Multimodal Differential Network for Visual Question Generation

Badri N. Patro **Sandeep Kumar** **Vinod K. Kurmi** **Vinay P. Namboodiri**
Indian Institute of Technology, Kanpur
{badri,sandepkr,vinodkk,vinaypn}@iitk.ac.in

Abstract

In this **Supplementary Material**, we provide details regarding the experimental setup used while training the proposed method and details about the datasets used. We also provide detailed explanation for variants of the proposed methods for generating natural question based on the image. We further provide additional results for the different variants used. We give the pseudocode for our method and also explain different fusion methods used in the Mixture module.

1 Introduction

Section 3 will provide details about training configuration for MDN, Section 4 will explain the various Proposed Methods and we also provide a discussion in section 2 regarding some important questions related to our method. We report BLEU1, BLEU2, BLEU3, BLEU4, METEOR, ROUGE and CIDER metric scores for VQG-COCO dataset. We present different experiments with Tag Net in which we explore the performance of various tags (Noun, Verb, and Question tags) and different ways of combining them to get the context vectors.

2 Discussion

2.1 How are exemplars improving Embedding

In Multimodel differential network, we use exemplars and train them using a triplet loss. It is known that using a triplet network, we can learn a representation that accentuates how the image is closer to a supporting exemplar as against the opposing exemplar (Hoffer and Ailon, 2015; Frome et al., 2007). The Joint embedding is obtained between the image and language representations. Therefore the improved representation helps in obtaining an improved context vector. Further we show that this also results in improving VQG.

Context	Meth	BLEU-1	Meteor	Rouge	CIDer
Image	-	23.2	8.6	25.6	18.8
Caption	-	23.5	8.6	25.9	24.3
Tag-n	JM	22.2	10.5	22.8	50.1
Tag-n	AtM	22.4	8.6	22.5	20.8
Tag-n	HM	**24.8**	**10.6**	24.4	**53.2**
Tag-n	AM	24.4	10.6	23.9	49.4
Tag-v	JM	23.9	10.5	24.1	**52.9**
Tag-v	AtM	22.2	8.6	22.4	20.9
Tag-v	HM	**24.5**	**10.7**	24.2	**52.3**
Tag-v	AM	24.6	10.6	24.1	49.0
Tag-wh	JM	22.4	10.5	22.5	48.6
Tag-wh	AtM	22.2	8.6	22.4	20.9
Tag-wh	HM	**24.6**	**10.8**	24.3	**55.0**
Tag-wh	AM	24.0	10.4	23.7	47.8

Table 1: Analysis of different Tags for VQG-COCO-dataset. We analyse noun tag (Tag-n), verb tag (Tag-v) and question tag (Tag-wh) for different fusion methods namely joint, attention, Hadamard and addition based fusion.

Context	BLEU-1	Meteor	Rouge	CIDer
Tag-n3-add	22.4	9.1	22.2	26.7
Tag-n3-con	**24.8**	10.6	24.4	**53.2**
Tag-n3-joint	22.1	8.9	21.7	24.6
Tag-n3-conv	24.1	10.3	24.0	47.9
Tag-v3-add	24.1	10.2	23.9	46.7
Tag-v3-con	24.5	10.7	24.2	**52.3**
Tag-v3-joint	22.5	9.1	22.1	25.6
Tag-v3-conv	23.2	9.0	24.2	38.0
Tag-q3-add	24.5	10.5	24.4	51.4
Tag-q3-con	24.6	**10.8**	24.3	**55.0**
Tag-q3-joint	22.1	9.0	22.0	25.9
Tag-q3-conv	24.3	10.4	24.0	48.6

Table 2: Combination of 3 tags of each category for hadamard mixture model namely addition, concatenation, multiplication and 1d-convolution

2.2 Are exemplars required?

We had similar concerns and validated this point by using random exemplars for the nearest neighbor for MDN. (k=R in table 5) In this case the method is similar to the baseline. This suggests that with random exemplar, the model learns to ignore the cue.

