Spatial Quality Aware Network for Video-Based Person Re-identification

Yujie Wang[1], Biao Leng[2(✉)], and Guanglu Song[1]

[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing, China
lengbiao@buaa.edu.cn

Abstract. Person re-identification in video is challenging in computer vision. Most methods adopt feature aggregation to get a video-level representation. However, almost all of them do it on the final feature embedding, which neglects the spatial difference among feature maps. To address this problem, we proposed an effective approach, named Spatial Quality Aware Network (SQAN) for video-based person re-identification. SQAN distributes a score for each pixel in a feature map. Then scores are normalized across all frames and the weighted sum is used to aggregate them. To deal with overfitting, we also proposed a semantic dropout strategy. Experiments show that our proposed method is competitive with state-of-the-art methods in performance.

Keywords: Person re-identification · Deep learning · Feature aggregation

1 Introduction

Person re-identification(re-id), which is widely applicated in smart video surveillance, aims to identify a probe person from a gallery person set via visual information. Most previous works [1–5] focus on image-based re-id, i.e. given a probe person's image, the system should return the most similar person across the gallery person set. Impressive progress has been achieved in the image-based person re-id area. However, in video surveillance scenario, one person's information is encoded not only in individual frames but the correspondence among frames. Empirical evidences [6] confirm that the video-based re-id is superior to the others. However, many challenges still exist.

Due to the length of a video is variable, the feature representation of video is not fixed, which makes the comparison between videos hard. Most methods resort to feature aggregation to build a fix length representation of the video. The direct way to aggregate features is to fetch the max or average value among frame-level features [5], i.e. max/average pool. The max pool only maintains the most salient part of features, while the average pool neglects the importance differences among frame features. These information's loss degrades the robustness of the algorithm.

D. Liu et al. (Eds.): ICONIP 2017, Part III, LNCS 10636, pp. 34–43, 2017.
https://doi.org/10.1007/978-3-319-70090-8_4

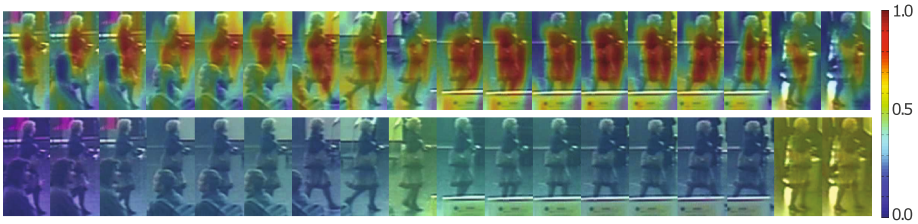

Fig. 1. The color represents the quality score, and warmer is higher quality score. The high quality score means this part should contribute more information for final video representation. The first row illustrates the spatial quality scores generated by our proposed method. The second row illustrates frame-level scores of QAN [7]. This figure shows that our proposed method can fully exploit spatial information, even the QAN determines a frame should have a low score. Best view in color. (Color figure online)

To overcome these weaknesses, Y. Liu et al. [7] proposed Quality Aware Network(QAN) to estimate the frames' quality. It generates a score for each frame and uses weighted sum of scores and corresponding feature embeddings to aggregate frame-level features. However, their method assumes all pixels in a feature map have the same score. This assumption losses the spatial differences of frames, as Fig. 1 shows to us.

In this paper, we mainly focus on aggregating features in spatial across frames. We consider the impact of different pixels in feature maps to improve person re-id performance. To achieve our goal, we proposed a network, named Spatial Quality Aware Network (SQAN). The SQAN has two branches and supports end-to-end training. The first branch is to learn a representation in frame-level. The second branch is to learn quality scores for different pixels of a feature map. Then the outputs of two branches will be aggregated to form a compact video-level representation. Note that in the second branch, we adopt an unsupervised like manner to learn scores, which means it does not depend on the human-made score label. What's more, to overcome the overfitting problem, we proposed an effective dropout strategy.

We evaluate our method in two datasets, iLIDS-VID and MARS. Experiments indicate that the proposed method is effective and is competitive with state-of-the-art methods.

In a word, the main contributions of the paper are as follows:

- The major contribution is that we proposed a Spatial Quality Aware Network (SQAN), which fully exploits spatial information of frames in a video. It shows big improvement in person re-id task than former method.
- The minor contribution is the proposed semantic dropout strategy that is used to effectively regularize spatial information.
- Experiments show that our proposed method reaches competitive performance compared with state-of-the-art methods.

