

Non-local NetVLAD Encoding for Video Classification

Yongyi Tang[†] Xing Zhang[‡] Jingwen Wang[†] Shaoxiang Chen[‡]
Lin Ma[†] Yu-Gang Jiang[‡]

[†]Tencent AI Lab [‡]Fudan University

{yongyi.tang92,skyezx2018,jaywongjaywong,forwchen}@gmail.com
forest.linma@gmail.com ygj@fudan.edu.cn

Abstract. This paper describes our solution for the 2nd YouTube-8M video understanding challenge organized by Google AI. Unlike the video recognition benchmarks, such as Kinetics and Moments, the YouTube-8M challenge provides pre-extracted visual and audio features instead of raw videos. In this challenge, the submitted model is restricted to 1GB, which encourages participants focus on constructing one powerful single model rather than incorporating of the results from a bunch of models. Our system fuses six different sub-models into one single computational graph, which are categorized into three families. More specifically, the most effective family is the model with non-local operations following the NetVLAD encoding. The other two family models are Soft-BoF and GRU, respectively. In order to further boost single models performance, the model parameters of different checkpoints are averaged. Experimental results demonstrate that our proposed system can effectively perform the video classification task, achieving 0.88763 on the public test set and 0.88704 on the private set in terms of GAP@20, respectively. We finally ranked at the fourth place in the YouTube-8M video understanding challenge.

1 Introduction

Understanding video content is a major challenge for numeric applications including video classification [24], video captioning [28, 30], video localization [6, 11], video attractiveness analysis [8], and so on. Especially with the exponential increment of online videos, video tagging, video retrieval and recommendation is of great demand. Therefore, developing reliable video understanding algorithms and systems has received extensive attentions in the area of computer vision and machine learning.

In order to recognize video content, convolutional neural networks(CNNs) [27, 24, 5, 23] and/or recurrent neural networks based methods [10, 26] have achieved state-of-the-arts results. Those methods [24] take the advantages of deep learning methods on static image content as well as the video motion containing temporal information to perform video analysis. However, prior works only perform on those video benchmarks with limited number of videos for model evaluations

such as UCF-101 [25], HMDB-51 [15], and ActivityNet [4] datasets. Recently, several large-scale video datasets are constructed, including the Kinetics dataset [17] developed by DeepMind and the Moments in Time dataset developed by MIT-IBM [21] with about a million of videos. However, for practical video applications such as YouTube and Netflix, such number of videos is still relatively small and not suitable for large-scale video understanding. Nowadays, Google AI releases a large-scale video dataset named YouTube-8M [2], which contains about 8 million YouTube videos with multiple class tags.

For the 1st Youtube-8M video understanding challenge, several techniques including context gating [20], multi-stage training [29], temporal modeling [19], and feature aggregation [7] have been proposed for video classification. However, the excellent performances of prior works mainly attribute to ensemble the results from a bunch of models, which is not practical in real-world applications due to the heavy computational expense. Therefore, the 2nd YouTube-8M video understanding challenge focus on learning video representation under budget constraints. More specifically, the model size of submission is restricted to 1GB, which encourages the participants to explore compact video understanding models based on the pre-extracted visual and audio features.

In this report, we propose a compact system that meets the requirements and achieves superior results in the challenge. We summarize the contributions as follows. First, we stack the non-local block with the NetVLAD to improve the video feature encoding. Experimental results demonstrate that the proposed non-local NetVLAD pooling method outperforms the vanilla NetVLAD pooling. Second, several techniques are employed for building the large-scale video classification system with limited number of parameters including weight averaging strategy of different checkpoints, model ensemble, and compact encoding of floating point number. Lastly, we show that the selected single models are complementary to each other which makes the whole system achieves a competitive result on the 2nd YouTube-8M video understanding challenge, ranked at the forth position.

2 Approach

The framework of our proposed system is shown in Fig. 1. In this work, we use three different families of video descriptor pooling methods for the video classification task, specifically the non-local NetVLAD, Soft-Bag-of-Feature (Soft-BoF), and GRU. In section 2.1, we introduce the details of the proposed NetVLAD incorporated with the non-local block with its variants introduced in section 2.2. The other two family models, namely the Soft-BoF and GRU, are introduced in section 2.3 and 2.4, respectively. The model ensemble is described in section 2.5.

