete image.

To perform objective evaluation, we propose a benchmark and diagnostic dataset called the CLEVR-Analogy dataset (CLEVR [2] for the Compositional Language and Elementary Visual Reasoning diagnostics dataset, upon which our dataset is built). CLEVR-Analogy has 479,900 Analogical VQA problems and 90,000 images, distributed in 4 broad categories and 9 analogical relations. Each Analogical VQA problem has 4 analogically related images (i.e. 2 pairs), 1 question, and 1 ground-truth answer.

We use CLEVR-Analogy to diagnose various baselines and find that existing methods struggle on analogies related to spatial relationships. The notable performance gap between the best performing VQA systems (64%) and human subjects (93%) calls for further research into this task.

The contributions of this paper are threefold:

1. We define a new task, Analogical VQA, which requires analogical reasoning on multiple images to correctly

*This work is done during Huang Ziqi's internship at A*STAR, with corresponding author Dr. Zhu Hongyuan. This work is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A2b0046).

answer a natural language question.
2. We propose the CLEVR-Analogy dataset for Analogical VQA and analogical reasoning study.
3. We conduct experiments on various baselines and analyze current systems' analogical reasoning abilities.

2. RELATED WORK

Visual Reasoning: VQA [3] is a representative task to measure visual reasoning abilities because it requires reasoning the questions over one image to infer the correct answer and has a well-defined evaluation metric. Different neural solutions have been explored to solve VQA. For example, [4–7] uses modular or graph networks to model individual steps in the reasoning process. Our task is a new variant of VQA, as it requires not only understanding language and visual content but also learning and applying the analogy rules to address the question. Visalogy [8] studies analogical reasoning using the task "image A is to image B as image C is to what". Our Analogical VQA task is different from Visalogy in two ways: (1) Our task is in the form of VQA while Visalogy outputs a ranking of candidate images; (2) Visalogy only uses a simple attribute change of a single entity to model analogical relations but we define a much wider range of analogy rules and apply them on multiple objects every image.

Visual Question Answering: There are different approaches to solve VQA problems. Joint embedding [3] approaches first extract image features by a CNN and question features by a recurrent model like LSTM [9], then fuse both features to obtain a multimodal embedding. Various multimodal pooling techniques [10–12] are explored to better express the fusion of visual and textual information. Attention mechanism enables [13, 14] to focus on local image regions selectively and sequentially. Symbolic approaches [5, 6] further improve VQA system performance by specializing various sub-tasks using different modular networks or functional programs, but require engineering a predefined inventory of modules.

Synthetic Datasets: Synthetic datasets have several advantages over collected real-world data: (1) no cost on data collection or labeling; (2) lower contextual biases; (3) highly controllable in data content and distribution. Our dataset CLEVR-Analogy is most closely related to the VQA synthetic dataset CLEVR [2], upon which several benchmarks for change captioning [15], causal reasoning [16] and spatiotemporal reasoning [17] are constructed. Unlike CLEVR [2], which provides exactly one image in each VQA round, our dataset has a group of analogically related images for each Analogical VQA problem and requires reasoning over the underlying analogy to correctly answer the question. RAVEN [18] benchmarks abstract visual reasoning abilities by image matrices, inspired by Raven's Progressive Matrices (RPM), a type of human intelligence test. Unlike RAVEN [18], whose task is multiple-choice questions and whose dataset uses simple 2D gray-scale geometric shapes in a discrete set of positions, our task requires answering free-form questions and increases the visual complexity using a 3D, colored and continuous position setting.

3. THE ANALOGICAL VQA TASK

In this section, we introduce the Analogical VQA task.

For each Analogical VQA problem, there are n pairs of images $\mathbf{I} = \{\{I_1^a, I_1^b\}, \{I_2^a, I_2^b\}, ..., \{I_i^a, I_i^b\}, ..., \{I_n^a, I_n^b\}\}$. For each pair of images $\{I_i^a, I_i^b\}$, there are a series of manipulation rules $\mathbf{R}$ applied on I_i^a to form I_i^b, denoted as $I_i^b = \mathbf{R}(I_i^a)$. All n pairs of images share a common manipulation rule $\mathbf{R}^*$, which will be discussed in Section 4.2.

Content of all images are visible except for the second image I_n^b in the final pair $\{I_n^a, I_n^b\}$, whose content is incomplete by removing some or all objects. Hence, the actual set of input images is $\mathbf{I}^* = \{\{I_1^a, I_1^b\}, \{I_2^a, I_2^b\}, ..., \{I_i^a, I_i^b\}, ..., \{I_n^a, I_n^{b*}\}\}$, where I_n^{b*} is the incomplete version of I_n^b. There is one natural language question $\mathbf{Q}$ asking about the content of the ground-truth image I_n^b.

