Supplementary Material for
Crossing the Gap: Domain Generalization for Image Captioning

Yuchen Ren[12], Zhendong Mao[13] [*], Shancheng Fang[1], Yan Lu[2], Tong He[2],
Hao Du[1], Yongdong Zhang[13] and Wanli Ouyang[2]
[1]University of Science and Technology of China, Hefei, China
[2]Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

1. Method details

In this section, we describe in detail some modules that bridge the domain gap, including good-hard negative samples mining, memory bank, momentum update and the way to measure domain gaps.

1.1. Bridging domain gaps

To bridge domain gaps for image captioning, we use an encoder-decoder framework to translate the input image into a natural language description.

For each input source domain image $\mathbf{I}_i^j$, we extract a set of image regions [1] $\mathbf{X}_i^j$ by a pre-trained Faster-RCNN [14]. We use Transformer [16] based refined encoder to refine visual features. Through n layers of attention, we get refined visual features $\mathbf{V}$. The decoder is similar to the refined encoder, using a multi-layer multi-head attention mechanism, but the attention mechanism is based on cross-attention with refined visual features.

1.1.1 Good-hard negative samples mining

We choose the '**good-hard**' negative samples with the smaller sim_t of the anchor sample as the good-hard negative sample, and our language-guided triplet sampling can easily implement this strategy. As shown in Fig. 1, we only need to set an upper bound n_u and a lower bound n_l for the negative samples sim_t.

1.1.2 Memory Bank

A large set of negative samples can have an important impact on feature learning, as revealed by recent work [3, 7, 20]. However, the number of negative samples is limited by the mini-batch size. Especially when mining hard negative samples, it is difficult to be efficient. Inspired by [7, 20], we employ the external memories as a bank to store the visual feature representations v^k by the encoder (or linguistic feature representations by the decoder), and corresponding caption c^k, the memory bank Mem with a size of mbs can be formulated as

$$Mem = \{M^1, \cdots, M^s | M^k = (v^k, c^k)\}. \quad (1)$$

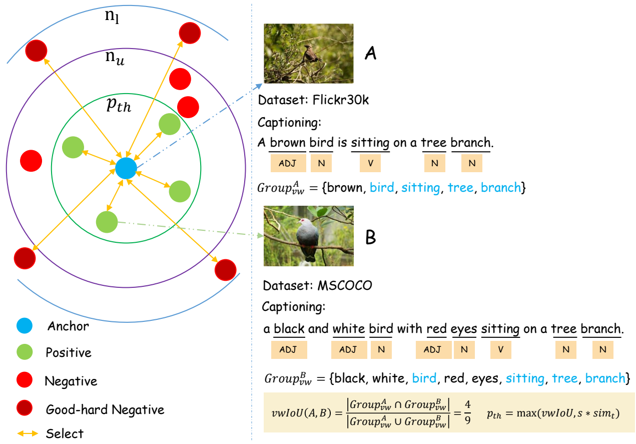

n_l
n_u
p_{th}
A
Dataset: Flickr30k
Captioning:
A brown bird is sitting on a tree branch.
ADJ
N
V
N
N
$Group_{vw}^A$ = {brown, bird, sitting, tree, branch}
B
Dataset: MSCOCO
Captioning:
a black and white bird with red eyes sitting on a tree branch.
ADJ
ADJ
N
ADJ
N
V
N
N
$Group_{vw}^B$ = {black, white, bird, red, eyes, sitting, tree, branch}
$vwIoU(A,B) = \frac{|Group_{vw}^A \cap Group_{vw}^B|}{|Group_{vw}^A \cup Group_{vw}^B|} = \frac{4}{9}$
$p_{th} = \max(vwIoU, s * sim_t)$
Anchor
Positive
Negative
Good-hard Negative
Select

Figure 1. The schematic diagram of language-guided triplet sampling. We show a circle centered on the anchor. The smaller the $1/sim_t$ of the sample and the anchor, the closer to the anchor in the circle.

To maintain the memory bank, we add the newest feature representations and corresponding visual words group while removing the oldest content for each batch.

