A NEIGHBOR-AWARE APPROACH FOR IMAGE-TEXT MATCHING

Chunxiao Liu[1,2], *Zhendong Mao*[1,2] *, *Wenyu Zang*[1,2], *Bin Wang*[1,2,3]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Xiaomi AI Lab
{*liuchunxiao*, *maozhendong*, *zangwenyu*}@iie.ac.cn, wangbin11@xiaomi.com

ABSTRACT

Image-text matching has received a large amount of interest since it associates different modalities and improves the understanding of image and natural language. It aims to retrieval semantic related images based on the given text query, and vice versa. Existing approaches have achieved much progress by projecting the image and text into a common space where data with different semantics can be distinguished. However, they process all the data points uniformly, while neglecting that data in a neighborhood are harder to distinguish due to their visual similarity or syntactic structural similarity. To address this issue, we propose a neighbor-aware network to image-text matching where an intra-attention module and neighbor-aware ranking loss are proposed to jointly distinguish data with different semantics, more importantly, semantic unrelated data in a neighborhood can be distinguished. The intra-attention attends to discriminative parts by comparing data with different semantics and magnifying difference between them, especially subtle difference between data in a neighborhood. The neighbor-aware ranking loss function utilizes the magnified difference to explicitly and effectively discriminate data in a neighborhood. We conduct extensive experiments on several benchmarks and show that the proposed approach significantly outperforms the state-of-the-art.

Index Terms— Image-text matching.

1. INTRODUCTION

Image-text matching has recently attracted much attention in computer vision, which aims to retrieval semantic related images based on the given text query, and vice versa. It associates different modalities and improves the understanding of image and natural language. Images with their corresponding detailed textual descriptions are considered as matched, otherwise mismatched.

Existing approaches can be roughly categorized into many-to-many approaches and one-to-one approaches. Many-to-many approaches[1–6] learn latent alignment between objects in the image and words in the text, which requires external object detection tools pre-trained on large-scale datasets. One-to-one approaches [7–17] learn the correspondence between the whole image and text without external object detection tools. In this work we focus on one-to-one approach. Existing one-to-one approaches typically project the image and text into a latent common space where semantic relationships between different modalities can be measured through distance computation. Previous works employ multiple neural network to improve feature representations such that semantic related data are close to each other, otherwise not, such as

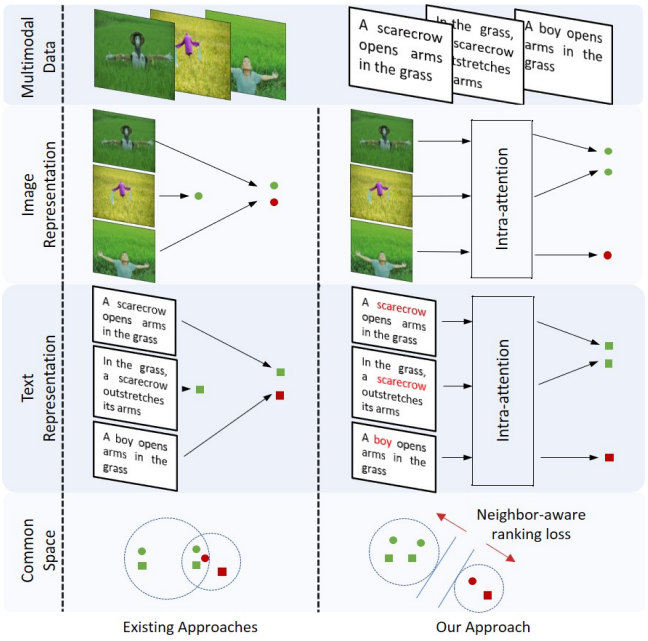

Fig. 1. The above images are visually similar, while the first two images share the same semantic and the third not. In prior works, representation of the first image is closer to the third instead of the second one. The same problem exists in text side, which causes similar semantics being indistinguishable in the common space. Our work modifies that based on intra-attention and distinguish them using neighbor-aware loss.

*Zhendong Mao is the corresponding author.