2.1 Non-local NetVLAD

Vector of Locally Aggregated Descriptors (VLAD). VLAD [14] is a popular descriptor pooling method for instance level retrieval [14] and image

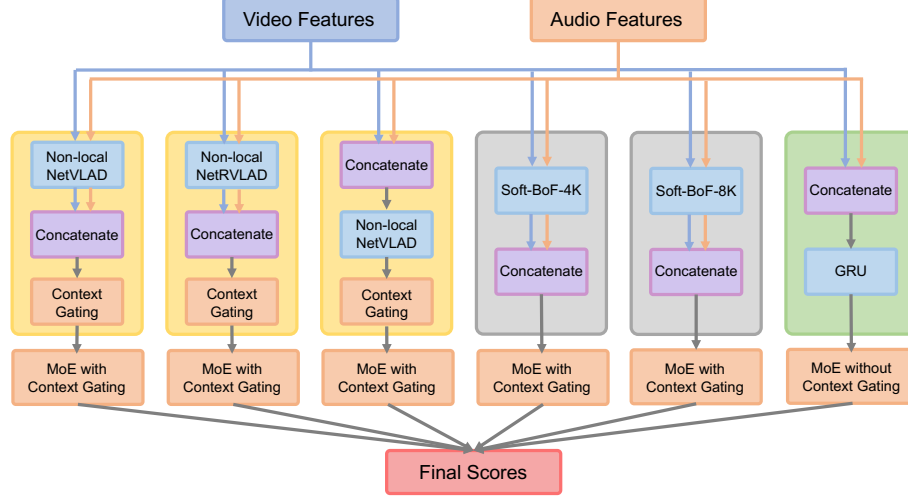


Fig. 1. The framework of our proposed system for video classification.

classification [12], as it captures the statistic information about the local descriptors aggregated over the image. Specifically, the VLAD summarizes the residuals of descriptors and its corresponding cluster center. Formally, given N D -dimensional descriptors $\{\mathbf{x}_i\}$ as input, and K cluster centers $\{\mathbf{c}_k\}$ as VLAD parameters, the pooling output of VLAD is $K \times D$ -dimensional representation V . Writing V as a $K \times D$ matrix, the (j, k) element of V can be computed as follows:

$$V(j, k) = \sum_{i=1}^N a_k(\mathbf{x}_i)(x_i(j) - c_k(j)), \quad (1)$$

where the $a_k(\mathbf{x}_i)$ indicates the hard assignment of the descriptor \mathbf{x}_i to k -th visual word \mathbf{c}_k . Thus, each column of matrix V records the sum of residuals of the descriptors. Intra-normalization and inter-normalization are performed after VLAD pooling.

NetVLAD Descriptor. However, the VLAD algorithm involves a hard cluster assignment that is non-differentiable. Thus the vanilla VLAD encoding is not appropriate for deep neural network that requires computing gradients for back-propagation. To address this problem, Arandjelovic *et al.* proposed the NetVLAD [3] with soft assignment $\bar{a}_k(\mathbf{x}_i)$ of descriptors \mathbf{x}_i to multiple cluster centers \mathbf{c}_k , *i.e.*,

$$\bar{a}_k(\mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i + b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T \mathbf{x}_i + b_{k'}}}, \quad (2)$$

where $\{\mathbf{w}_k\}$, $\{b\}$ and $\{\mathbf{c}_k\}$ are the learnable parameters of the NetVLAD descriptor.

Non-local NetVLAD Descriptor. As described above, the VLAD descriptor uses cluster centers \mathbf{c}_k to represent features, while NetVLAD further uses soft-assignment to construct the local feature descriptors. To enrich the information of NetVLAD descriptors, we model the relations between different local cluster centers. We employ the non-local block proposed by Wang *et al.* [31], which has already demonstrated the relation modeling ability in action recognition task. Here, we empirically adopt the embedded Gaussian function to compute the non-local relations:

$$f(\mathbf{v}_i, \mathbf{v}_j) = e^{\theta(\mathbf{v}_i)^T \phi(\mathbf{v}_j)}. \quad (3)$$