1.1.3 Momentum update

After we construct the memory bank, we can improve the efficiency of metric learning through better negative example mining. However, it still suffers from non-synchronization of feature representations because of the rapidly-changed encoder. Following [7], we use a momentum update by a momentum encoder to feed visual feature representations to the memory bank. The momentum encoder parameter θ_{me} is updated as the encoder parameter θ_e is optimized by a back-propagation of the loss function following

$$\theta_{me} = m * \theta_{me} + (1 - m) * \theta_e, \quad (2)$$

[*]Corresponding authors.

Dataset	Domain	# Images	# Caps per image	Caps length
MSCOCO [11]	Common	132 K	5	10.5
VizWiz [6]	Assistive	70 K	5	13.0
Flickr30K [21]	Social	31 K	5	12.4
CUB-200 [19]	Avian	12 K	10	15.2
Oxford-102 [12]	Floral	8 K	10	14.1

Table 1. Statistics of five domains in DGIC.

where $m \in [0, 1)$ is a momentum coefficient to adjust the updated degree, and θ_{me} evolves more smoothly than θ_e. In addition to the encoder, we can also use momentum update to the decoder to feed text features to memory for metric learning. We find in the ablation experiment that the features of the encoder and decoder are very different, and the effect of the encoder is better than that of the decoder.

1.2. Measuring the domain gaps for DGIC

As introduced in the main paper, we measure domain gaps by using the Maximum Mean Discrepancy (MMD) [5] theory. The MMD distance between domains $\mathcal{D}^S$ and $\mathcal{D}^T$ can be measured according to the following equation:

$$\mathrm{MMD}(\mathcal{D}^S, \mathcal{D}^T)^2 = \|\mu_{\mathbb{P}_s} - \mu_{\mathbb{P}_t}\|^2_{\mathcal{H}}\,, \tag{3}$$

where $\mu_{\mathbb{P}_s} := \mathbb{E}_{\mathbf{s}\sim\mathcal{D}^S}[\phi(\mathbf{s})]$ and $\mu_{\mathbb{P}_t} := \mathbb{E}_{\mathbf{t}\sim\mathcal{D}^T}[\phi(\mathbf{t})]$ are the samples projected in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, and $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathcal{H}$ represents a mapping operation. We can use the kernel embedding technique to represent the arbitrary distribution when the kernel $k(\cdot, \cdot)$ meet to be characteristic, which the mapping to the RKHS $\mathcal{H}$ is injective [15]. Then we employ a kernel function $k(\mathbf{s}_i, \mathbf{t}_j) = \phi(\mathbf{s}_i)\phi(\mathbf{t}_j)^\intercal$ induced by $\phi(\cdot)$, and the MMD distance can be reformulated as:

$$\mathrm{MMD}(\mathcal{D}^S, \mathcal{D}^T)^2 = \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s} k(\mathbf{s}_i, \mathbf{s}_j) + \tag{4}$$
$$\frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t} k(\mathbf{t}_i, \mathbf{t}_j) - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(\mathbf{s}_i, \mathbf{t}_j),$$

where n_s, n_t represent sample size in the source and target domains, respectively.

In this work, we use the RBF characteristic kernel to compute both visual and linguistic features with $k(\mathbf{s}_i, \mathbf{t}_j) = exp(-\frac{1}{2\alpha}\|\mathbf{s}_i - \mathbf{t}_j\|^2)$ and $\alpha = 1; 5; 10$.

2. Experiments and analysis

In this section, we show more details about the DGIC benchmark, implementation details and ablation experiments.

2.1. The DGIC Benchmark

2.1.1 Datasets

The DGIC benchmark consists of 253K images and 1,365K captions, sourced from MSCOCO [11], VizWiz [6], Flickr30K [21], CUB-200 [13, 19] and Oxford-102 [12, 13]. The statistics of these datasets are shown in Tab. 1. The details about datasets are below:

MSCOCO. The MSCOCO dataset [11] is a widely used dataset and covers the common domain. It contains more than 120,000 images in total, and each of these images comes with at least 5 human-annotated sentences as the ground truth captions. We follow the data split provided by Karpathy *et al.* [8], where the training, validation, and test splits include 113,287, 5,000, and 5,000 images, respectively.

VizWiz. The VizWiz dataset [6] consists of 39,181 images originating from people who are blind. The VizWiz is the specific assistive domain dataset towards assistive technologies, and each of these images is paired with 5 ground-truth captions. Similar to [8], we divide it into 23,429 images for training, 3,750 images for validation, and 4,000 images for test.