We formulate the Analogical VQA task as $\mathbf{A} = \mathbf{M}(\mathbf{I}^*, \mathbf{Q})$, that is, given n pairs of images $\mathbf{I}^*$ and one natural language question $\mathbf{Q}$, the model $\mathbf{M}$ predicts the answer $\mathbf{A}$.

In order to correctly solve an Analogical VQA problem, the model $\mathbf{M}$ should first analogically reason on the n pairs of input images to discover the common mapping relation $\mathbf{R}^*$, then apply the mapping relation $\mathbf{R}^*$ on image I_n^a to reason the full content of the ground-truth image $I_n^b = \mathbf{R}^*(I_n^a)$ to reach the answer.

4. THE CLEVR-ANALOGY DATASET

The CLEVR-Analogy dataset studies analogical reasoning on a group of images. It provides diagnostic tools to evaluate different analogical reasoning abilities. Unlike the existing dataset on visual analogy [8], our dataset is free of object or scene bias. CLEVR-Analogy is the first dataset for the Analogical VQA task. We provide 479,900 Analogical VQA problems (see Table 1(a)) categorized into 9 different types according to analogical relations among images in one problem (see Fig. 2 for examples of the 9 analogy rules).

4.1. Images

The CLEVR-Analogy dataset is built upon CLEVR [2], adopting similar object attribute settings. There are two object *sizes* (small, large), eight *colors* (gray, red, blue, green, brown, purple, cyan, yellow), two *materials* (shiny metal, matte rubber), and three *shapes* (cube, sphere, cylinder). Objects are placed on a light gray plane. We provide ground-truth annotations of attributes (*size*, *color*, *material*, *shape*, *orientation*, and *position*) for every object, and pair-wise spatial relationships (left, right, front, behind) for all object pairs in the same image.

Table 1: CLEVR-Analogy Dataset statistics. In table (b), *Problems* refer to Analogical VQA problems.

Split	Total	Train	Validation	Test
Images	90,000	72,000	9,000	9,000
Questions	479,900	383,916	48,000	47,984
Problems	479,900	383,916	48,000	47,984

(a) Dataset size by split.

Type	Attribute	Symmetry	One Object	Array
Images	20,000	10,000	40,000	20,000
Questions	119,911	60,000	239,989	60,000
Problems	119,911	60,000	239,989	60,000

(b) Dataset size by analogy rule category.

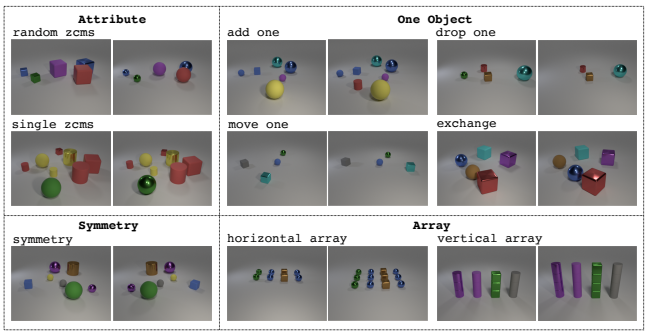

Fig. 2: Examples of analogy rules. There are 9 analogy rules, summarized into 4 broad categories. For each one of the 9 analogy rules, we present a pair of images as an example.

4.2. Analogy Rules

The number of image pairs n per problem is a hyperparameter that can be tuned during data generation. Smaller n means fewer reference pairs available for mapping relation discovery, while larger n increases computation consumption. When generating CLEVR-Analogy, we set $n = 2$ (see Fig. 1). Both pairs of images in one problem share a common mapping relation from the first to the second image instantiated from an analogy rule. The dataset is instantiated from 9 analogy rules (see Fig. 2), falling into 4 broad categories (see Table 1(b)), where each examines one aspect of analogical reasoning abilities. Below are the descriptions of the analogy rules.

(1) `Attribute`: One of the attributes (*size*, *color*, *material*, *shape*) changes its value from A to B in all image pairs.
- `random zcms`: All the objects are subject to change. In Fig. 1, all objects change from large to small size.
- `single zcms`: Only one object changes in a pair.

(2) `Symmetry`: It has only one sub-category `symmetry`.
- `symmetry`: All objects move to their symmetric positions. The line of symmetry is either horizontal (left to right) or vertical (front to behind) in a problem.

