Semi-Siamese Network for Content-based Video Relevance Prediction

Zongxian Li[1,2*], Sheng Li[1*], Lantian Xue[1,2], Yonghong Tian[1,2†]
[1] National Engineering Laboratory for Video Technology, School of EE&CS, Peking University, Beijing, China
[2] Pengcheng Laboratory, Shenzhen, China

***Abstract*—The intractable "cold-start" problem often encountered in existed video recommendation systems when a new video is coming with minimal users' feedback since most existing recommendation algorithms heavily rely on the users' implicit feedbacks. This paper presents a novel idea for solving the "cold start" problem by analyzing the video content itself. The Harmonic Sampling is designed in our work for utilizing the rank information automatically during the sampling procedure , and a Semi-Siamese network is proposed for overcoming the asymmetric training samples. The proposed method demonstrated its effectiveness in dealing with "cold-start" problem, achieving superior performance over the Content-based Video Relevance Prediction Dataset.**

I. Introduction

The video streaming has become one prevalent Internet services with the development of network techniques, which acted as the most vital role of Internet multimedia services. For better helping users to discover the video they would enjoy, the video recommendations represented by video relevance prediction has attracted increasing interest in recent years.

Current video recommendation algorithms can be roughly divided into two major aspects, collaborative-based filtering [1], [2], [3], [4] and content-based recommendation [5], [6], [7], [8]. Typical collaborative-based filtering methods are mainly implemented by using the users' implicit feedbacks, considering the similarities from users' reviews of videos or some inherent attributes from the meta-information(actors, directors *et al.*). On the other hand, content-based methods mainly focus on the spatial or the temporal feature not only for images but also captions or the voice. The collaborative-based filtering methods are more welcomed by video streaming providers before the rise of deep learning techniques. However, when a new video is coming with minimal users' feedbacks, the "cold-start" problem is often inevitable in the collaborative-based filtering method, and it is hard to analyze the similarity among videos without enough information.

With the dramatic feature extraction abilities brought from the deep neural networks [9], the content-based recommendation algorithms are increasingly recognized for solving the "cold-start" problem. As the core part of the video recommendation system, the video relevance prediction aims to give the recommended video lists when a new video is coming. As shown in Fig. 1, the Video Recommendation System is designed to recommend videos which are relevant to the input video A without any users' feedbacks.

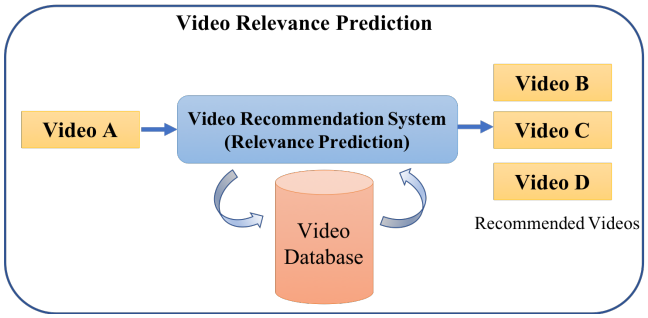

Fig. 1. The overview of the Video Relevance Prediction. When a new video is coming with minimal users' feedback, the Video Recommendation System is used for analyzing the input video and output videos which are relevant to the input video.

In this paper, a Semi-Siamese network is presented for solving the "cold-start" problem in a content-based way. We designed a new Harmonic Sampling strategy which is able to utilize the rank information from the training set automatically during the sampling procedure. For better forming the training pairs, the correlations between positive and negative samples are entirely mined by applying our Harmonic Sampling strategy. For each sample video, the number of unrelated videos is much larger than the number of related videos, resulting in a serious unbalance issue among the training set. Moreover, the correlations between video relevance pairs are not always bidirectional, which means that the video A is relevant to B while the B is irrelevant to A in many cases. The Semi-Siamese network is proposed which trains the Siamese network from two different stages, for better fitting the asymmetric training samples and gaining a significant improvement for video relevance prediction.

The major contributions of this work can be summarized from three aspects as listed:

- The Harmonic Sampling is designed for utilizing the rank information automatically during the sampling procedure.
- The Semi-Siamese network and its training method are proposed, handling the asymmetric training samples which occupied nearly 40% in the training set.
- For obtaining a stable convergence during the training procedure, a new feature penalty is designed and added into the origin contrastive loss function.

∗. Zongxian Li and Sheng Li contribute equally to this work.
† Corresponding author: Yonghong Tian (email: yhtian@pku.edu.cn).

