

Supplementary – Creativity: Generating Diverse Questions using Variational Autoencoders

Unnat Jain*
UIUC
uj2@illinois.edu

Ziyu Zhang*
Northwestern University
zzhang@u.northwestern.edu

Alexander Schwing
UIUC
aschwing@illinois.edu

1. Quantitative Results

In the following we present additional quantitative results, some of which were already mentioned in the paper. We report average BLEU, oracle BLEU, average METEOR, oracle METEOR, unique questions (UQ) and unseen unique questions for the VQG-COCO, the VQG-Flickr, and the VQG-Bing test sets. Fig. 1 shows the average and oracle BLEU scores for the three test sets. Fig. 2 shows the average and oracle METEOR scores for the same. For diversity metrics, Fig. 3 shows the percentage of unique questions for different sampling schemes. Fig. 4 shows the percentage of unique questions generated by our model which are unseen in training. More specifically in Tab. 1, Tab. 2, and Tab. 3 we report these metrics **averaged** over all the epochs. In Tab. 4, Tab. 5, and Tab. 6 we report the **maximum** of these metrics over all the epochs. For most of the metrics we observe a uniform distribution within $[-20, 20]$ with 500 samples to perform best.

2. Qualitative Results

In Fig. 5, Fig. 6 and Fig. 7 we illustrate images and some questions that our model generated. Lighter boxes are for more *literal* questions which are based on object shape, color or count and can be easily answered by looking at the image. Darker colored boxes are for *inferential* questions, which need prior (human-like) understanding of the objects or scene. The questions with **bold ticks** (✓) are questions generated by our VQG model which never occurred during training (what we refer to as ‘unseen’ questions). We demonstrate the diversity of our model by showing a variety of literal to inferential questions as well as ‘unseen’ questions.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.331	0.37	0.188	0.207	1.78	6.54
N1, 500	0.328	0.376	0.187	0.211	2.04	7.44
U10, 100	0.305	0.447	0.178	0.254	2.04	7.44
U10, 500	0.295	0.468	0.175	0.269	12.52	16.22
U20, 100	0.295	0.486	0.172	0.281	17.02	13.66
U20, 500	0.283	0.519	0.168	0.307	33.41	19.6

Table 1: VQG-COCO Summary of metrics. Metrics **averaged** over the epochs.

Within those plots we also show some failure cases. We observe our model to face one of the following challenges: *recognition*, *co-occurrence* or *natural language* based challenges. To repeat, we term failures due to incorrect recognition (attributed to weak feature learning or description) as recognition based failures. Cases where a question is incorrectly generated due to its frequent occurrence with a particular object category are called co-occurrence based failures. Generated sentences with mistakes in the language structure are referred to as natural language based failures. We give examples of each for all three datasets.

* indicates equal contributions.

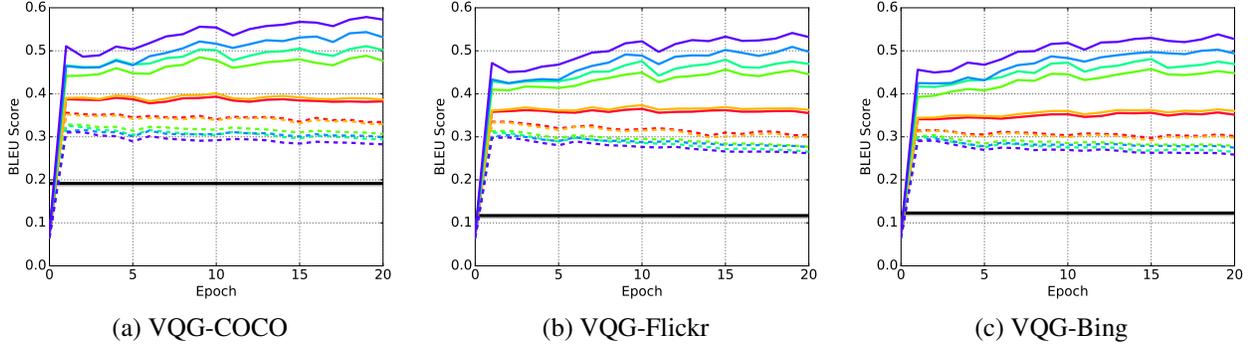


Figure 1: **BLEU Score**: Oracle-BLEU and average-BLEU score over epochs. Experiments with various sampling procedures and results compared to the performance of the baseline model [1] as line in **black bold** color. (Legend same as METEOR plots)

