Using Small VAE-Inspired Models for VQA

Donglin Wang [1] **Nicolas Fleece** [1] **Galen O'Shea** [2]

Abstract

VQA is a challenging dataset where the machines are asked to answer questions based on pictures. VQA is particularly challenging because requires machine to display intelligence across two modalities. More importantly, it requires the model to efficiently combine the two modalities to come up with reasonable answers. Although many state-of-the-art methods deliver high performance on the test accuracy, these method can't easily be transferred to other tasks. This is because they often have strong assumptions over the relationship between pictures, questions, and answers. In addition, they use special techniques that can be cumbersome to implement and fine tune. Therefore, we propose a series of baseline that have been used for multi-modal learning and apply them to VQA quesitons. Although we don't expect SOTA results, our goal is to show that small multi-modal model can also perform well on VQA. We experimented with three different architectures to test.

1. Introduction

VQA dataset is a massive dataset that was proposed by Agrawal et al. The dataset is an expansion on the COCO 2014 dataset. Agrawal et al. collected a series of questions and answers through Amazon Mechanical Turk from human participants. There are a total of 760 thousands unique questions, 250 thousands unique images, and 10 million answers from human responders. Many questions can be mapped to the same image and vice versa. There are many state of the art (SOTA) methods that achieved test accuracies that surpassed even human on the task. However all of these methods have some significant drawbacks.

Firstly, they make domain-specific and non-transferable assumptions on the correlation between the salient objects in the image and words in the sentences. Therefore, they heavily depend on pre-trained object detection models such as Fast-RCNN. Not only this, some methods even go as far as to assume the probabilistic dependence between the salient object in the pictures. For examples, Wang et al. assumes that the objects that are frequently present in both the text and images often obey a type of do dependency. Namely, instead of assuming that there is a direct correlation between the visual data, Wang et al. claimed that there exist a latent causal variable z that should be taken into account when training over the text. Such latent variable z depend heavily on the list of available object classes and their frequencies in the image dataset. Therefore, this formulation is heavily biased towards the dataset that the Fast-RCNN is trained on.

Another family of SOTA method depends on some sort of self-supervision scheme. They rely on the robustness of the object detection algorithm to identify the object in the scene. Then, the model would attempt to find the words in the sentences that directly matches the lable of the identified objects. If there exist some objects in the sentences that are not present in the set of object detectable by the object-detecting model, the VQA model would not function well. Therefore, there is a heavy dependency between SOTA VQA methods and object detection datasets (e.g. Visual Genome) and pre-trained object detection models. In addition, the self-supervision mechanism in these model are computationally expensive. This means that they are iteratively generating masking and labeling questions recursively. This makes it unattractive devices that don't have enough computational power.

The third family of VQA models takes advantage of the self-attention mechanism (I47, I48). However, the attention mechanism can be computationally expensive. For starters, it is not easy to implement attention mechanism over images. Most SOTA VQA models implement some form of "top-down and bottom-up" approach. The "bottom-up" part refers to the construction of image regional image features. This is accomplished by using some form of pre-trained models, such as Fast RCNN, in order to obtain a set of boudning box object classifications. Therefore, the SOTA methods involves a lot of "pre-training" where the object detection modesl are pre-trained to a certain accuracy before the actual training begin. In a way, the SOTA VQA dataset has are often object-detection models with extra infused

[1]Faculty of Engineering, School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada. [2]Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada.

language features. The "top-down" part of the mechanism refers to the attention mechanism that choose which part of the image to focus on.

Given the shortcomings of these models, we decided to take a simpler approach towards the VQA task. Namely, instead of designing models specifically tailoring to the VQA task, we want to use multi-modal models that are know to generalize well across different types of input. This way, we would design a model that has some degree of transferability to other tasks. For the scope of this project, we only focused on the yes/no questions from a subset of the training dataset.