Flickr30k. The Flickr30k [21] dataset is used as the social domain. There are 31,783 images in this dataset, and each image is also annotated with 5 sentences. We adapt the data split from [8], where 29,000 images are used for training, 1,014 images are used for validation, and 1,000 images are used for test.

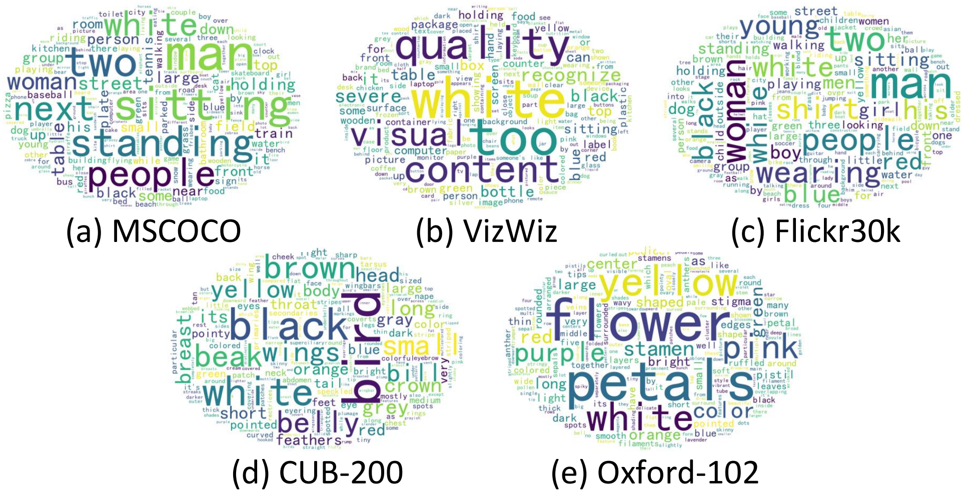
CUB-200. The CUB-200 dataset [19] is the avian domain dataset. It includes 11,788 images of birds in total, which is paired with 10 sentence annotations for each image by Reed *et al.* [13]. Similar to [8], we divide it into 9,788 images for training, 1,000 images for validation, and 1,000 images for test.

Oxford-102. The Oxford-102 dataset [12] is the floral domain dataset. It includes 8,189 images of flowers in total, which is paired with 10 sentence annotations for each image by Reed *et al.* [13]. Similar to [8], we divide it into 6,189 images for training, 1,000 images for validation, and 1,000 images for test.

We also illustrate the differences in label spaces among MSCOCO, VizWiz, Flickr30k, CUB-200, Oxford-102 by using word clouds (see Fig. 2).

2.1.2 Implementation details

For input image representations, we obtain a 2048-D Bottom-up [1] feature for each region with a pre-trained Faster-RCNN [14]. We set the dimensionality of each layer to 512, the number of encoder and decoder layers to 6, and the number of heads to 8. To achieve language-guided semantic metric learning, we set the sentence scaling factor s to 0.29, the positive threshold value p_{th} to 0.2, the upper bound n_u of negative sampling to 0.02, the lower bound n_l of negative sampling to 0.01, the margin δ of the triplet loss to 2, the cross-entropy loss weight α to 1, the inter-domain metric learning weight β to 0.01, the intra-domain metric learning weight γ to 0.01, the size mbs of the memory bank to 800, and the momentum coefficient m of mo-



(a) MSCOCO (b) VizWiz (c) Flickr30k

(d) CUB-200 (e) Oxford-102

Figure 2. Word clouds for five datasets of DGIC.

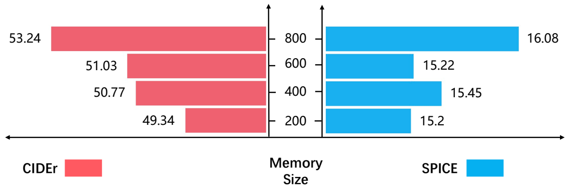


Figure 3. Comparison of different memory bank size.

m	0.999	0.9	0.8
CIDEr	**53.24**	52.00	51.61
SPICE	**16.08**	16.04	15.81

Table 2. Comparison of different momentum coefficient m.

mentum update to 0.99. We use the Adam optimizer [9] to train our model with the learning rate defined as 0.05. We use 20,000 warmup steps, and use a batch size of 50. We train the model with cross-entropy loss and triplet loss for 50 epochs. To implement our framework, we utilize the PyTorch library on NVIDIA A100 GPUs. We prune the vocabulary by dropping words with a frequency of less than 5 and adding special Unkown (UNK), Begin-Of-Sentence (BOS), and End-Of-Sentence (EOS) tokens. In order to simplify the implementation, we employ a joint vocabulary containing words in both the source domains and the target domain similar to [22].