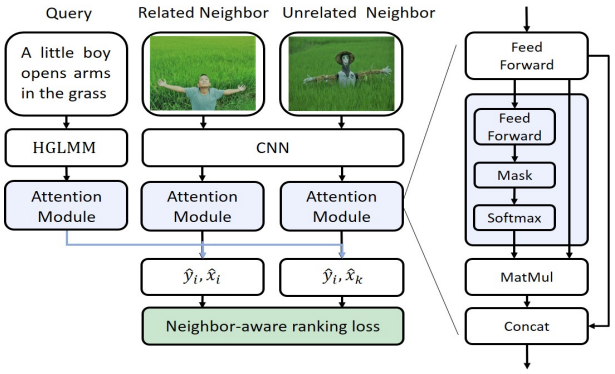

Fig. 2. Our approach mainly consists of intra-attention module and neghbor-aware ranking loss. The intra-attention takes features as input and learns to attend on discriminative parts. The neighbor-aware ranking loss utilizes the attended features to explicitly discriminate neighbors with different semantics.

multimodal convolutional neural networks (m-CNNs) [15], multimodal Recurrent Neural Network (m-RNN) [16], recurrent residual fusion (RRF) [17] and so on. Some others focus on optimization[8, 9, 17, 18]. For example, [8, 9] apply a ranking loss that forces semantic related images to be ranked higher than semantic unrelated images for each text query, so it is when given an image query. Among these works, all the data are processed uniformly whether they are neighbors.

However, data in a neighborhood are harder to distinguish which is neglected by existing approaches. It mainly arises from that data in a neighborhood are similar in content-level (e.g. visual appearance similar for the image and syntactic structure similar for the text) instead of semantic-level, which leads to content similar but semantic different data points becoming neighbors. As shown in Fig 1, the above three images are visually similar, while the first two images are semantic related and the third one is semantic unrelated to them. As a result, the representation of the first image is closer to the third instead of the second, which causes their semantics being indistinguishable and inferior matching results. Therefore, more focus is required to data in a neighborhood to learn more discriminative features since they only differ in subtly parts, which facilitates to distinguish data with different semantics more effectively, see the right of Fig 1.

To address this issue, we propose a neighbor-aware network that consists of an intra-attention module and a neighbor-aware ranking loss to jointly learn more discriminative features. It enables to discriminate data with different semantics, more importantly, semantic unrelated data in a neighborhood can be distinguished that ensures accurate image-text matching. The intra-attention is applied to learn image and text representations respectively by comparing each data and its semantic unrelated neighbors in detail and then magnifies their subtle difference. The neighbor-aware ranking loss emphasizes on neighbors and explicitly distinguish them using the magnified difference. We conduct extensive experiments on several benchmarks, showing that the proposed approach outperforms the state-of-the-art.

The main contributions of our work are listed as follows:

1. We introduce a neighbor-aware network that employs an intra-attention module to magnify difference between data with different semantics, especially data in a neighborhood.

2. We propose a neighbor-aware ranking loss function that explicitly distinguishes neighbors with different semantics.

3. Extensive experiments on benchmarks show that the proposed method significantly outperforms the state-of-the-art for image-text matching.

2. APPROACH

The architecture of neighbor-aware network is illustrated in Fig.2, it mainly consists of an intra-attention module and a neighbor-aware ranking loss that jointly learn the difference between data with different semantics, which will be discussed in section 2.1 and 2.2, respectively.

2.1. Architecture

We first extract image features using pre-trained CNN and text features using fisher vector following [9, 17, 19], denoted as $v_i \in R^{1\times N}$, and $t_i \in R^{1\times M}$. We then generate feature map after a fully connected layer. The intra-attention module is build on the feature map. Here, we briefly introduce this module in the image branch, text branch is similar to it.

The intra-attention module is designed to assign different importance to different features using global information as a reference. We first generate attention mask by performing MLP (Multilayer Perceptron) on the acquired feature map h_i

$$x_i = f(W_i h_i + b_i) \tag{1}$$

where W_i and b_i are parameters to be learned, $f(\cdot)$ denotes multiple nonlinearity transformations including fully connected layers and ReLU activation. x_i is the attention mask for i-th image, which represents the relative importance of each feature. The final attention mask value a_i can be obtained by normalizing this mask into [0,1]

$$a_i = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}} \tag{2}$$

Then, the feature map is reweighted by element-wise product of the feature map and attention mask, which obtains the reweighted feature map h_i^{atten}, that is

$$h_i^{atten} = a_i \odot h_i. \tag{3}$$

Different from conventional approaches, we do not feed the attended map into the next layer directly. Attended feature map is fused with previous feature map by concatenation, which decreases information loss since there might exist bias in attention mask. The fused feature $[h_i; h_i^{atten}]$ is fed into

Table 1. Matchinig results on Flickr8K, the bests are in bold.