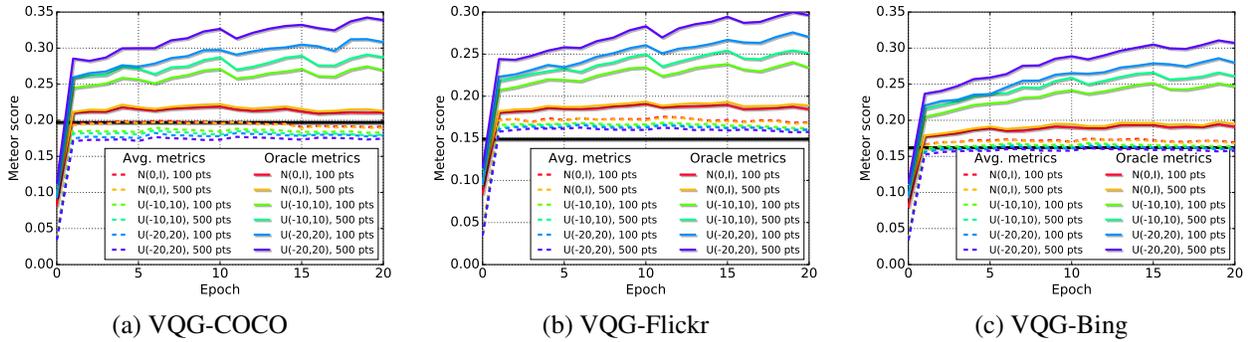


Figure 2: **METEOR Score**: Oracle-METEOR and average-METEOR score over epochs. Experiments with various sampling procedures and results compared to the performance of the baseline model [1] (line in **black** color).

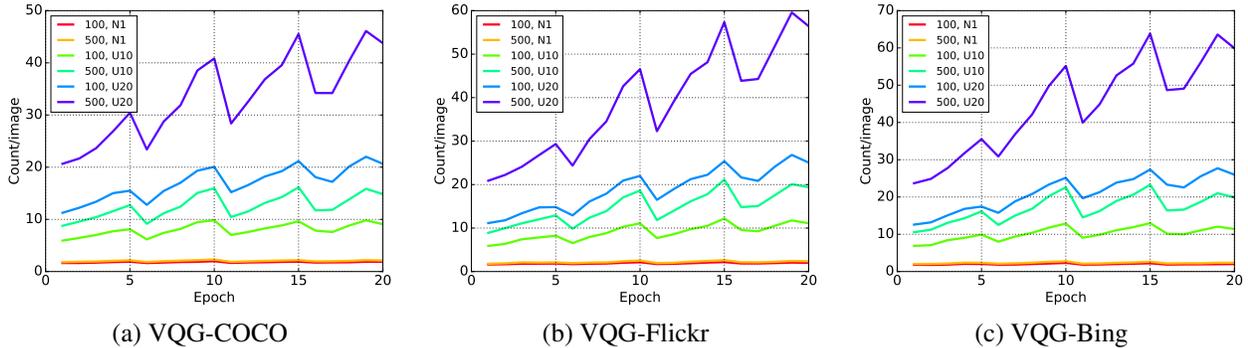


Figure 3: **Generative strength**: Number of unique questions averaged over the number of images. Shows that sampling the latent space by Uniform distribution leads to more unique questions per image.