2. Background

2.1. ImageNet and VG pre-training

2.2. Variational Auto-Encoder

The Variational Auto-Encoder [1] assumes that a given dataset $X = \{x_1, ..., x_n\}$ is generated by an underlying continuous random variable z. More specifically, it assumes that the data is generated by first sampling a value from the prior distribution $p_\theta(z)$ and then generating a particular x_i using the conditional distribution $p_\theta(x|z)$. Realistically, it is intractable to find the true posterior $p_\theta(x|z)$. Therefore, we come up with an estimation $q_\phi(z|x)$ with a well-understood distribution (such as Gaussian) to approximate the posterior. To optimize the evidence lower bound of this estimation, the author of VAE derived the following loss:

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = & E_{q_\phi(z|x)}\left[\lambda \log p_\theta(x \mid z)\right] \\ & - \beta KL\left[q_\phi(z \mid x), p(z)\right]\end{aligned} \tag{1}$$

One problem with the original VAE is that it does not allow us control the output once the model is trained. If the latent space is small, one can manage to plot the output of the VAE over the entire latent space. However, if the latent space is large, it is very hard to find a correspondence between the latent space and a class label. The Conditional VAE [12] aims to solve this problem by "conditioning" the model with a class label y. However, the "condition" for a model does not have to be a class label. As we will see in Section 5, the Conditional-VAE is the inspiration for our first model where the condition is the LSTM feature of the questions. Below is the loss function proposed by the original author of the Conditional-VAE.

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} = & E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z})\right] \\ & - KL\left(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right)\end{aligned} \tag{2}$$

where

$$\mathcal{L}_{\text{cross-entropy}}\left(\hat{\mathbf{y}}, \mathbf{y}\right) = -\sum_i y_i \log\left(\hat{y}_i\right) \tag{3}$$

However, for the purpose of this project, we have modified the loss function so that it becomes binary cross entropy loss instead of the original VAE loss. This means ditching pixel-wise loss in the first term and replacing it with ... DIRECT COPY!!!!!

In addition, we also come up with what we call "combined loss" where we append the binary cross entropy loss to the CVAE loss. The combined loss can be written as:

$$\begin{aligned}\mathcal{L}_{\text{CVAE}} = & E_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z})\right] \\ & - KL\left(q_\phi(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right) \\ & + L_{binary}(p_\theta(\mathbf{y} \mid \mathbf{x}, \mathbf{z}), \mathbf{y})\end{aligned} \tag{4}$$

2.3. Multi-Modal Learning

Multi-Modal learning is a subset of learning algorithms where, instead of mapping one form of data into another, multi-modal learning aims to come up with a generalized model that can handle multiple forms of data. Specifically, we choose to use variants of VAE because it is known for its ability to learn representative features.

DIRECT COPY!!!!! There are many models that take advantage of VAEs in order to achieve multi-modal inference. Some VAE architectures attempt to approximate the joint distribution of multi-modal datasets directly. For example, Suzuki et al. [3] attempted to approximate the joint distribution $p(x_1, ..., x_n)$ directly from a multi-modal inference network $q(z|x_1, ..., x_n)$. In addition, Vedantam et al. [4] introduced a two-step process to train multi-modal VAE. Specifically, Vedantam et al. first trained a network $q(z|x_1, ..., x_n)$ on data that has all available modalities. Then, they trained a number of smaller networks between different bi-modal pairs. END DIRECT COPY!!!!!

3. Processing Limitations

Since this task involves processing large amounts of image and text data, we ran into some limitations that the papers related to this area had not ran into. Our main issue was processing power, we were mainly running our models on google colab, where we had access to one Tesla P-100 GPU. In addition to this we also had access to one RTX-3090 which we were able to run models on. While these GPU's are very fast and good for training models, there is still an inherent limitation since the SOTA models are often trained on several of these high-end GPU's. Due to this performance limitation, we were only able to use a subset of the original data for training, since it would often take hours or days to train on this subset of data with the systems we had access

to.

4. Data Processing

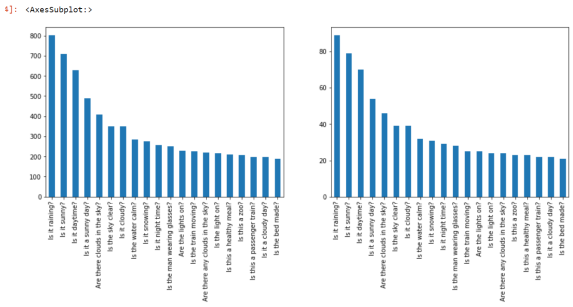

Figure 1. Baseline architecture

Due to the lack of computation power, we only have enough time to use a subset of the dataset. This turns out to be a tricky task because we have to keep the yes and no balanced while finding enough photo example for each question. If we just randomly sample a fixed percentage of the dataset, it is unlikely that we have a diverse selection of picture for a particular questions and vice versa. With some experimentation, we found that randomly sampling 15% of the dataset produce almost all distinct image-question pairs. This means that randomly sampled data would not provide enough examples for the model to train on. Therefore, we selected the top 200 most frequent question and all their corresponding picture. For pre-processing, we used first normalized the pictures and then resized the images to 224x224x3 images.