Method	Flickr8K					
	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
m-CNN [15]	24.8	53.7	67.1	20.3	47.6	61.7
m-RNN [16]	14.5	37.2	48.5	11.5	31.0	42.4
HM-LSTM [1]	27.7	-	68.6	24.4	-	68.1
DSPE [8]	30.1	60.4	73.7	23.0	51.3	64.8
Ours	**37.2**	**68.1**	**79.1**	**27.7**	**59.6**	**71.8**

a fully connected layer followed by L2 normalization. The distance between image and text in the common space can be computed as $d(\hat{x}_i, \hat{y}_i)$ using Euclidean distance, where $\hat{x}_i$ and $\hat{y}_i$ are image and text representations, respectively.

2.2. Neighbor-aware ranking loss

Different from existing ranking loss [8, 9] applied in previous works, our approach pays more attention to neighbors that belong to different semantics, which will be more effective to distinguish different semantics. It consists of an inter-modal and intra-modal neighbor-aware ranking loss, where one controls the semantic relations between different modalities and another one controls that in the same modality. They will be presented in section 2.1.1 and section 2.2.2.

2.2.1. Inter-modal neighbor-aware ranking loss

Definition: given a data x, its semantic unrelated neighbors are a set of data points whose features are close to the feature of x, its semantic related neighbors are ground truth.

Given a training i-th image, we first extract its feature as v_i. Next, we sample a semantic related image and extract its feature v_j. Then, we sample another k-th semantic unrelated image whose feature is v_k, and calculate the difference of distance between them to v_i, namely

$$n_{ijk} = d(v_i, v_k) - d(v_i, v_j) \tag{4}$$

We then effectively pushing semantic unrelated data away through treating neighbors and non-neighbors separately, which is determined by n_{ijk} value. If it is smaller than a threshold α, k-th image is treated as a semantic unrelated neighbor of i-th image based on our definition. We enlarge distance between i-th and k-th image, forcing it to be larger than that between i-th and j-th image by a fixed margin n.

$$L_1 = \sum_{i,j,k} max\left[0, d(\hat{x_i}, \hat{x_j}) - d(\hat{x_i}, \hat{x_k}) + n\right] \tag{5}$$

where $\hat{x_i}$, $\hat{x_j}$, and $\hat{x_k}$ are representations of the given image, its semantic related and unrelated neighbor in common space.

If n_{ijk} is larger than α, it indicates k-th image is a non-neighbor. We relax its constraint by setting the margin as n_{ijk}, then previous loss L_1 becomes

$$L_1 = \sum_{i,j,k} max\left[0, d(\hat{x_i}, \hat{x_j}) - d(\hat{x_i}, \hat{x_k}) + n_{ijk}\right] \tag{6}$$

Table 2. Matching results on Flickr30K, the bests are in bold.

Method	Flickr30K					
	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
m-CNN [15]	33.6	64.1	74.9	26.2	56.3	69.6
m-RNN [16]	35.4	63.8	73.7	22.8	50.7	63.1
HM-LSTM [1]	38.1	-	76.5	27.7	-	68.8
DSPE [8]	40.3	68.9	79.9	29.7	60.1	72.1
VSE++ [11]	43.7	-	82.1	32.3	-	72.1
RRF [17]	47.6	77.4	87.1	35.4	68.3	79.9
CMPM + CMPC [10]	49.6	76.8	86.1	37.3	65.7	75.5
Ours	**55.1**	**80.3**	**89.6**	**39.4**	**68.8**	**79.9**

Note that the weight of loss for non-neighbors is lower than that for neighbors. When given a training text, it is analogous to L_1. The inter-modal neighbor-aware ranking loss is the combination of restraints on both image and text side.

2.2.2. Intra-modal neighbor-aware ranking loss

Given a training image, we obtain its representation $\hat{x_i}$ in the common space. Let $\hat{y_j}$ and $\hat{y_k}$ denote the representation of one of its semantic related texts and semantic unrelated texts. We enforce the distance from $\hat{x_i}$ to $\hat{y_j}$ to be smaller than that to $\hat{y_k}$ by a margin m ($m >= n$). When given a training text, it is similar to that. The overall intra-modal loss is

$$\begin{aligned} L_2 = &\sum_{i,j,k} max\left[0, d(\hat{x_i}, \hat{y_j}) - d(\hat{x_i}, \hat{y_k}) + m\right] \\ &+ \sum_{i,j,k} max\left[0, d(\hat{y_i}, \hat{x_j}) - d(\hat{y_i}, \hat{x_k}) + m\right] \end{aligned} \tag{7}$$

Note that we emphasis on neighbors by computing this loss of each training data and selecting top 10 to minimize since neighbors are more likely to produce high loss while training. The overall loss function is a weighted combination of inta-modal and inter-modal neighbor-aware ranking loss, where the weight λ is a trade-off.