Figure 1: These are some more examples from the VQG-COCO dataset which provide a comparison between the questions generated by our model and human annotated questions. (b) is the human annotated question for the first row-fourth column, & fifth column image and (a) for the rest of images.

2.3 Are captions necessary for our method?

This is not actually necessary. In our method, we have used an existing image captioning method (Karpathy and Fei-Fei, 2015) to generate captions for images that did not have them. For VQG dataset, captions were available and we have used that, but, for VQA dataset captions were not available and we have generated captions while training. We provide detailed evidence with respect to caption-question pairs to ensure that we are generating novel questions. While the caption generates scene description, our proposed method generates semantically meaningful and novel questions. Examples for Figure 1 of main paper: First Image:- Caption- A young man skateboarding around little cones. Our Question- Is this a skateboard competition? Second Image:- Caption- A small child is standing on a pair of skis. Our Question:- How old is that little girl?

2.4 Sampling Exemplar: KNN vs ITML

Our method is aimed at using efficient exemplar-based retrieval techniques. We have experimented with various exemplar methods, such as ITML (Davis et al., 2007) based metric learning for image features and KNN based approaches. We observed KNN based approach (K-D tree) with Euclidean metric is a efficient method for finding exemplars. Also we observed that ITML is computationally expensive and also depends on the training procedure. The table provides the experimental result for Differential Image Network variant with k (number of exemplars) = 2 and Hadamard method:

Meth	Exemplar	BLEU-1	Meteor	Rouge	CIDer
KNN	IE(K=2)	23.2	8.9	**27.8**	22.1
ITML	IE(K=2)	22.7	9.3	24.5	22.1

Table 3: VQG-COCO-dataset, Analysis of different methods of finding Exemplars for hadamard model. ITML vs KNN based methods. We see that both give more or less similar results but since ITML is computationally expensive and the dataset size is also small, it is not that efficient for our use. All these experiment are for the differential image network for K=2 only.

2.5 Question Generation approaches: Sampling vs Argmax

We obtained the decoding using standard practice followed in the literature (Sutskever et al., 2014). This method selects the argmax sentence. Also, we evaluated our method by sampling from the probability distributions and provide the results for our proposed MDN-Joint method for VQG dataset as follows:

Meth	BLEU-1	Meteor	Rouge	CIDer
Sampling	17.9	11.5	20.6	22.1
Argmax	36.0	23.4	41.8	50.7

Table 4: VQG-COCO-dataset, Analysis of question generation approaches:sampling vs Argmax in MDN-Joint model for K=5 only. We see that Argmax clearly outperforms the sampling method.

Meth	Exemplar	BLEU-1	Meteor	Rouge	CIDer
AM	IE(K=1)	21.8	7.6	22.8	22.0
AM	IE(K=2)	22.4	8.3	23.4	16.0
AM	IE(K=3)	22.1	8.8	24.7	24.1
AM	IE(K=4)	23.7	9.5	**25.9**	25.2
AM	IE(K=5)	**24.4**	**11.7**	25.0	27.8
AM	IE(K=R)	18.8	6.4	20.0	20.1
HM	IE(K=1)	23.6	7.2	25.3	21.0
HM	IE(K=2)	23.2	8.9	**27.8**	22.1
HM	IE(K=3)	24.8	9.8	27.9	28.5
HM	IE(K=4)	27.7	9.4	26.1	**33.8**
HM	IE(K=5)	**28.3**	**10.2**	26.6	31.5
HM	IE(K=R)	20.1	7.7	20.1	20.5
JM	IE(K=1)	20.1	7.9	21.8	20.9
JM	IE(K=2)	22.6	8.5	22.4	28.2
JM	IE(K=3)	24.0	9.2	24.4	29.5
JM	IE(K=4)	28.7	10.2	24.4	32.8
JM	IE(K=5)	**30.4**	**11.7**	**26.3**	38.8
JM	IE(K=R)	21.8	7.4	22.1	22.5

Table 5: VQG-COCO-dataset, Analysis of different number of Exemplars for addition model, hadamard model and joint model, R is random exemplar. All these experiment are for the differential image network. k=5 performs the best and hence we use this value for the results in main paper.