(3) `One Object`: In both image pairs, one object is added, removed, moved or swaps position with another object.
- `add one`: The same new object is added in both pairs.
- `drop one`: The same object is removed in both pairs.
- `move one`: One object changes position. Start and end positions of the moved objects are the same in both pairs.
- `exchange`: Two objects swap positions. In both pairs, the positions of the two swapped objects are the same.

(4) `Array`: Objects are arranged in a 2D array following an obvious pattern. From the first to second image each pair, one additional row or column is added to the object array. Since there are 4 directions in which we can add a line to an array, we use the same direction for both pairs in one problem.
- `horizontal array`: The array lies horizontally.
- `vertical array`: The array stands vertically.

The analogy rule for each problem is annotated. Since the final image in each problem might be fully or partially incomplete, we also provide its ground-truth rendered image (e.g. Fig. 1(c)) and annotations for easier reference.

5. EXPERIMENTS

5.1. Models

We model Analogical VQA as multiclass classification over a predefined set of candidate answers, and report classification accuracy on the test set for each baseline (see Fig. 4).

(1) Q-type: The baseline predicts the most frequent training-set answer for each question type.

(2) LSTM QA: The questions are processed with learned word embeddings followed by an LSTM [9]. The final LSTM hidden state is passed to a multi-layer perceptron (MLP) that predicts a distribution over candidate answers.

(3) BoW: For the 4 input images, we use CNN to extract 4 feature maps, and perform Global Average Pooling [19] on each feature map to obtain 4 image vectors, which are then passed to a self-attention layer to produce 4 attended vectors. We take the average of all attended vectors to obtain the image embedding. For the question, we flatten its word embeddings to a vector and pass it to 2 fully-connected layers to obtain a question embedding. The image and question embeddings are multiplied elementwise and passed to an MLP to predict a softmax distribution over candidate answers like [3].

(4) Vanilla: Images are encoded the same way as the BoW baseline using CNN, and questions are encoded the same way as the LSTM QA baseline using LSTM. The image and question representations are multiplied elementwise and passed to an MLP to predict answer distributions like [3].

(5) MCB: Images and questions are encoded the same way as the Vanilla baseline, but the image and question representations are pooled using Multimodal Compact Bilinear Pooling (MCB) [10, 11] before answer classification.

(6) SA: We use LSTM and CNN to extract question and image features, then obtain a final image representation using the architecture in Fig. 3. Finally, we use the question vector

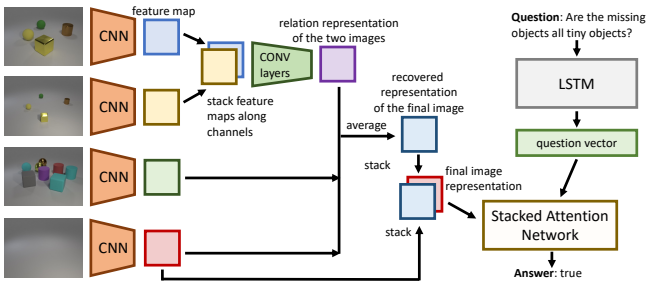

Fig. 3: Network architecture of the SA baseline.

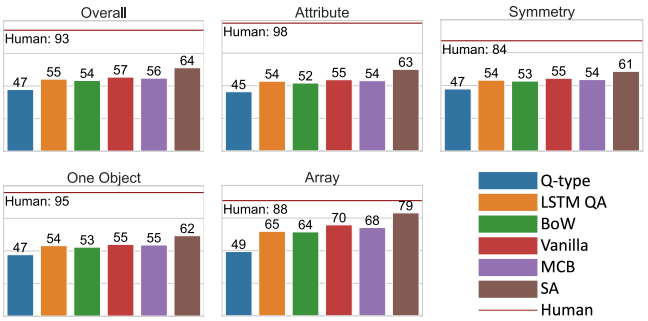

Fig. 4: Testing accuracy (%) of each model against human performance. *Overall* denotes the performance on the entire test set, while the other 4 charts are for the 4 analogy rule categories respectively.

to query the final image representation via soft attention twice using Stacked Attention Network [13] to predict the answer.

(7) Human: We randomly sample test set problems for human evaluation, and report the average accuracy of all human subjects.

Experimental Setup. We split CLEVR-Analogy into 80% for training, 10% for validation, and 10% for testing (see Table 1(a)). There are 2 pairs of images in each Analogical VQA problem (i.e. 4 images per problem). We train all baselines for 20 epochs on the training set and report test accuracy on the epoch with the highest validation accuracy.