2 Related Work

The proposed SQAN mainly builds upon deep learning based person re-identification and dropout strategy. Below, we review the related works in these two aspects.

Deep learning based person re-identification. Along with the rapid development of deep learning, many attempts have been made to apply deep models into person re-id. Wu et al. [8] proposed that hand-crafted histogram feature is complementary to Convolutional Neural Network(CNN) feature. Liu et al. [7] designed a quality generate unit to distribute different weights to frames, then use the weighted sum of them to represent a video. What's more, some methods adopt Recurrent Neural Network(RNN) and its variants to learn video-level feature for video based re-id task. McLaughlin et al. [5] use CNN to extract frame-level features from the frame and optical flow, then RNN is used to aggregate features across frames. Yan et al. [9] use Long-Short Term Memory network [10] to aggregate frame-level features into video-level feature.

To fully exploit frames' information, we proposed spatial quality aware network (SQAN). It can be seen as an extension of QAN proposed by [7]. Our SQAN fully exploits the spatial differences across frames, which is omitted by QAN.

Dropout strategy. Dropout [11] is a widely used method in deep learning to relief overfitting problem, which is mostly severe when training data is not enough. Due to the small scale in most existing person re-id datasets, this method should be useful in person re-id task. The traditional dropout [11] randomly set some values to zero for the given inputs. Geng et al. [12] proposed pairwise-consistent dropout, which is used for dropping the values in same positions among multiple input feature vectors. Tompson et al. [13] proposed a method to regularize for convolution layers, which sets all the values across the randomly selected channels of the feature map into zero.

However, [11,12] don't consider the spatial correlation and semantic structure of feature maps. [13] only consider the spatial correlation in randomly selected channels. Thus, we propose a semantic dropout strategy, which drops values in a feature map and all the values in the same position across channels will be dropped too. See details in Sect. 3.3

3 Proposed Method

3.1 Architecture Overview

Recent work [7] shows great improvement on person re-id by granting a score to each frame of a video. However, it considers every part of a frame owns the same weight. It ignores the useful information in some parts of a frame with a low score. To make the best use of useful information from all frames, we designed a network, named Spatial Quality Aware Network (SQAN). The core part of it is spatial quality generate module. It gives a score for each pixel of a frame's feature map. Note that a pixel in high level representation feature map corresponds to

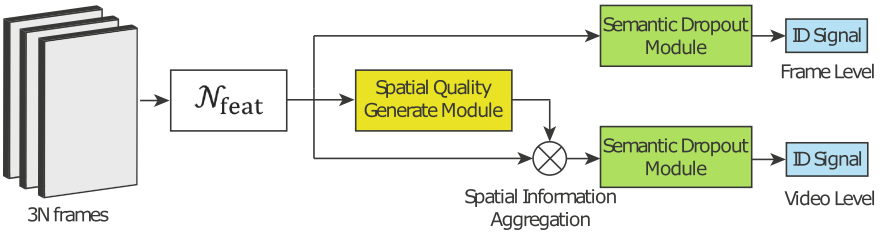

Fig. 2. The proposed Spatial Quality Aware Network (SQAN). The network's input is 3N frames, where N is the sampled frame number of one person's video. There are 2N frames belong to the same identity, while N frames belong to another identity. $\mathcal{N}_{feat}$ is the feature extraction network. The Spatial Quality Generate Module and Semantic Dropout Module are introduced at Sects. 3.2 and 3.3. The final representation is the feature after spatial information aggregation.

a specific part in original frame. And this operation can be seen as an quality evaluation to a specific part of original frame. Then the scores are normalized across feature maps in a video. Finally, these feature maps are aggregated to represent a video. Besides, we design a semantic dropout strategy to overcome overfitting. See Fig. 2 for details.

3.2 Spatial Quality Generate Module

Given the input video V with N frames of a person. Let $I_i (i = 1, \ldots, N)$ to represent its frames. The module's target is to output scores for each pixel of feature maps. A deep neural network $\mathcal{N}_{feat}$ is used to extract frame-level feature. In this paper, we use GoogLeNet [14] as $\mathcal{N}_{feat}$. To encode spatial information, the last 7×7 feature maps is used to learn the score, formulated as $f_{7\times7} = \mathcal{N}_{feat}(I)$. We use f to represent $f_{7\times7}$ below for simplicity. The Spatial Quality Generate Module(SQGM) includes three layers: a $1 \times 1 \times 512$ convolution layer,