References

[1] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende. Generating natural questions about an image. In *Proc. ACL*, 2016. 2

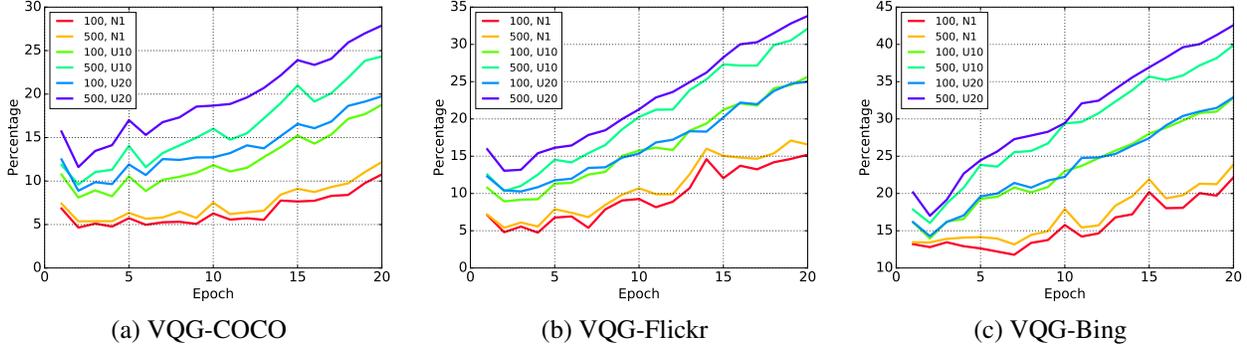


Figure 4: **Inventiveness:** $\frac{\text{Unique questions which were never seen in training set}}{\text{Total unique questions for that image}}$ averaged over the number of images. This too suggests that sampling from the uniform distributions for the latent space generates more diverse questions.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.305	0.346	0.165	0.181	1.88	9.64
N1, 500	0.302	0.351	0.165	0.185	2.18	10.87
U10, 100	0.283	0.417	0.160	0.221	9.07	16.31
U10, 500	0.275	0.436	0.158	0.234	14.73	20.59
U20, 100	0.278	0.453	0.157	0.245	18.93	16.66
U20, 500	0.267	0.483	0.154	0.267	39.01	22.6

Table 2: VQG-Flickr Summary of metrics. Metrics **averaged** over the epochs.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.295	0.336	0.165	0.183	1.98	15.56
N1, 500	0.292	0.342	0.164	0.187	2.31	17.00
U10, 100	0.277	0.415	0.159	0.228	10.17	23.43
U10, 500	0.267	0.436	0.157	0.242	16.94	28.83
U20, 100	0.272	0.452	0.155	0.252	21.06	23.65
U20, 500	0.261	0.482	0.152	0.273	44.65	30.73

Table 3: VQG-Bing Summary of metrics. Metrics **averaged** over the epochs.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.356	0.393	0.199	0.219	1.98	10.76
N1, 500	0.352	0.401	0.198	0.222	2.32	12.19
U10, 100	0.328	0.488	0.19	0.275	9.82	18.78
U10, 500	0.326	0.511	0.186	0.291	16.14	24.32
U20, 100	0.316	0.544	0.183	0.312	22.01	19.75
U20, 500	0.311	0.579	0.177	0.342	46.1	27.88

Table 4: VQG-COCO Summary of metrics. These metric values are the **maximum** over the epochs.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.335	0.365	0.176	0.191	2.17	15.2
N1, 500	0.333	0.374	0.174	0.193	2.63	17.1
U10, 100	0.314	0.456	0.168	0.241	12.21	25.65
U10, 500	0.31	0.479	0.167	0.254	21.14	32.12
U20, 100	0.304	0.509	0.166	0.276	26.83	24.98
U20, 500	0.299	0.541	0.163	0.3	59.57	33.81

Table 5: VQG-Flickr Summary of metrics. These metric values are the **maximum** over the epochs.

Sampling	Avg. Bleu	Oracle Bleu	Avg. Meteor	Oracle Meteor	UQ	Unseen UQ
N1, 100	0.316	0.357	0.175	0.194	2.32	22.15
N1, 500	0.315	0.364	0.173	0.198	2.76	22.87
U10, 100	0.304	0.457	0.168	0.252	12.99	32.84
U10, 500	0.299	0.481	0.166	0.266	23.3	39.84
U20, 100	0.296	0.503	0.164	0.286	27.71	32.91
U20, 500	0.291	0.538	0.161	0.311	63.83	42.58

Table 6: VQG-Bing Summary of metrics. These metric values are the **maximum** over the epochs.