978-1-7281-0397-6/19/$31.00 ©2019 IEEE

II. Related work

Collaborative filtering-based recommendation

The collaborative filtering-based recommendation is first proposed by Tapestry *et al.* [10], which is implemented by using the explicit opinion of people from large communities. After that, the GroipLens Research System is designed [11], providing a pseudonymous collaborative filtering solution. On the other hand, the email-based system Ringo [5] is proposed to generate the recommendation music and movies. Recently, a framework of tightly coupled collaborative filtering and deep neural network [12] is proposed by Wei *et al.* for solving the "cold-start" problem in video relevance prediction.

Content-based recommendation

Traditional content-based algorithms still heavily rely on the users' preferences and feedbacks, by explicitly asking them to make a judgment of the item [14] or analyzing the user activities [15]. Moreover, the meta-data such as actors, directors *et al.* are also considered [16]. However, when a new video is coming without any user information or the meta-data, the current content-based method will usually fall into the "cold-start" problem. For overcoming the "cold-start" issue in a content-based way, HULU proposed to solve the problem by using the triplet network [17], [18].

The proposed Semi-Siamese network is designed specifically for overcoming the "cold-start" problem, analyzing the new video itself(visual feature) and recommending the video lists which are relevant to the input video.

III. Method and Formulation

A. Problem definition

The main purpose of the video relevance prediction is to train a model which is able to learn the relationship between videos and recommend a series of videos which are relevant to the given video. Specifically, when visual feature X_i(extracted from its visual content) of a video V_i is fed into the model, a relevance list of V_i consisting of M_i well-ordered candidates defined as r_i = [r_{i1},r_{i2},...,r_{ij},...,r_{iM_i}] will be returned. j indicates the importance of the candidates(rank information). We defined that Video B is relevant to Video A if B is in the relevance list of A.

B. Proposed method

1) Sampling method: The training data fed into the siamese network are usually desired to be formed as (X_a, X_b, $label$). X_a and X_b refers to visual features of two videos V_a and V_b respectively, and it will be labeled as 1 if the latter video V_b is in V_a's relevance list and 0 otherwise. Hard Negative sampling, the widely-used data sampling strategy, is to select all positive samples directly and then select the same amount of hard negative samples. However, as stated in Sec. I, the serious unbalance issue exists in the training set.

A straightforward idea for overcoming the unbalance issue is to increase the coverage of the negative samples directly and select the positive samples repeatedly until the amount is balanced, which is indeed effective while the useful rank information about the relevance list is totally ignored. For better sampling the training data and utilizing the rank information in the given relevance list of video V_i, the Harmonic Sampling strategy is designed and can be formed as follows:

$$\lambda_{ij} = \frac{\frac{1}{j^p}}{\sum_{n=1}^{M_i} \frac{1}{n^p}} \tag{1}$$

M_i refers to the number of videos which are relevant to the video V_i, where $0 < j \leq M_i$, λ_{ij} indicates the ratio of the positive sample (X_i, $X_{r_{ij}}$, 1) among all samples sampled from the relevance list r_i and it can be regarded as the frequency which this positive sample need to be re-sampled. The sampling process can be completed in this way no matter how many positive samples for each video are needed for balance. Moreover, the sampling intensity can be controlled by adjusting the hyperparameter p.

2) Semi-Siamese network: The Siamese network trained with "pairwise" manners are widely-used for solving the retrieval problem, which is first proposed in [19] and applied in face authentication tasks. The training process of Siamese Network can essentially be regarded as mapping features from the original space to a new space, which makes the highly correlated videos closer to each other and pushes away the unrelvant pairs. However, this similarity metric is generally performed bidirectionally and symmetrically, which is different from the data distribution of video relevance datasets. For example, it is very common that video A is relevant to the video B while B is not relevant to A. A carefully statistic analysis will be conducted in Sec IV-B.

The Semi-Siamese network is proposed for better fitting both symmetrical and asymmetrical data. The training procedure of the proposed method can be divided into two stages. In the first stage, it is exactly the same as the general Siamese Network training, and the symmetrical related samples are selected to train the model. And then, the pre-trained Siamese network is copied directly into two independent models, and it will be fine-tuned by using the asymmetrically related samples.