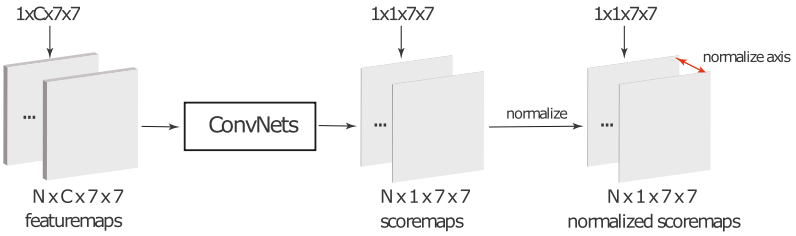

Fig. 3. Spatial Quality Generate Module. The input of this module is the spatial feature maps of a video, which contains N feature maps and C channels for each. Both height and width of feature maps are 7 for example. After passing through a ConvNets, N score maps are produced. Then the score maps are normalized across N feature maps, to be the final output of this module.

a $3\times3\times512$ convolution layer, and a $1\times1\times1$ convolution layer. Each convolution layer is followed by a batch normalization layer. And the activation function is ReLU [15]. See Fig. 3 for details. After passing through the layers, we get N corresponding score maps $S_i(i = 1, \ldots, N)$. Then we normalize the score maps as below:

$$S^{x,y}_{norm_i} = \frac{e^{S^{x,y}_i}}{\sum_j e^{S^{x,y}_j}} \tag{1}$$

Note that $S^{x,y}_{norm_i}$ is the normalized score at position (x, y) of frame i's feature map.

Then we calculate the weighted sum of f and normalized score maps as the final video representation F:

$$F^{x,y,c} = \sum_{i=1}^{N} S^{x,y}_{norm_i} \cdot f^{x,y,c}_i \tag{2}$$

Note that $F^{x,y,c}$ and $f^{x,y,c}_i$ are values at position (x, y) in channel c of final feature map and input feature map separately.

3.3 Semantic Dropout Module

Overfitting is a severe problem in model optimization, especially in small datasets. Dropout is an effective method to relief this problem. The most common dropout strategy drops values randomly and is mostly applied to the feature vector. [13] proposed a convolution dropout strategy and it drops values across some randomly selected channels. However, in SQAN, the corresponding vector at each pixel in f is highly semantic. Our intuition is to make the representation more robust via dropping some pixel-wise vectors and letting the remains can also represent a person. Thus, we propose a dropout strategy, named Semantic Dropout. It drops randomly selected pixel's values in a feature map. And all the values in the same position across all channels will be dropped too. See Fig. 4 for details.

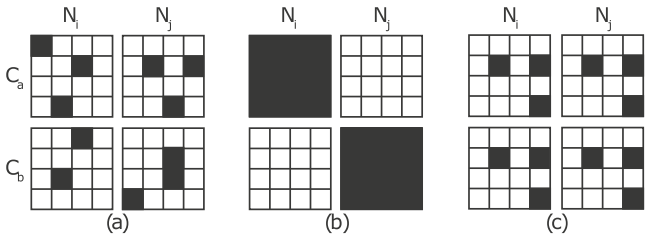

Fig. 4. The difference among three kinds of dropout strategies. (a) is the normal dropout. (b) is spatial dropout [13]. (c) is our proposed semantic dropout. C_a and C_b means two channels. N_i and N_j means two feature maps from different frames of a video.

Note that it is important to drop f and F in the same dropout pattern, i.e. the dropped pixels should be same. Because F is aggregated from f, if they adopt different dropout pattern, the optimized target will be inconsistent.

3.4 Multi-Loss Supervised Training

We hope the aggregated feature can not only classify identities, but the distance between different identities is distant. So we use triplet loss [16] to deal with this issue. The overall loss of SQAN can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{softmax_1} + \mathcal{L}_{softmax_2} + \mathcal{L}_{softmax_3} + \mathcal{L}_{trp} \tag{3}$$

$$\mathcal{L}_{trp} = \sum_{V_a, V_p, V_n} [\|\mathcal{N}(V_a) - \mathcal{N}(V_p)\|_2^2 - \|\mathcal{N}(V_a) - \mathcal{N}(V_n)\|_2^2 + margin]_+ \tag{4}$$

Note that $\mathcal{L}_{softmax_1}$,$\mathcal{L}_{softmax_2}$,$\mathcal{L}_{softmax_3}$ are the original GoogLeNet softmax loss. $\mathcal{L}_{trp}$ represent triplet loss. $\mathcal{N}(V)$ is the representation for video V. V_a and V_p are same identity's video, while V_n is one different identity's video. What's more, the function $[z]_+ = max(z, 0)$ and $margin$ is a hyper-parameter which is set to 1.2 in all experiments.