Implementation Details. Images are normalized and resized to 224×224 before feature extraction. The CNN that extracts image features is ResNet-101 [20], pretrained on ImageNet and fixed during training. We use the 14×14×1024 feature maps from the last layer of conv4 stage. LSTMs are 2-layer with cell size of 512 or 1024. MLPs use ReLU and dropout [21] and have 1 hidden layer of size 1000. All models are trained using Adam [22] with learning rate 5e-4.

5.2. Analysis by Analogy Rules

(1) `Attribute`: In this type, each object changes at most one of its non-spatial attributes. The accuracy of SA is higher than that of **`Symmetry`** and **`One Object`**. Human subjects also perform the best on **`Attribute`**. This can be attributed to perfect one-to-one correspondence of objects in both images every pair, which eases models and human subjects from reasoning on object mapping or spatial relationship changes.

(2) `Symmetry`: In this type, two images in a pair share low pixel-level similarity due to the symmetry spatial manipulation. Most baselines including Human perform the worst on **`Symmetry`**. We believe that existing models fail to identify the concept of spatial symmetry, or to learn the true semantics of spatial relationships between a pair of images. One possible reason is that, for **`Symmetry`**, objects in I_n^a and I_n^b are at completely mirrored positions with huge spatial structure change, which makes it difficult for models and human subjects to reason and imagine I_n^b. For other types, most objects in the final image I_n^b remain spatially unchanged compared to its paired image I_n^a so VQA can be performed on I_n^b by referencing to I_n^a as a guidance.

(3) `One Object`: The baseline that performs the best on **`One Object`** is SA because it uses attention mechanism to focus on the specific object(s) that changes. Among sub-types of **`One Object`**, SA performs the worst on `exchange` because `exchange` involves position change of 2 objects while other sub-types involve only one.

(4) `Array`: All baselines except for Human perform the best on this type, where objects are arranged in arrays. In all other types, object positions appear more random and chaotic, adding complexity to spatial relationships. Q-type and LSTM QA performing the best on **`Array`** suggests that it has question-conditional biases higher than other types, partially contributing to better performance of other models. The Vanilla, MCB, and SA baselines all perform 1 to 2 points better on `vertical array` than `horizontal array` because heavier object occlusions in `horizontal array` add difficulty in detecting object quantities and attributes.

In summary, models perform relatively better on analogies related to local manipulations but still struggle to discover the underlying spatial relationships among objects or between two images. For example, models are better at discovering the lower-level fact that *"this object changes its size from large to small"* than understanding higher-level notions like *"two objects **swapped** their positions"* or *"object positions in two images are **symmetric**"*. Understanding the true semantics of higher-level analogies requires summarizing, comparing, and extrapolating multiple lower-level facts.

6. CONCLUSIONS

We introduce the new task of Analogical VQA and inspect recent state-of-the-art VQA approaches in analogical reasoning on a group of images to correctly answer a question. We also present CLEVR-Analogy, a dataset designed as a benchmark and a diagnostic evaluation tool for the Analogical VQA task. We hope that CLEVR-Analogy will help future research in Analogical VQA and enable broader reasoning tasks.

7. REFERENCES

[1] Dedre Gentner, Keith J Holyoak, and Boicho N Kokinov, *The Analogical Mind: Perspectives From Cognitive Science*, MIT Press, 2001.

[2] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1988–1997.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[4] Jiaxin Shi, Hanwang Zhang, and Juanzi Li, "Explainable and explicit visual reasoning over scene graphs," in *CVPR*, 2019.

[5] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick, "Inferring and executing programs for visual reasoning," in *ICCV*, 2017.

[6] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum, "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding," in *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[7] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu, "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision," in *International Conference on Learning Representations*, 2019.

[8] Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi, "Visalogy: Answering visual analogy questions," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. 2015, vol. 28, pp. 1882–1890, Curran Associates, Inc.

[9] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, Nov. 2016, pp. 457–468, Association for Computational Linguistics.

[11] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 317–326.

[12] T. Do, H. Tran, T. Do, E. Tjiputra, and Q. Tran, "Compact trilinear interaction for visual question answering," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 392–401.

[13] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 21–29.

[14] Huijuan Xu and Kate Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," *european conference on computer vision*, 2016.

[15] Huk Dong Park, Trevor Darrell, and Anna Rohrbach, "Robust change captioning," *ICCV*, pp. 4623–4632, 2019.

[16] Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum, "Clevrer: Collision events for video representation and reasoning," in *ICLR 2020*, 2020.

[17] Rohit Girdhar and Deva Ramanan, "CATER: A diagnostic dataset for compositional actions & temporal reasoning," in *ICLR 2020*, 2020.

[18] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu, "Raven: A dataset for relational and analogical visual reasoning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," *ICLR*, 12 2014.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[22] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2014.