5. Model Architectures

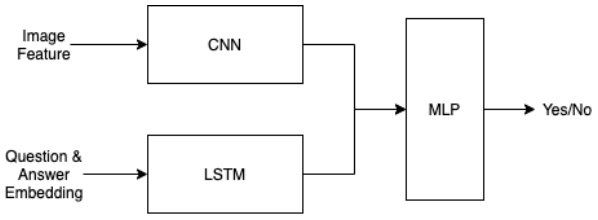

Figure 2. Baseline architecture

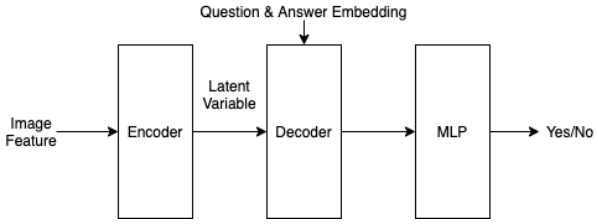

Figure 3. Conditional-VAE inspired architecture

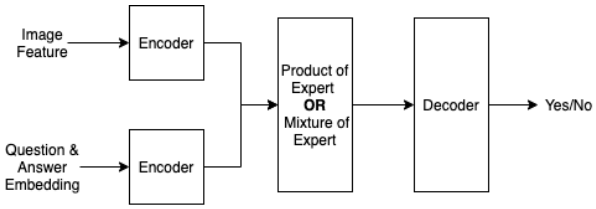

Figure 4. Multi-Modal VAE inspired architecture

As the baseline, we used the architectures proposed by the authors of the VQA dataset. We start by converting the words in the raw text for the questions with their GloVe embeddings. The words that are not found are converted to zeros.

The second architecture we adopted if inspired by the Conditional VAE. For this, we try to ”condition” the VAE according to the questions that are asked. For the textual input, we used a LSTM to and used the last hidden states as the features. For the images, we used a series of convolutional layers. The two are then concatenated and reparameterized into a latent vector of a fixed size. At this point, we have two variants. For the first variant, the latent variable is fed directly into a dense network for classification. In the second variant, the latent variable is first reconstructed into a ”latent image” via several deconvolutional layers before feeding into a classification network. We also have a variant where we used the ”Combined VAE Loss” to train the network.

The third architecture we considered is the Multi-Modal VAE proposed by Wu and Goodman in 2018. This model uses the Product of Expert (PoE, ANOTHER CITATION) approach. In this model, the text and image went through the image encoder and LSTM respectively. After the latent variable is sampled, the latent variable for text and image are multiplied together before feeding into the decoder. There are several advantages to this approach. Firstly, the dimensionality of the input to the decoder no longer depend on the size of the output features for each encoder. Therefore, the number of parameters of the model would decrease, and the decoder would have less dimensionality to deal with. Secondly, Wu and Goodman showed that models with comparable number of parameters often generalize when the features are multiplied rather than concatenated. Like the CVAE-inspired model, we trained two other varieties: one without the decoder and one with the combined loss.

5.1. Optimization Techniques

We found that drop out and batch normalization layers helped with the training and validataion accuracy of our models. We suspect that the dropout layers helped with the model significantly becausse the drop out forces the LSTM and the convolutional layers learn the representation of the

images and text. Because our models often use two independent modules to encode the images and text, it is likely that each of the modules would overfit on their perspective inputs. Therefore, the dropout act as a regularization where it grounds the two modalities together instead of letting them overfit on their own.