4 Experiment

4.1 Datasets and Evaluation Protocol

iLIDS-VID. The iLIDS-VID [17] dataset contains 600 videos of 300 randomly sampled people. Each person has one pair of video from two camera views. Each video is comprised of 23 to 192 image frames, with an average length of 73 for each. The challenges of this dataset largely lie in clothing similarities, lighting and viewpoint changes across camera views, complicated background, and occlusions.

The evaluation on iLIDS-VID is the same as previous methods [9]. The dataset is randomly divided into training set and testing set by half, with no overlap between them. During testing, the sequence of the first camera is regarded as the query, while sequences from the second camera as the gallery set. The widely used cumulative matching characteristic (CMC) curve is employed for measuring the performance of methods on this dataset. To ensure statistically reliable evaluation, we repeat the procedure 10 times and use the average performance as the result.

MARS. MARS [6] is a recently released large scale video-based re-id dataset. It contains 1,261 identities and around 20,000 video sequences. The dataset has 1,191,003 images in total from six different cameras and each identity has 13.2 sequences on average. Different from the iLIDS-VID dataset, it has no manually annotated bounding boxes. Each sequence is automatically obtained by pedestrian detector and tracker. Besides, the dataset also contains 3,248 distractor sequences.

For the sake of large scale MARS dataset, the train/test split is fixed with 631 and 630 identities respectively. We use mean average precision score (mAP) and cumulative matching characteristic (CMC) to evaluate methods, which are recommended in [6]. The evaluation mode is video-to-video, single query.

4.2 Implementation Details

Our implementation is based on the open source deep learning framework Caffe [18]. All our experiments were carried on a NVIDIA TITAN X GPU with 12GB of onboard memory. The network is trained with stochastic gradient descent (SGD) end-to-end. The learning rate is set to 1e-3. The total iterations are 15,000 for iLIDS-VID and 250,000 for MARS. The weight decay is set to 0.002. The batch size is fixed to 24, and 8 frames are randomly sampled for anchor, positive and negative classes in triplet loss. As for SDM, the dropout ratio is set to 0.3, which means a pixel vector will be selected to drop in a probability of 30%.

4.3 Ablation Study on iLIDS-VID

Table 1 compares the results of different variants of SQAN. We remark that in this table all results are obtained in the same experiment settings, except (a). So the differences are contributed by the method itself.

Method (a) is the original GoogLeNet with Batch Normalization. It only has image level softmax supervision and uses average pool to aggregate features. What's more, it only has 4,000 iterations for it converge rapidly. It reaches 61.3% CMC1 performance, which is similar to previous works.

Method (b) is QAN proposed by [7]. The "QGM" means the frame-level quality generate module, which distribute a score to each frame. It improves 11% in CMC1 and we think the improvement is from two aspects: one is the video-level supervision and the other is the frame-level quality score.

Method (c) is our proposed method with SQGM. It brings about 14% improvement comparing with QAN. This shows to us that spatial information can not be omitted. Parts in a frame with low score may be important and vice versa.

Method (d) is the final version of SQAN. It introduces SDM based on (c) and further gains 1.3% in CMC1. All these results show the effectiveness of our proposed methods.

4.4 Comparison with State-of-the-art Methods

To further judge the effectiveness, we also compare our methods with other state-of-the-art methods. We evaluate our methods both in iLIDS-VID and MARS.

Table 1. Ablation Study on iLIDS-VID

	(a)	(b)	(c)	(d)
+QGM		✓		
+SQGM			✓	✓
+SDM				✓
CMC1	61.3	68.0	76.4	**77.7**
CMC5	83.3	86.8	93.4	**94.3**
CMC10	88.7	89.8	97.3	**97.4**
CMC20	92.4	97.4	99.1	**99.8**

Table 2 shows the results on iLIDS-VID. SQAN achieves higher CMC than most of the other methods, only a little bit lower than current state-of-the-art method PAM-LOMO+KISSME.