The entire training procedure is illustrated in Fig.2, where Fig.2-(a)(b), Fig.2-(c)(d) represent the first and second stages, respectively. Fig2-(a)(c) are the network architecture, and Fig2-(b)(d) are abridged general views(only for explaining) of the proposed model. In stage-1, the input sample $(X_a, X_b, label)$ is symmetrical, which means video V_a and V_b are both relevant to each other and the distance defined as $d_{(a,b)}$ is considered equals to $d_{(b,a)}$. As shown in (b), the shared model maps the two different video features X_a and X_b into two new features F_a and F_b. At that time, $d_{(a,b)}$ equals to $d_{(b,a)}$ since they pass through a same model, the original model model-O. In stage-2, as samples are asymmetrical, $(X_p, X_q, label)$ and $(X_q, X_p, label)$ cannot be treated in the same way. We would pass the feature of the formal video in each sample through model-F, and the latter one through model-L. After fine-tuning, both the model-F(denoted as orange line) and model-L(denoted as green line) are different from the model-O(denoted as gray dotted line) as shown in

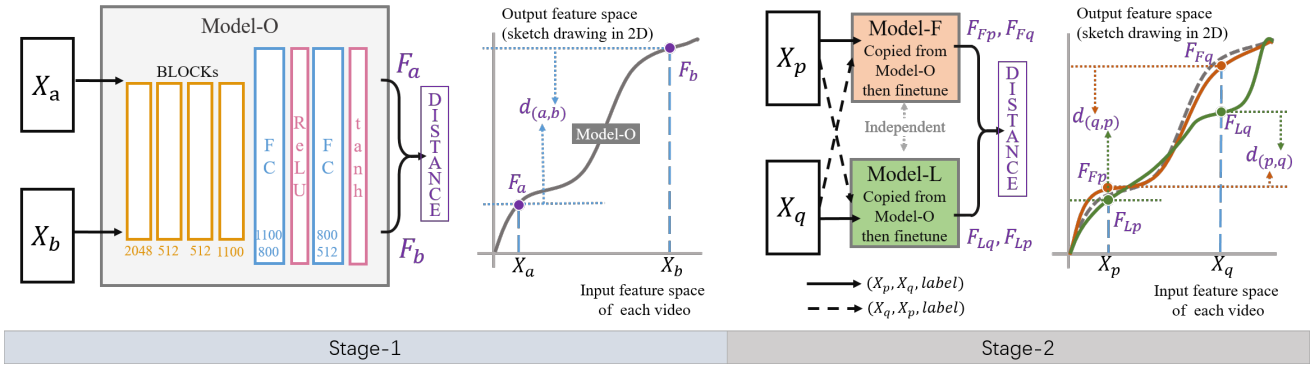

Fig. 2. The training procedure of the proposed Semi-Siamese network. The parameter of the model-F and model-L are completely independent in the second stage. Each block in (a) is consist of a fully-connected layer, a ReLu activation and a dropout. (b) and (d) are the visualization of the training procedure in 2 stage, respectively.

Fig2-(d), hence a same video would get different features when passing through different models, resulting in the difference of $d_{(p,q)}$ and $d_{(q,p)}$. From the comparison of Fig.2-(b) Fig.2-(d), it can be seen that the model is able to predict asymmetric relationships based on the symmetric relationship learned in the first stage.

3) Loss function: The basic contrastive loss[20] is adopted in our work:

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)\,max\,(m-d,0)^2 \qquad (2)$$

Where $d = ||F_1 - F_2||_2$, m refers to the pre-defined margin and y is the label. During the training procedure, the negative samples will be gradually pushed away from each other in the trained feature space. As the amount of negative samples is much larger than positive samples, the essential characteristics of the data still exist even though the imbalance issue is alleviated to some extent by using the Harmonic Sampling strategy. The distance between 2 negative samples will keep increasing even though it has already larger than the pre-defined margin value, which will result in a decrease of the loss value without the performance improvement even drop.

For obtaining a stable convergence during the training procedure, the distance between positive and negative samples should be controlled within a reasonable range. A novel feature penalty is designed and added into the original contrastive loss function and can be defined as follows:

$$L = \frac{1}{2N}\sum_{n=1}^{N} yd^2 + (1-y)\,max\,(m-d,0)^2 + \varphi\,(|F_1|_2 + \ |F_2|_2) \qquad (3)$$

Where φ is a hyperparameter used to adjust the degree of the constraint. F_1 and F_2 refers to the output features by forwarding the proposed Semi-Siamese network. The designed feature penalty works as a regularization, which leads a constraint of the feature distance between fed training pairs.