3. EXPERIMENT

3.1. Experimental setup

Datasets. We evaluate our approach on Flickr8K [20], Flickr30K [21] and MSCOCO [22] for image-to-text and text-to-image matching tasks. For Flickr8K, we employ the standard split that 6k, 1k, 1k images are used for training, validation, and testing. Each image corresponds to five texts. For Flickr30K, we split the benchmark into 29k training images, 1024 validation images and 1k test images following [3, 16, 17]. For MSCOCO, it consists of 83k training images and 41k validation images, we choose 1000 images from training and validation sets for testing following [8, 9]. Five corresponding texts are selected for each image.

Evaluation. We adopt the widely used Recall@K (K = 1,5,10) as evaluation, it is the percentage of queries for which at least one correct result is retrieved among top K results.

Table 3. Ablation study on Flickr8K, Flickr30K and MSCOCO benchmarks, the best results are in bold.

Method	Flickr8K						Flickr30K						MSCOCO					
	Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image			Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VGG + Att	31.2	62.7	73.7	23.2	52.5	65.4	42.8	72.6	82.5	31.3	61.7	72.5	52.1	83.1	91.3	41.0	74.9	86.9
VGG + Loss	31.8	60.8	75.2	23.7	52.4	65.9	44.0	71.9	82.3	31.2	61.7	73.0	53.3	82.0	91.4	42.0	76.0	86.9
VGG + Att + Loss	31.9	62.8	75.8	23.7	52.7	66.0	43.3	74.8	83.0	31.6	61.9	73.7	54.3	83.2	92.4	41.2	75.8	87.2
RES + Att + Loss	**37.2**	**68.1**	**79.1**	**27.7**	**59.6**	**71.8**	**55.1**	**80.3**	**89.6**	**39.4**	**68.8**	**79.9**	**61.3**	**87.9**	**95.4**	**47.0**	**80.8**	**90.1**

Table 4. Matching results on MSCOCO, the bests are in bold.

Method	MSCOCO					
	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
m-CNN [15]	42.8	73.1	84.1	32.6	68.6	82.8
m-RNN [16]	41.0	73.0	83.5	29.0	42.2	77.0
HM-LSTM [1]	43.9	-	87.8	36.1	-	86.7
DSPE [8]	50.1	79.7	89.2	39.6	75.2	86.9
VSE++ [11]	58.3	-	93.3	43.6	-	87.8
RRF [17]	56.4	85.3	91.5	43.9	78.1	88.6
CMPM + CMPC [10]	56.1	86.3	92.9	44.6	78.8	89.0
Ours	**61.3**	**87.9**	**95.4**	**47.0**	**80.8**	**90.1**

Settings. We make comparisons with most representative approaches, where [17] and [8] are the most close works to our approach, [10, 11] are most recent works. Others [1, 15, 16] are benchmarks. We implement the proposed approach in Tensorflow [23], we train the network using Adam optimization, and the mini-batch size is 1500, the learning rate is 0.0001 without decay. The representations in the common space are set to be 512. We also employ dropout with probability 0.5 to avoid overfitting.

Feature extraction. To get visual features, we extract 4096 and 2048 dimensional activations from pre-trained VGGNet19 [24] and ResNet [25] model respectively. To get textual features, we exploit the unsupervised Fisher Vector (FV) approach [26] to extract 6000-dimensional features.

3.2. Experimental results

Comparison with the state-of-the-art. We compare our approach with the state-of-the-art on benchmarks. As shown in Table 1,2,4, our approach marginally outperforms the state-of-the-art on all the benchmarks, which indicates the effectiveness of our approach. It is observed that our method generally achieves more improvement on R@1 than R@10 since semantic unrelated neighbors are pushed away, which benefits to retrieve correct results accurately. It validates the effectiveness of our method in distinguishing data in a neighborhood.

Ablation study. To systematically evaluate the effectiveness of different components, ablation studies are designed as shown in Table 3. Components include: 1) VGG / RES: refers to extract image features using VGGNet or ResNet. 2) Att: represents the proposed intra-attention module. 3) Loss: the proposed neighbor-aware ranking loss. From the table we can obtain the conclusion that either employing the intra-attention or neighbor-aware loss can improve the performance compared with baselines, integrating them can further improve it, which indicate they complement each other. Despite that there is a slight drop when combing the attention module and loss function, the overall performance improves. Note that using ResNet as feature extraction achieves significant improvement since it reserves more semantic information.