	<p>What is the object in the picture?</p> <p>What is the color scheme of the photo?</p> <p>What type of plane is this?</p> <p>What do the wheels do?</p>	<p>What are the men doing?</p> <p>Is this a black and white photo?</p> <p>What do you call the plane in the background?</p> <p>What year was this taken?</p>		<p>What is the shape of the plate?</p> <p>What kind of food is this?</p> <p>What is the glass for?</p> <p>What kind of sandwich is this?</p>	<p>Is there a spoon in the glass?</p> <p>What is the green vegetable on the plate?</p> <p>Is the glass half full or half empty?</p> <p>What is the drink in the glass?</p>
	<p>What color is the bike?</p> <p>What is the name of the street?</p> <p>What is the name of the bike?</p> <p>What is the purpose of the blue bike?</p>	<p>What color are the bike tires?</p> <p>What is the weather like in this scene?</p> <p>Is the umbrella for sale?</p> <p>Is this a good place to put a bike?</p>		<p>What color is the building?</p> <p>What is the name of the bridge?</p> <p>What is the name of the tall building in the background?</p> <p>Is this a modern building?</p>	<p>Are there any people in the photo?</p> <p>What is the name of the hotel?</p> <p>What is the tallest building used for?</p> <p>Is this a big city?</p>
	<p>What is the color of the table the plate is on?</p> <p>Is the sandwich cut in half?</p> <p>Is the glass full?</p> <p>Is the food in the foreground of a vegetarian?</p>	<p>What color is the tablecloth?</p> <p>How many plates are there?</p> <p>What is the red vegetable on the plate?</p> <p>Is the food tasty?</p>		<p>Is she talking to a cell phone in her hand?</p> <p>What color is her hair?</p> <p>Is the woman a good phone for the day?</p> <p>Is the girl a good model?</p>	<p>What is the girl doing with her right hand?</p> <p>Is the person on the phone a man or woman?</p> <p>Is the girl using a flip phone or a smartphone?</p> <p>Is the woman happy?</p>
	<p>Is the boy holding a donut?</p> <p>Is this girl wearing a necklace?</p> <p>Is this a photo of a party?</p> <p>Are they eating at a restaurant?</p>	<p>Is the girl holding a plate?</p> <p>Is the girl holding a donut?</p> <p>Is this a family?</p> <p>Is this person eating a healthy meal?</p>		<p>How many riders are in this picture?</p> <p>Is there a fence?</p> <p>Is this a race?</p> <p>Is the rider a professional?</p>	<p>What kind of animals are these?</p> <p>Are the horses running?</p> <p>Is the rider in the background a man or a woman?</p> <p>Is this a professional photo?</p>
	<p>What does the graffiti say?</p> <p>Is this person holding a snowboard?</p> <p>Is this photo indoors?</p> <p>Is this person in danger of falling off of the ground?</p>	<p>What is the man doing in the air?</p> <p>Is this person on the ground or in the air?</p> <p>Is this person skiing on a street?</p> <p>Is this person in a competition?</p>		<p>What is the cat sitting in?</p> <p>Are the flowers in a garden?</p> <p>Are the flowers all the same?</p> <p>Are the flowers artificial?</p>	<p>Are there any roses?</p> <p>How many different types of flowers are there?</p> <p>What is the vase made of?</p> <p>Are the flowers alive?</p>

Figure 5: Examples of VQG-COCO with more questions, generated by our VQG algorithm. Darker colored boxes contain questions which are more inferential.

Recognition based failures (blue box): Due to similar appearance a woman is recognized as a boy and in another image the shadow on a wall is recognized as graffiti.

Co-occurrence based failure (pink box): Frequent occurrence of tablecloth based questions with food images generates a similar question in this image, even without a tablecloth.

Natural language based failure (red box): Correct subjects like woman, phone and day are combined in an incorrect language structure.