3 Dataset and Training Details

3.1 Dataset

We conduct our experiments on two types of dataset: VQA dataset (Antol et al., 2015), which contains human annotated questions based on images on MS-COCO dataset. Second one is VQG-COCO dataset based on natural question (Mostafazadeh et al., 2016).

3.1.1 VQA dataset

VQA dataset(Antol et al., 2015) is built on complex images from MS-COCO dataset. It contains a total of 204721 images, out of which 82783 are for training, 40504 for validation and 81434 for testing. Each image in the MS-COCO dataset is associated with 3 questions and each question has 10 possible answers. So there are 248349 QA pair for training, 121512 QA pairs for validating and 244302 QA pairs for testing. We used pre-trained caption generation model (Karpathy et al., 2014) to extract captions for VQA dataset.

3.1.2 VQG dataset

The VQG-COCO dataset(Mostafazadeh et al., 2016), is developed for generating natural and engaging questions that are based on common sense reasoning. This dataset contains a total of 2500 training images, 1250 validation images and 1250 testing images. Each image in the dataset contains 5 natural questions.

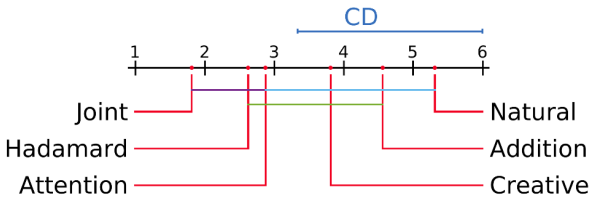

Figure 2: The mean rank of all the models on the basis of BLEU score are plotted on the x-axis. Here Joint refers to our MDN-Joint model and others are the different variations of our model and Natural-(Mostafazadeh et al., 2016), Creative-(Jain et al., 2017). Also the colored lines between two models represent that those models are not significantly different from each other.

3.2 Training Configuration

We have used RMSPROP optimizer to update the model parameter and configured hyper-parameter values to be as follows: learning rate $= 0.0004$, batch size $= 200, \alpha = 0.99, \epsilon = 1e-8$ to train the classification network . In order to train a triplet model, we have used RMSPROP to optimize the triplet model model parameter and configure hyper-parameter values to be: learning rate $= 0.001$, batch size $= 200, \alpha = 0.9, \epsilon = 1e-8$. We also used learning rate decay to decrease the learning rate on every epoch by a factor given by:

$$Decay\_factor = exp\left(\frac{log(0.1)}{a * b}\right)$$

where values of a=1500 and b=1250 are set empirically.

4 Ablation Analysis

While, we advocate the use of multimodal differential network (MDN) for generating embeddings that can be used by the decoder for generating questions, we also evaluate several variants of this architecture namely (a) Differential Image Network, (b) Tag net and (c) Place net. These are described in detail as follows:

4.1 Differential Image Network

For obtaining the exemplar image based context embedding, we propose a triplet network con-

Context	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr
Natural2016	19.2	-	-	-	19.7	-	-
Creative2017	35.6	-	-	-	19.9	-	-
Image Only	20.8	14.1	8.5	5.2	8.6	22.6	18.8
Caption Only	21.1	14.2	8.6	5.4	8.5	25.9	22.3
Tag-Hadamard	24.4	15.1	9.5	6.3	10.8	24.3	55.0
PlaceCNN-Joint	25.7	15.7	9.9	6.5	10.8	24.5	**56.1**
Diff.Image-Joint	30.4	20.1	14.3	8.3	11.7	26.3	38.8
MDN-Joint (**Ours**)	**36.0**	24.9	16.8	10.4	**23.4**	**41.8**	50.7
Humans2016	**86.0**	-	-	-	**60.8**	**-**	-

Table 6: Full State-of-the-Art comparison on VQG-COCO Dataset. The first block consists of the state-of-the-art results, second block refers to the baselines mentioned in State-of-the-art section of main paper and the third block provides the results for the best method for different ablations of our method.

sist of three network, one is target net, supporting net and opposing net. All these three networks designed with convolution neural network and shared the same parameters.