Qualitative results. In Fig.3, we offer visualizations on text-to-image matching using our model and DSPE, which is the closest to our work. We retrieve top 1 image for each text query. The four columns correspond text query, DSPE results, our approach results, and ground truth, respectively. It is observed that almost all the incorrect results in this figure are visually similar to the ground truth, which indicates data in a neighborhood that are with different semantics will to some extent impact matching result, and our proposed approach can address this issue by learning more discriminative features.

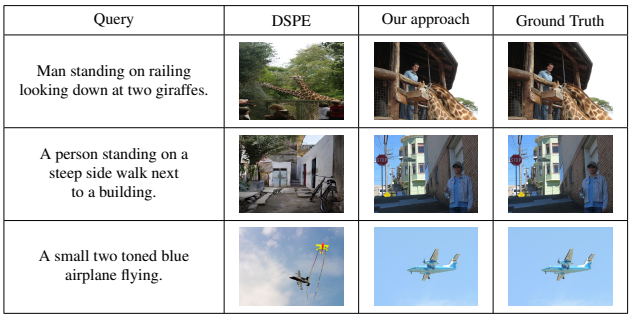

Fig. 3. Qualitative results for text-to-image on MSCOCO.

4. CONCLUSION

In this work, we propose a neighbor-aware network to distinguish different semantic data, especially data in a neighborhood. It employs intra-attention to magnify the subtle difference between neighbors, which is utilized by the neighbor-aware ranking loss to further increase the distance between neighbors that are with different semantics. Experimental results show that the proposed approach will significantly improve the image-text matching performance.

5. ACKNOWLWDGWMENT

This work is supported by the National Natural Science Foundation of China (grants No. 61502477)

References

[1] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua, "Hierarchical multimodal lstm for dense visual-semantic embedding," in *ICCV*, 2017, pp. 1899–1907.

[2] Yan Huang, Wei Wang, and Liang Wang, "Instance-aware image and sentence matching with selective multimodal lstm," pp. 7254–7262, 2016.

[3] Andrej Karpathy, Armand Joulin, and Li Fei-Fei, "Deep fragment embeddings for bidirectional image sentence mapping," vol. 3, pp. 1889–1897, 2014.

[4] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He, "Stacked cross attention for image-text matching," *CoRR*, vol. abs/1803.08024, 2018.

[5] Yan Huang, Qi Wu, and Liang Wang, "Learning semantic concepts and order for image and sentence matching," *CoRR*, vol. abs/1712.02036, 2017.

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh, "Hierarchical question-image co-attention for visual question answering," *CoRR*, vol. abs/1606.00061, 2016.

[7] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *ICML*, 2013, pp. III–1247.

[8] Liwei Wang, Yin Li, and Svetlana Lazebnik, "Learning deep structure-preserving image-text embeddings," in *CVPR*, 2016, pp. 5005–5013.

[9] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik, "Learning two-branch neural networks for image-text matching tasks," *TPAMI*, vol. PP, pp. 1–1, 2017.

[10] Ying Zhang and Huchuan Lu, "Deep cross-modal projection learning for image-text matching," in *ECCV*, 2018, pp. 707–723.

[11] Fartash Faghri, David J. Fleet, Jamie Kiros, and Sanja Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *BMVC*, 2018.

[12] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim, "Dual attention networks for multimodal reasoning and matching," *CVPR*, pp. 2156–2164, 2017.

[13] Yiling Wu, Shuhui Wang, and Qingming Huang, "Learning semantic structure-preserved embeddings for cross-modal retrieval," in *MM*, 2018, pp. 825–833.

[14] Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," *CoRR*, vol. abs/1711.06420, 2017.

[15] Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li, "Multimodal convolutional neural networks for matching image and sentence," in *ICCV*, 2015, pp. 2623–2631.

[16] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille, "Deep captioning with multimodal recurrent neural networks," *CoRR*, vol. abs/1412.6632, 2014.

[17] Yu Liu, Yanming Guo, Erwin M. Bakker, and Michael S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *ICCV*, 2017, pp. 4127–4136.

[18] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *TPAMI*, vol. 36, pp. 521–35, 2014.

[19] Benjamin Eliot Klein, Guy Lev, Gil Sadeh, and Lior Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," *CVPR*, pp. 4437–4446, 2015.

[20] Micah Hodosh, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," in *AAAI*, 2015, pp. 4188–4192.

[21] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," *ICCV*, pp. 2641–2649, 2015.

[22] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.

[23] Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, and Matthieu Devin, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," *CVPR*, pp. 770–778, 2016.

[26] Benjamin Eliot Klein, Guy Lev, Gil Sadeh, and Lior Wolf, "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation," *CoRR*, vol. abs/1411.7399, 2014.