	Is this a pond? ✓	Is this water frozen? ✓		What are they flying? ✓	What is in the sky? ✓
	What is the white stuff on the rocks? ✓	What is the name of the river? ✓		Are the kites flying in the air? ✓	How many kites are in the air? ✓
	Is the water clean? ✓	Is the water deep? ✓		What is the weather like? ✓	Are the kites the same? ✓
	Is the water cold? ✓	Is this water safe to drink? ✓		Is this a good place to fly a kite? ✓	Is this a toy festival? ✓
	How many women are in this photo? ✓	Are the girls wearing the same color? ✓		How many boats? ✓	Is this a harbor? ✓
	the woman in the middle wearing a dress? ✓	Are these people happy? ✓		Is this a sunny day? ✓	Is this a good place to go swimming? ✓
	Are they in a public place? ✓	Are they all the same gender? ✓		What is the purpose of this body of water? ✓	Is the water calm? ✓
	Are the people all related? ✓	Are they all in love? ✓		Is the sun setting? ✓	Is this a big or a small town? ✓
	How many bicycles are in the picture? ✓	How many people are in the picture? ✓		What color is the child's hair? ✓	Is the girl wearing a helmet? ✓
	Are the bikes on display? ✓	What color are the bikes? ✓		What color is the road? ✓	What is the red object in the street? ✓
	Are these bicycles or are the same type of bike? ✓	Are they selling bikes? ✓		What is the girl riding? ✓	What color is the skateboard? ✓
	Is the bike moving? ✓	Are the bikes on the left side of the photo an antique? ✓		What is the name of the street the building is in front of? ✓	Is this a skate park? ✓
	Is the spoon in the image dirty? ✓	Is this a vegetable? ✓		What is the man holding? ✓	Is this man wearing a shirt? ✓
	Is the bowl full? ✓	Is the bowl dirty? ✓		Is the man wearing glasses? ✓	Is this man in a zoo? ✓
	Is this a healthy meal? ✓	Is the water in the bowl edible? ✓		Is this man dressed for the weather? ✓	Is the man holding his dog? ✓
	Is this a typical breakfast dish? ✓	Is the bowl of this dish a dessert or a vegetable? ✓		Is this a good place for a dog? ✓	Is this man happy? ✓
	How many candles are there? ✓	What is on the cake? ✓		How many windows are on this building? ✓	How many cars are parked on the street? ✓
	What is the cake being served with? ✓	What is the girl cutting with? ✓		How many stories is the building on the right? ✓	What color is the building? ✓
	What is in the girl's hair? ✓	What is the pattern of the girl's shirt? ✓		Is the building made of brick? ✓	Is this house made of wood? ✓
	What is the occasion that the people are celebrating? ✓	What is the occasion for this occasion? ✓		Is that a new building? ✓	Is that a church in the background? ✓

Figure 6: Examples of VQG-Flickr with more questions, generated by our VQG algorithm. Darker colored boxes contain questions which are more inferential.

Recognition based failures (blue box): An orange scooter is perceived as a red skateboard in one of the images.

Co-occurrence based failures (pink box): Frequent occurrence of spoon based questions with food images generates a similar question in one of the images, which doesn't even have a spoon. Similar is the case for candle questions in birthday images. This cake doesn't have a candle.

Natural language based failures (red box): Correct subject like bicycles is incorrectly framed in a question. Similar is the case with word 'occasion' in the birthday image.



Figure 7: Examples of VQG-Bing with more questions, generated by our VQG algorithm. Darker colored boxes contain questions which are more inferential.

Recognition based failures (blue box): Image on the top left (which is difficult for even humans to recognize) is of a tortoise/turtle with its eggs. The image looks very similar to objects like grapes, vines, snakes. We observe a recognition based failure for the image with a train station. The dark track and platform are recognized as road and sidewalk respectively.

Co-occurrence based failures (pink box): Frequent occurrence of bird based questions with tree images generates a similar question in one of the images, which doesn't even have a bird. Similarly, license plate question pops up in the car image. This car view doesn't have a license plate view.

Natural language based failures (red box): Correct subjects like trail and mountains are incorrectly framed in a question.