Specifically, given the NetVLAD descriptor \mathbf{v}_k corresponding to cluster centers \mathbf{c}_k , the non-local NetVLAD descriptor $\hat{\mathbf{v}}_k$ of cluster k is formulated as:

$$\hat{\mathbf{v}}_i = \mathbf{W}\mathbf{y}_i + \mathbf{v}_i, \quad (4)$$

where $\mathbf{y}_i = \frac{1}{Z(\mathbf{v})} \sum_{\forall j} f(\mathbf{v}_i, \mathbf{v}_j)g(\mathbf{v}_j)$. For implementation, the non-local NetVLAD is formulated as:

$$\mathbf{y} = \text{softmax}(\mathbf{v}^T \mathbf{W}_\theta^T \mathbf{W}_\phi \mathbf{v})g(\mathbf{v}), \quad (5)$$

where $g(\mathbf{v})$ is a linear transformation.

2.2 Non-local NetVLAD Model and its Variants

Note that in our system, we use three variant non-local NetVLAD methods, which are demonstrated to be complementary with each other.

Late-fused Non-local NetVLAD (LFNL-NetVLAD). The first model is the late-fused non-local NetVLAD (LFNL-NetVLAD). The pre-extracted visual feature and audio feature are encode independently by the non-local NetVLAD pooling method. Afterwards, these two non-local NetVLAD features, encoding visual and audio modalities, are concatenated into a vector, which is followed by the context gating module.

Please note that context gating is introduced by Miech *et al.* [20], which transforms the input feature into a new representation and captures feature dependencies and prior structure of output space. Context gating is defined as:

$$\mathbf{z} = \text{sigmoid}(\mathbf{W}\mathbf{y}) \odot \mathbf{y}, \quad (6)$$

where \odot indicates elements-wise multiplication.

As shown in Fig.1, the mixture of experts (MoE) model [16] equipped with video level context gating is used for the multi-label video classification.

Late-fused Non-local NetRVLAD (LFNL-NetRVLAD). In addition, the NetRVLAD that drops the computation of cluster centers is proposed in [20], which can be considered as self-attended local feature representation. Formally, the NetRVLAD can be defined as:

$$V'(j, k) = \sum_{i=1}^N \bar{a}_k(\mathbf{x}_i) x_i(j), \quad (7)$$

where the soft assignment $\{\bar{a}_k\}$ are computed by Eq. 2. Similarly, the video and audio features pass through non-local NetRVLAD pooling and perform concatenation, followed by one context gating module and the MoE equipped with video level context gating.

Early-fused Non-local NetVLAD (EFNL-NetVLAD). Early fusion that concatenates the video and audio feature before non-local NetVLAD pooling is used to build another model. The early-fused feature lies in different feature space resulting in different expressive ability compared with the late-fused representation. The frame level context gating and video level MoE with context gating are also used in this model.

2.3 Soft-Bag-of-Feature Pooling

For bag-of-feature encoding, we utilize soft-assignment of descriptors to feature clusters [22] to obtain the distinguishable representation. Also, we perform late fusion of Soft-BoF with 4K and 8K clusters, which are named as Soft-BoF-4K and Soft-BoF-8K, respectively. Those outputs only followed by the video level MoE with context gating.

2.4 Gated Recurrent Unit

Recurrent neural networks, especially the Gated Recurrent Unit (GRU) [9], have been investigated for video understanding [10, 20, 19, 7]. We stacked two layers of GRU of 1024 hidden neurons for each layer. The experimental results demonstrate that the GRU model is complementary with the non-local NetVLAD and Soft-BoF families resulting a significant improvement after model ensemble.

2.5 Model Ensemble

Model ensemble is a common way for boosting final results in different challenges [20, 7, 29, 19, 32]. The superior improvement may attribute to the various feature expressions of different models. Thus, model ensemble helps to finalize a robust result and relief over-fitting. We perform model ensemble based on the six different models as mentioned. Experimental results along with implementation details will be introduced in the following.

Table 1. Single model performances on our split validation set.

Model	LFNL-NetVLAD	EFNL-NetVLAD	LFNL-NetRVLAD
GAP@20	0.8703	0.8674	0.8687
Model size	593M	427M	478M
Model	Soft-BoF-4K	Soft-BoF-8K	GRU
GAP@20	0.8525	0.8512	0.8568
Model size	109M	143M	243M

Table 2. Single averaged model performances on our split validation set.