The weights of this network are learnt through end-to-end learning using a triplet loss. The aim is to obtain latent weight vectors that bring the supporting exemplar close to the target image and enhances the difference between opposing examples. More formally, given an image x_i we obtain an embedding g_i using a CNN that we parameterize through a function $G(x_i, W_c)$ where W_c are the weights of the CNN. This is illustrated in figure 3.

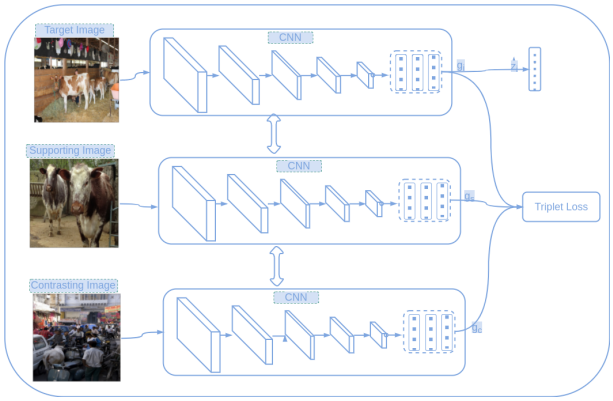

Figure 3: Differential Image Network

4.2 Tag net

The tag net consists of two parts Context Extractor & Tag Embedding Net. This is illustrated in figure 4.

Extract Context: The first step is to extract the caption of the image using NeuralTalk2 (Karpathy et al., 2014) model. We find the part-of-speech(POS) tag present in the caption. POS taggers have been developed for two well known corpuses, the Brown Corpus and the Penn Treebanks. For our work, we are using the Brown Corpus tags. The tags are clustered into three category namely Noun tag, Verb tag and Question tags (What, Where, ...). Noun tag consists of all the noun & pronouns present in the caption sentence and similarly, verb tag consists of verb & adverbs present in the caption sentence. The question tags consists of the 7-well know question words i.e., why, how, what, when, where, who and which. Each tag token is represented as a one-hot vector of the dimension of vocabulary size. For generalization, we have considered 5 tokens from each category of the Tags.

Tag Embedding Net: The embedding network consists of word embedding followed by temporal convolutions neural network followed by max-pooling network. In the first step, sparse high dimensional one-hot vector is transformed to dense low dimension vector using word embedding. After this, we apply temporal convolution on the word embedding vector. The uni-gram, bi-gram and tri-gram feature are computed by applying convolution filter of size 1, 2 and 3 respectability. Finally, we applied max-pooling on this to get a vector representation of the tags as shown figure 4. We concatenated all the tag words followed by fully connected layer to get feature dimension of 512. We also explored joint networks based on concatenation of all the tags, on element-wise addition and element-wise multiplication of the tag vectors. However, we observed that convolution over max pooling and joint concatenation gives

Algorithm 1 Multimodal Differential Network

procedure MDN(x_i)
Finding Exemplars:
  $x_i^+, x_i^- := KD - Tree(x_i)$
  $c_i, c_i^+, c_i^- := Extract\_caption(x_i, x_i^+, x_i^-)$
Compute Triplet Embedding:
  $g_i, g_i^+, g_i^- := Triplet\_CNN(x_i, x_i^+, x_i^-)$
  $f_i, f_i^+, f_i^- := Triplet\_LSTM(c_i, c_i^+, c_i^-)$
Compute Triplet Fusion Embedding :
  $s_i = Triplet\_Fusion(g_i, f_i, Joint)$
  $s_i^+ = Triplet\_Fusion(g_s, f_s, Joint)$
  $s_i^- = Triplet\_Fusion(g_c, f_c, Joint)$
Compute Triplet Loss:
  $Loss\_Triplet = triplet\_loss(s_i, s_i^+, s_i^-)$
Compute Decode Question Sentence:
  $\hat{y} = Generating\_LSTM(s_i, h_i, c_i)$
  $loss = Cross\_Entropy(y, \hat{y})$
end procedure