Table 2. Comparison of SQAN and other state-of-the-art methods on iLIDS-VID

	CMC1	CMC5	CMC10	CMC20
SQAN	77.7	94.3	97.4	99.8
QAN	68.0	86.8	89.8	97.4
CNN+RNN [5]	58.0	84.0	91.0	96.0
TDL [19]	56.3	87.6	95.6	98.3
FrameExtraction+CNN [20]	60.2	85.1	-	94.2
PAM-LOMO+KISSME [21]	79.5	95.1	97.6	99.1

Table 3 shows the results on MARS. Comparing with the results on iLIDS-VID, SQAN has a big gap below the state-of-the-art method on MARS. We can compensate it by adding XQDA and re-ranking [22]. The we can get 75.8% CMC1 and 67.4% mAP. This performance is close to the state-of-the-art TriNet [4] and better than those methods with similar additions. But we argue that the intrinsical reason of bad performance is the properties of attention like schema. The attention schema for feature aggregation has the assume that semantic part in each frame should be aligned. For a more realistic and not cropped dataset like MARS, the misalignment problem is more frequent and severe than it on iLIDS-VID. This problem will be left for our future work.

Table 3. Comparison of SQAN and other state-of-the-art methods on MARS.

	CMC1	CMC5	mAP
SQAN	67.5	80.3	41.0
SQAN+XQDA+re-ranking[a]	75.8	85.5	67.4
CNN+XQDA [6]	65.3	82.0	47.6
FrameExtraction+CNN [20]	55.5	70.2	-
CaffeNet+XQDA+re-ranking [22]	67.8	-	58.0
ResNet50+XQDA+re-ranking [22]	73.9	-	68.5
TriNet [4]	79.8	91.4	67.7

[a]We use the released code by [22] for XQDA and re-ranking

5 Conclusion and Future Work

In this paper, we propose a Spatial Quality Aware Network (SQAN) for person re-identification. The proposed method can distribute a quality score to each pixel of a frame's feature map, then the weighted feature maps are aggregated across frames to represent the video of a person. What's more, we also propose a dropout strategy, named semantic dropout, which effectively reduces the impact of overfitting. Experiments show the effectiveness of our method and our method is competitive with state-of-the-art methods in performance.

SQAN is a fine-grained spatial information aggregation model. It may suffer from the severe misalignment problem in a video. Thus, how to integrate alignment method into SQAN will be explored in our future work.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (No. 61472023) and the State Key Laboratory of Software Development Environment (No. SKLSDE-2016ZX-24).

References

1. Chen, D., Yuan, Z., Hua, G., Zheng, N., Wang, J.: Similarity learning on an explicit polynomial kernel feature map for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1565–1573 (2015)
2. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2197–2206 (2015)
3. Su, C., Yang, F., Zhang, S., Tian, Q., Davis, L.S., Gao, W.: Multi-task learning with low rank attribute embedding for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3739–3747 (2015)
4. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
5. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1325–1334 (2016)

6. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: MARS: a video benchmark for large-scale person re-identification. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 868–884. Springer, Cham (2016). doi:10.1007/978-3-319-46466-4_52
7. Liu, Y., Yan, J., Ouyang, W.: Quality aware network for set to set recognition. arXiv preprint arXiv:1704.03373 (2017)
8. Wu, S., Chen, Y.C., Li, X., Wu, A.C., You, J.J., Zheng, W.S.: An enhanced deep feature representation for person re-identification. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–8. IEEE (2016)
9. Yan, Y., Ni, B., Song, Z., Ma, C., Yan, Y., Yang, X.: Person re-identification via recurrent feature aggregation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 701–716. Springer, Cham (2016). doi:10.1007/978-3-319-46466-4_42
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
11. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
12. Geng, M., Wang, Y., Xiang, T., Tian, Y.: Deep transfer learning for person re-identification. arXiv preprint arXiv:1611.05244 (2016)
13. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 648–656 (2015)
14. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
15. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: Aistats, vol. 15, p. 275 (2011)
16. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
17. Wang, T., Gong, S., Zhu, X., Wang, S.: Person re-identification by video ranking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 688–703. Springer, Cham (2014). doi:10.1007/978-3-319-10593-2_45
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093 (2014)
19. You, J., Wu, A., Li, X., Zheng, W.S.: Top-push video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1345–1353 (2016)
20. Zhang, W., Hu, S., Liu, K.: Learning compact appearance representation for video-based person re-identification. arXiv preprint arXiv:1702.06294 (2017)
21. Khan, F.M., Brèmond, F.: Multi-shot person re-identification using part appearance mixture. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 605–614. IEEE (2017)
22. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding. arXiv preprint arXiv:1701.08398 (2017)