5.2. Results

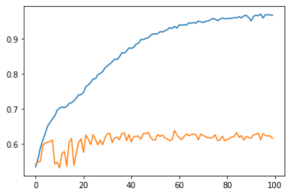

Figure 5. Baseline training

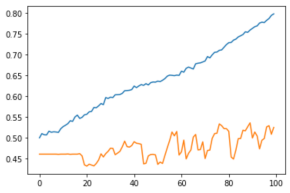

Figure 6. CVAE training

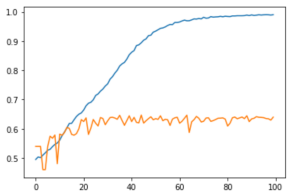

Figure 7. MVAE training

6. Conclusion

7. Future Directions

We discovered that Wu and Goodman [13] have proposed a good architecture that makes VAE robust across modalities. In their paper, they proposed the Multi-Modal VAE (MVAE) where they used a Product of Expert (PoE) to approximate the joint distribution across different modes. Before the PoE approach, for other multi-modal methods, we need to specify 2^N inference network $q(z|X)$ for every subset of modality $X \subset \{x_1, x_2, ..., x_N\}$. However, Wu and Goodman showed that the joint distribution $p\left(z \mid x_1, \ldots, x_N\right)$ is proportional to the products of conditional distribution over each modality. Namely:

$$
\begin{aligned}
p\left(z \mid x_1, \ldots, x_N\right) &= \frac{p\left(x_1, \ldots, x_N \mid z\right) p(z)}{p\left(x_1, \ldots, x_N\right)} \\
&\propto \frac{\prod_{i=1}^{N} p\left(z \mid x_i\right)}{\prod_{i=1}^{N-1} p(z)}
\end{aligned} \tag{5}
$$

Since in most cases, we cannot directly compute the probability distribution $p(z|x_i)$, we can approximate it using $q(z|x_i)p(z)$. Therefore, the expression above can be written as:

$$
\begin{aligned}
p\left(z \mid x_1, \ldots, x_N\right) &\propto \frac{\prod_{i=1}^{N} p\left(z \mid x_i\right)}{\prod_{i=1}^{N-1} p(z)} \\
&\approx \frac{\prod_{i=1}^{N} \left[\tilde{q}\left(z \mid x_i\right) p(z)\right]}{\prod_{i=1}^{N-1} p(z)} \\
&= p(z) \prod_{i=1}^{N} \tilde{q}\left(z \mid x_i\right)
\end{aligned} \tag{6}
$$

This gives use a simple product between the features across the modalities.

Model	Loss	Accuracy	AUC	Precision	Recall
CVAE	0.9107	0.5524	0.6143	0.6789	0.3323
CVAE +	3.3885	0.4573	0.4880	0.0000	0.0000
MVAE	0.8973	0.6056	0.6929	0.7389	0.4082
MVAE +	1.1100	0.6326	0.7027	0.7381	0.4882
CNN * +	1.2665	0.7762	0.8401	0.7965	0.7891
CNN * D +	1.0443	0.7682	0.8356	0.7818	0.7946
MVAE * +	1.4014	0.7779	0.8318	0.7797	0.8233
MVAE * D +	0.6255	0.7213	0.7944	0.7945	0.6560

References

[1] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[2] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In Advances in Neural Information Processing Systems, pages 3581–3589, 2014.

[3] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. arXiv preprint arXiv:1611.01891, 2016.

[4] Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. arXiv preprint arXiv:1705.10762, 2017.

[5] Romain Mormont, Pierre Geurts, and Raphaël Marée. Comparison of deep transfer learning strategies for digital pathology. In Proceedings of the IEEE Conference on Com-

puter Vision and Pattern Recognition Workshops, pages 2262–2271, 2018.

[6] Pulkit Agrawal, Ross B. Girshick, and Jitendra Malik. Analyzing the performance of multilayer neural networks for object recognition. In European Conference on Computer Vision (ECCV), 2014.

[7]. Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Return of the devil in the details: delving deep into convolutional nets. In British Machine Vision Conference, 2014

[8] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 580–587, 2014.

[9] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Advances in Neural Information Processing Systems, pages 3320–3328, 2014.

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, pages 91–99, 2015

[11] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In Advances in neural information processing systems, pages 379–387, 2016.

[12] Sohn, Kihyuk, Honglak, Lee, and Xinchen, Yan. "Learning Structured Output Representation using Deep Conditional Generative Models." . In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2015.

[13] Mike Wu and Noah Goodman. "Multimodal Generative Models for Scalable Weakly-Supervised Learning." (2018).