---

procedure TRIPLET FUSION(g_i,f_i, $flag$)
  g_i:Image feature,14x14x512
  f_i: Caption feature,1x512
Match Dimension:
  $G_{img} = reshape(g_i)$,196x512
  $F_{caps} = clone(f_i)$ 196x512
If flag==Joint Fusion:
  $Ajnt = \tanh(W_{ij}G_{img} \boxdot (W_{cj}F_{cap} + b_j))$
  $S_{emb} = \tanh(W_A A_{jnt} + b_A)$,
  [$\boxdot = *$ (MDN-Mul), $\boxdot = +$ (MDN-Add)]
If flag==Attention Fusion :
  $h_{att} = \tanh(W_I G_{img} \oplus (W_C F_{cap} + b_c))$
  $P_{att} = \text{Softmax}(W_P h_{att} + b_P)$
  $V_{att} = \sum_i P_{att}(i) G_{img}(i)$
  $A_{att} = V_{att} + f_i$
  $S_{emb} = \tanh(W_A A_{att} + b_A)$
Return S_{emb}
end procedure

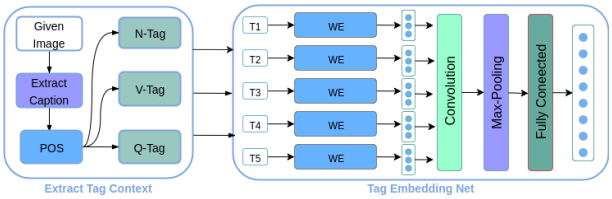

Figure 4: Illustration of Tag Net

better performance based on CIDer score.

$$F_C = \text{Tag\_CNN}(C_t)$$

Where, T_CNN is Temporally Convolution Neural Network applied on word embedding vector with kernel size three.

4.3 Place net

Visual object and scene recognition plays a crucial role in the image. Here, places in the image are labeled with scene semantic categories(Zhou et al., 2017), comprise of large and diverse type of environment in the world, such as (amusement park, tower, swimming pool, shoe shop, cafeteria, rain-forest, conference center, fish pond, etc.). So we have used different type of scene semantic categories present in the image as a place based context to generate natural question. A place365 is a convolution neural network is modeled to classify 365 types of scene categories, which is trained on the place2 dataset consist of 1.8 million of scene images. We have used a pre-trained VGG16-places365 network to obtain place based context embedding feature for various type scene categories present in the image. The context feature F_C is obtained by:

$$F_C = w * \text{p\_conv}(I) + b$$

Where, p_conv is Place365_CNN. We have extracted $conv5$ features of dimension 14x14x512 for attention model and FC8 features of dimension 365 for joint, addition and hadamard model of places365. Finally, we use a linear transformation to obtain a 512 dimensional vector.

We explored using the CONV5 having feature dimension 14x14 512, FC7 having 4096 and FC8 having feature dimension of 365 of places365.

5 Analysis of Tag Net

5.1 Analysis of Tag Context

Tag is language based context. These tags are extracted from caption, except question-tags which

is fixed as the 7 'Wh words' (What, Why, Where, Who, When, Which and How). We have experimented with Noun tag, Verb tag and 'Wh-word' tag as shown in tables. Also, we have experimented in each tag by varying the number of tags from 1 to 7. We combined different tags using 1D-convolution, concatenation, and addition of all the tags and observed that the concatenation mechanism gives better results.