Averaged Model	LFNL-NetVLAD	LFNL-NetRVLAD	EFNL-NetVLAD
GAP@20	0.8716	0.8704	0.8704
Averaged Model	Soft-BoF-4K	Soft-BoF-8K	GRU
GAP@20	0.8574	0.8563	0.8612

3 Experiments

3.1 YouTube-8M Dataset

The YouTube-8M dataset [2] adopted in the 2nd YouTube-8M Video Understanding Challenge is the 2018 version with higher-quality, more topical annotations, and a cleaner annotation vocabulary. It contains about 6.1 million videos, 3862 class labels and 3 labels per video on average. Because of the large scale of the dataset, the video information is provided as pre-extracted visual and audio features at 1 FPS.

3.2 Implementation Details

The provided dataset is divided into training, validation and test subsets with around 70%, 12% and 18% of videos. But in our work, we keep around 100K videos for validation, and the remaining videos of training and validation subset are used for training due to the observation of improvement. We found that the performance on our validation set was 0.02-0.03 lower than the test set on the public leader board. We report the Global Average Precision (GAP) metric at top 20 with our split validation subset and the public test set shown on the leader board.

For most of the models, we empirically used 1024 hidden states except for the GRU model which adopted 1200 hidden states. We trained every single model independently with our training split on Tensorflow [1]. The Adam optimizer[18] with 0.0002 as the initial learning rate was employed throughout our experiments. Training procedures converged around 300k. After finishing the training procedure, we built a large computational graph of model ensemble, and the parameters within this graph were imported from the independent models. The

averaged score of each sub-model was the final score of our system. Further fine-tuning for the system may improve the final score. In the submission, we simply used model-wise averaging due to the lack of time.

3.3 Single Model Evaluation

In this section, we evaluate the six single models used in our system as shown in Table 1. For the LFNL-NetVLAD model, we deployed 64 clusters with 8 MoE in video level model achieving 0.8702 GAP@20 in our validation set, while the vanilla NetVLAD achieves 0.8698 under the same settings. Also, 64 clusters were adopted in the EFNL-NetVLAD and the LFNL-NetRVLAD since we found this setting keeps the balance between model size and performance. And the MoE of these two models were 2 and 4, respectively. The model size of non-local NetVLAD models are around 500M, which takes a large portion of the parameters in our system.

We also adopted the GRU model with model size as 243M and two smaller Soft-BoF models with 4K and 8K clusters, respectively, since we found that those models are complementary to the non-local NetVLAD models. The MoE of these three models were set to 2.

In order to further boost the single model performance, we employed linear model averaging that utilizing the average of multiple checkpoint to improve single model performance inspired by Stochastic Weight Averaging method [13]. The final GAP@20 of each model is shown in Table 2, which shows that linear model averaging can significantly improve single model performance especially for the GRU and Soft-BoF models with over 0.005 improvements.

3.4 Tricks for Compact Model Ensemble

Recall that the challenge requires less than 1GB model for final submission. We thus adopted several techniques for improving model abilities under the limited parameters including using 'bfloat16' format of parameters and repeatedly random sampling.

At first, we trained the network with float32 in Tensorflow [1], which means that it takes 4 bytes for every parameter. To make our model meet the model size requirement, we used a tensorflow-specific format, 'bfloat16', in the ensemble stage, which is different from IEEE's float16 format. The bfloat16 is a compact 16-bit encoding of floating point number with 8 bits for exponent and 7 bits for mantissa. We found that using 'bfloat16' format can accelerate the process without significant performance decrease, with its benefits on halving the model size which makes ensembling multiple models become possible. As results, we performed ensemble with the models mentioned in Table 2 into one computational graph as our final model as shown in Table 3.

Further, since feature sub-sampling were used in our sub-models for better generalization, we performed multiple running with different feature sub-sampling in the same system to produce the final classification result. By averaging the 10 times repeated results, the final performance gained about 0.0005

Table 3. Ensemble model performances on our split validation set. M1-M6 denote LFNL-NetVLAD, LFNL-NetRVLAD, EFNL-NetVLAD, Soft-BoF-4k, Soft-BoF-8k and GRU, respectively.