2.1.3 Parameter analysis

Memory bank size. Fig. 3 shows that our method has better performance with a large memory bank size, but a large memory bank size will lead to the computational burden due to the calculation of triplet loss through visual features and captions. Therefore, we need to make a trade-off between performance and computational cost. So, we set our memory bank size as 800 in our implementation.

Momentum coefficient. Tab. 2 shows that our method has better performance with a large momentum coefficient due to the better feature consistency. Therefore, we set momentum coefficient as 0.999 in our implementation.

Dataset	Domain	# Images	# Caps per image	Caps length
COCOfog-no [11]	No Fog	132 K	5	10.5
COCOfog-mild	Mild Fog	132 K	5	10.5
COCOfog-normal	Normal Fog	132 K	5	10.5
COCOfog-heavier	Heavier Fog	132 K	5	10.5
COCOfog-severe	Severe Fog	132 K	5	10.5

Table 3. Statistics of fog sub-benchmark in DGIC.

Dataset	Domain	# Images	# Caps per image	Caps length
COCOro0 [11]	0 Degree Rotation	132 K	5	10.5
COCOro45	45 Degree Rotation	132 K	5	10.5
COCOro60	60 Degree Rotation	132 K	5	10.5
COCOro75	75 Degree Rotation	132 K	5	10.5
COCOro90	90 Degree Rotation	132 K	5	10.5

Table 4. Statistics of rotation sub-benchmark in DGIC.

3. Constructing synthetic sub-benchmarks

In this section, we propose two synthetic datasets to further explore domain generalization for image captioning.

3.1. Datasets

Domain generalization can be divided into homogeneous and heterogeneous depending on whether the different domain label spaces are identical or not [10,23]. In order to explore the homogeneous captioning setting, we propose two synthetic sub-benchmarks for DGIC. Each sub-benchmark is composed of five different levels of synthetic datasets, similar to Rotated MNIST [4] and the synthetic semantic segmentation dataset [17] for DG. When compositing different datasets for DGIC, we ensure that the synthetic operation does not change the caption semantics of the image (*e.g.*, partial cropping, puzzle transformation, color style transfer do not meet the conditions), and we use the Albumentations [2] to synthesize two sub-benchmarks of fog and rotation based on MSCOCO. The statistics of fog sub-benchmark and rotation sub-benchmark are shown in Tab. 3 and Tab. 4.

Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	no+mild+normal+heavier →very severe fog	28.69	14.65	7.87	4.58	10.23	26.83	39.07	10.47
LSML		**29.45**	**15.09**	**8.20**	**4.75**	**10.58**	**27.41**	**40.55**	**10.78**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	no+mild+normal+very severe→heavier fog	30.03	15.66	8.70	5.17	10.97	28.08	44.76	11.58
LSML		**30.21**	**16.05**	**8.96**	**5.28**	**11.14**	**28.49**	**46.55**	**12.20**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	no+mild+heavier+very severe→normal fog	31.97	17.30	9.80	5.84	12.21	29.44	52.94	14.40
LSML		**32.36**	**17.85**	**10.33**	**6.28**	**12.44**	**30.01**	**56.89**	**14.73**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	no+normal+heavier+very severe→mild fog	**34.53**	19.76	11.61	7.14	13.78	31.89	65.58	16.94
LSML		34.42	**19.88**	**11.86**	**7.36**	**13.82**	**32.05**	**67.82**	**17.39**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	mild+normal+heavier+ very severe→no fog	36.19	21.23	12.67	7.94	14.71	33.45	74.13	18.90
LSML		**36.30**	**21.56**	**13.08**	**8.29**	**14.88**	**33.70**	**76.06**	**19.22**

Table 5. Comparison with State-of-the-Arts domain generalization methods on five synthetic fog datasets on the DGIC sub-benchmark. The performance is evaluated by BLEU1-4, METEOR, ROUGE, CIDEr and SPICE.

Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	0+45+60+75→90 degree rotation	35.32	20.57	12.33	7.72	14.55	32.99	73.45	18.80
LSML		**35.41**	**20.72**	**12.48**	**7.85**	**14.60**	**33.28**	**75.59**	**18.91**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	0+45+60+90→75 degree rotation	**35.75**	20.94	12.56	7.83	14.58	33.44	74.62	18.65
LSML		35.70	**21.05**	**12.84**	**8.14**	**14.72**	**33.63**	**76.76**	**19.28**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	0+45+75+90→60 degree rotation	34.17	19.61	11.60	7.19	13.87	32.07	67.85	17.56
LSML		**34.44**	**19.91**	**11.85**	**7.33**	**13.94**	**32.41**	**68.94**	**17.68**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	0+60+75+90→45 degree rotation	34.25	19.53	11.46	7.02	13.83	32.15	67.39	17.33
LSML		**34.71**	**20.08**	**11.88**	**7.40**	**14.11**	**32.74**	**69.75**	**17.93**
Method	Source→Target	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE
EISNet	45+60+75+90→0 degree rotation	35.19	20.42	12.04	7.32	**14.42**	32.89	69.84	18.43
LSML		**35.35**	**20.57**	**12.34**	**7.68**	14.40	**33.05**	**72.05**	**18.95**

Table 6. Comparison with State-of-the-Arts domain generalization methods on five synthetic rotation datasets on the DGIC sub-benchmark.

3.2. Experiments

To demonstrate the performance of our method, we compare EISNet [18] with our proposed method with the same backbone (Transformer [16]) on these synthetic sub-benchmarks: (1) MSCOCO-Rotation (0, 45, 60, 75, and 90 degrees); (2) MSCOCO-Fog (no fog, mild fog, normal fog, heavier fog, and very severe fog). In terms of implementation, EISNet uses the same source from different domains as positive samples, and our LSML uses only inter-domain metric learning and visual words-guided triplet sampling to demonstrate the importance of discriminative features for cross-domain semantic learning and to prove the importance of visual words even in the case where there is almost no domain gap in the language. Tab. 5 and Tab. 6 show the better effectiveness of our language-guided semantic metric learning.

References

[1] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR (2018) 1, 2

[2] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information (2020) 3

[3] Chen, X., Fan, H., Girshick, R., He, K.: Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297 (2020) 1

[4] Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D.: Domain generalization for object recognition with multi-task autoencoders. In: ICCV (2015) 3

[5] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A.: A kernel method for the two-sample-problem. In: NeurIPS (2006) 2

[6] Gurari, D., Zhao, Y., Zhang, M., Bhattacharya, N.: Captioning Images Taken by People Who Are Blind. In: ECCV (2020) 2

[7] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: CVPR (2020) 1

[8] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015) 2

[9] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. In: ICLR (2015) 3

[10] Li, Y., Yang, Y., Zhou, W., Hospedales, T.: Feature-critic networks for heterogeneous domain generalization. In: ICML (2019) 3

[11] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: ECCV (2014) 2, 3

[12] Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008) 2

[13] Reed, S., Akata, Z., Lee, H., Schiele, B.: Learning deep representations of fine-grained visual descriptions. In: CVPR (2016) 2

[14] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks **39**(6), 1137–1149 (2017) 1, 2

[15] Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Lanckriet, G.R., Schölkopf, B.: Kernel choice and classifiability for rkhs embeddings of probability distributions. In: NeurIPS (2009) 2

[16] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 1, 4

[17] Volpi, R., Namkoong, H., Sener, O., Duchi, J., Murino, V., Savarese, S.: Generalizing to unseen domains via adversarial data augmentation. In: NeurIPS (2018) 3

[18] Wang, S., Yu, L., Li, C., Fu, C.W., Heng, P.A.: Learning from extrinsic and intrinsic supervisions for domain generalization. In: ECCV (2020) 4

[19] Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. rep., California Institute of Technology (2010) 2

[20] Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: CVPR (2018) 1

[21] Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL (2014) 2

[22] Zhao, W., Wu, X., Luo, J.: Cross-domain image captioning via cross-modal retrieval and model adaptation. IEEE TIP (2020) 3

[23] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 3