As we can see in the table 1 that taking Nouns, Verbs and Wh-Words as context, we achieve significant improvement in the BLEU, METEOR and CIDEr scores from the basic models which only takes the image and the caption respectively. Taking Nouns generated from the captions and questions of the corresponding training example as context, we achieve an increase of 1.6% in Bleu Score and 2% in METEOR and 34.4% in CIDEr Score from the basic Image model. Similarly taking Verbs as context gives us an increase of 1.3% in Bleu Score and 2.1% in METEOR and 33.5% in CIDEr Score from the basic Image model. And the best result comes when we take 3 Wh-Words as context and apply the Hadamard Model with concatenating the 3 WH-words.
Also in Table 2 we have shown the results when we take more than one words as context. Here we show that for 3 words i.e 3 nouns, 3 verbs and 3 Wh-words, the Concatenation model performs the best. In this table the conv model is using 1D convolution to combine the tags and the joint model combine all the tags.

5.2 Analysis of Context: Exemplars

In Multimodel Differential Network and Differential Image Network, we use exemplar images(target, supporting and opposing image) to obtain the differential context. We have performed the experiment based on the single exemplar(K=1), which is one supporting and one opposing image along with target image, based on two exemplar(K=2), i.e. two supporting and two opposing image along with single target image. similarly we have performed experiment for K=3 and K=4 as shown in table- 5.

6 Mixture Module: Other Variations

Hadamard method uses element-wise multiplication whereas Addition method uses element-wise addition in place of the concatenation operator of the Joint method. The Hadamard method finds a correlation between image feature and caption feature vector while the Addition method learns a resultant vector. In the attention method, the output S_i is the weighted average of attention probability vector P_{att} and convolutional features G_{img}. The attention probability vector weights the contribution of each convolutional feature based on the caption vector. This attention method is similar to work stack attention method (Yang et al., 2016). The attention mechanism is given by:

$$
\begin{aligned}
h_{att} &= \tanh(W_I G_{img} \oplus (W_C F_{cap} + b_c)) \\
P_{att} &= \text{Softmax}(W_P^T h_{att} + b_P) \\
V_{att} &= \sum_i P_{att}(i) G_{img}(i) \\
A_{att} &= V_{att} + f_i \\
s_i &= \tanh(W_A A_{att} + b_A)
\end{aligned}
\tag{1}
$$

where G_{img} is the 14x14x512-dimensional convolution feature map from the fifth convolution layer of VGG-19 Net of image X_i and f_i is the caption context vector. The attention probability vector P_{att} is a 196-dimensional vector. W_I, W_C, W_P are the weights and b_c, b_A, b_c is the bias for different layers. We evaluate the different approaches and provide results for the same. Here $\oplus$ represents element-wise addition.

6.1 Evaluation Metrics

Our task is similar to encoder -decoder framework of machine translation. we have used same evaluation metric is used in machine translation. BLEU(Papineni et al., 2002) is the first metric to find the correlation between generated question with ground truth question. BLEU score is used to measure the precision value, i.e That is how much words in the predicted question is appeared in reference question. BLEU-n score measures the n-gram precision for counting co-occurrence on reference sentences. we have evaluated BLEU score from n is 1 to 4. The mechanism of ROUGE-n(Lin, 2004) score is similar to BLEU-n,where as, it measures recall value instead of precision value in BLEU. That is how much words in the reference question is appeared in predicted question.Another version ROUGE metric is ROUGE-L, which measures longest common sub-sequence present in the generated question. METEOR(Banerjee and Lavie, 2005) score is another useful evaluation metric to calculate the similarity between generated question with reference

one by considering synonyms, stemming and paraphrases. the output of the METEOR score measure the word matches between predicted question and reference question. In VQG, it compute the word match score between predicted question with five reference question. CIDer(Vedantam et al., 2015) score is a consensus based evaluation metric. It measure human-likeness, that is the sentence is written by human or not. The consensus is measured, how often n-grams in the predicted question are appeared in the reference question. If the n-grams in the predicted question sentence is appeared more frequently in reference question then question is less informative and have low CIDer score. We provide our results using all these metrics and compare it with existing baselines.

References

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. 2007. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.

Andrea Frome, Yoram Singer, Fei Sha, and Jitendra Malik. 2007. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE.

Elad Hoffer and Nir Ailon. 2015. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out:Proceedings of the ACL-04 workshop*.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1802–1813.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pages 3104–3112.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.