Ensemble Model	Validation GAP@20	Public-Test GAP@20
M1 & M4	0.8752	-
M1 & M2	0.8778	-
M1 & M6	0.8782	0.8790
M1 & M4 & M6	0.8800	-
M1 & M2 & M4 & M6	0.8820	-
M1 & M2 & M3 & M4 & M6	0.8839	0.88678
M1 & M2 & M3 & M4 & M5 & M6	0.8842	-

Table 4. Performances of our model with different times of random averaging.

Ensemble Model	Validation GAP@20	Public-Test GAP@20
Our model run once	0.8842	-
Our model run 5 times	0.8846	0.88756
Our model run 10 times (final submission)	0.8847	0.88763

improvement on our validation set as shown in Table 4. In practice, we repeated the input feature several times, and averaged the results for each video. The final model size of our submission is 995M.

4 Conclusions

In this report, we proposed a compact large-scale video understanding system that effectively performs multi-label classification on the YouTube-8M video dataset with limited model size under 1GB. A non-local NetVLAD pooling method is proposed for constructing more representative video descriptors. Several models including LFNL-NetVLAD, LFNL-NetRVLAD, EFNL-NetVLAD, GRU, Soft-BoF-4K, and Soft-BoF-8K are incorporated in our system for model ensemble. To halve model size, bfloat16 format is adopted in our final system. Averaging multiple outputs after random sampling is also used in our system for further boosting the performance. Experimental results on the 2nd YouTube-8M video understanding challenge show that the proposed system outperforms most of the competitors, ranking the fourth place in the final result.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning.
2. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675 (2016)
3. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5297–5307 (2016)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 4724–4733. IEEE (2017)
6. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video. In: EMNLP (2018)
7. Chen, S., Wang, X., Tang, Y., Chen, X., Wu, Z., Jiang, Y.G.: Aggregating frame-level features for large-scale video classification. arXiv preprint arXiv:1707.00803 (2017)
8. Chen, X., Chen, J., Ma, L., Yao, J., Liu, W., Luo, J., Zhang, T.: Fine-grained video attractiveness prediction using multimodal deep learning on a large real-world dataset. In: WWW (2018)
9. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
10. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
11. Feng, Y., Ma, L., Liu, W., Zhang, T., Luo, J.: Video re-localization. In: ECCV (2018)
12. Gong, Y., Wang, L., Guo, R., Lazebnik, S.: Multi-scale orderless pooling of deep convolutional activation features. In: European conference on computer vision. pp. 392–407. Springer (2014)
13. Izmailov, P., Podoprikin, D., Garipov, T., Vetrov, D., Wilson, A.G.: Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407 (2018)
14. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. pp. 3304–3311. IEEE (2010)
15. Jhuang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T.: A large video database for human motion recognition. In: Proc. of IEEE International Conference on Computer Vision (2011)
16. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the em algorithm. Neural computation 6(2), 181–214 (1994)

17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Li, F., Gan, C., Liu, X., Bian, Y., Long, X., Li, Y., Li, Z., Zhou, J., Wen, S.: Temporal modeling approaches for large-scale youtube-8m video understanding. arXiv preprint arXiv:1707.04555 (2017)
20. Miech, A., Laptev, I., Sivic, J.: Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905 (2017)
21. Monfort, M., Zhou, B., Bargal, S.A., Andonian, A., Yan, T., Ramakrishnan, K., Brown, L., Fan, Q., Gutfrund, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. arXiv preprint arXiv:1801.03150 (2018)
22. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: Improving particular object retrieval in large scale image databases (2008)
23. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 5534–5542. IEEE (2017)
24. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NIPS (2014)
25. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
26. Tang, Y., Zhang, P., Hu, J.F., Zheng, W.S.: Latent embeddings for collective activity recognition. In: Advanced Video and Signal Based Surveillance (AVSS), 2017 14th IEEE International Conference on. pp. 1–6. IEEE (2017)
27. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
28. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning. In: CVPR (2018)
29. Wang, H.D., Zhang, T., Wu, J.: The monkeytyping solution to the youtube-8m video understanding challenge. arXiv preprint arXiv:1706.05150 (2017)
30. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning. In: CVPR (2018)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018)
32. Xiaoteng, Z., Yixin, B., Feiyun, Z., Kai, H., Yicheng, W., Liang, Z., Qinzhu, H., Yining, L., Jie, S., Yao, P.: Qiniu submission to activitynet challenge 2018. arXiv preprint arXiv:1806.04391 (2018)