IV. Experiments

In this section, We first introduced the Content-Based Video Relevance Prediction Dataset(CBVRP), which is used as a benchmark [17] in our work for evaluating our method. And then, we experimentally validated the performance of our Semi-Siamese network and a series of ablation experiments are conducted for analyzing each of its components for video relevance prediction. Since there is hardly previous experimental results on this new dataset, we only made a fair comparison with the baseline method which is provided by HULU, the maker of this challenge dataset. All evaluation experiments are developed using the widely-used $Pytorch$ deep learning framework [21] and run on the NVIDIA Titan Xp GPUs.

A. Dataset and evaluation metrics

1) Content-Based Video Relevance Prediction(CBVRP): For driving the study on the video relevance recommendation and explore effective solutions for overcoming the "cold-start" problem, the large Content-Based Video Relevance Prediction Datasets is released by the HULU, a world famous video streaming provider. Nearly 18,000 video trailers are released in the forms of the pre-extracted visual feature, including the C3D feature at the video level and the Inception feature at the frame level. All video trailers are divided into 2 separated tracks, TV-shows and Movies. Specially, C3D features with 512 dimension are extracted by using the state-of-the-art C3D architecture [22], and the Inception features with 2048 dimension are obtained from the Inception V3 networks [23], which are pre-trained on the ILSVRC2012 Dataset [24].

2) Evaluation metrics: A relevance list with N sample videos are provided as ground truth lists in the training set in CBVRP. Specially, for each sample video V_i, there are M_i ground truth relevance videos. The relevance list is defined as r_i = [r_{i1},r_{i2},...,r_{ij},...,r_{iM_i}], where j indicates the importance of the candidates. The recall rate recall@Top K are computed as an performance measurement.

$$recall@k = \frac{|r_i \cap \hat{r_i}|}{|r_i|} \qquad (4)$$

Where $\hat{r_i}$ refers to the predicted results.

Methods	Recall@k			
	k=50	k=100	k=200	k=300
Baseline-C3D	0.111	0.175	0.264	0.329
Baseline-Inception	0.111	0.172	0.262	0.331
Baseline-fusion	0.123	0.189	0.281	0.352
Semi-Siamese-C3D	0.148	0.225	0.315	0.380
Semi-Siamese-Inception	0.119	0.186	0.281	0.342
Semi-Siamese-fusion	0.161	0.237	0.353	0.402

TABLE I
TRACK 1 - TV-SHOWS: PERFORMANCE COMPARISON WITH THE BASELINE METHOD.

Methods	Recall@k			
	k=50	k=100	k=200	k=300
Baseline-C3D	0.101	0.143	0.206	0.257
Baseline-Inception	0.086	0.125	0.185	0.229
Baseline-fusion	0.107	0.154	0.219	0.269
Semi-Siamese-C3D	0.108	0.164	0.226	0.283
Semi-Siamese-Inception	0.089	0.133	0.198	0.240
Semi-Siamese-fusion	0.110	0.171	0.247	0.308

TABLE II
TRACK 2 - MOVIES: PERFORMANCE COMPARISON WITH THE BASELINE METHOD.

B. Data statistic analysis

A careful data statistic analysis of the CBVRP dataset is conducted. A new measurement "Betrayal Rate", denoted as γ is introduced for evaluating the "unidirectional" among the training set. Specially, considering the video V_i in the training set, γ_i can be defined as follows:

$$\gamma_i = \frac{\sum_{j=1}^{M_i} |C(ID(i,j),i) - 1|}{M_i} \tag{5}$$

The $ID(i,j)$ returns the id information of the j_{th} video in video V_i's relevance list. The $C(ID(i,j),i)$ will return 1 when V_i is in the relevance list of $V_{ID(i,j)}$. γ for all videos can be computed as follows:

$$\gamma_{all} = \frac{\sum_{i=0}^{N}\sum_{j=1}^{M_i} |C(ID(i,j),i) - 1|}{\sum_{i=0}^{N} Mi} \tag{6}$$

N is the total number of the videos in training set.

The betrayal rate γ_{all} for track TV-shows is 38.95% and is 42.69% for track Movies. The distribution of the Betrayal Rate for each track is showed in Fig 3.

C. Performance Evaluation

1) Baseline method: The baseline method is provided by the HULU, the maker of the CBVRP dataset. The basic triplet loss is utilized for relevance learning. For triplet construction, the video which is relevant to the sample is regarded as a positive and will be considered as a negative otherwise.

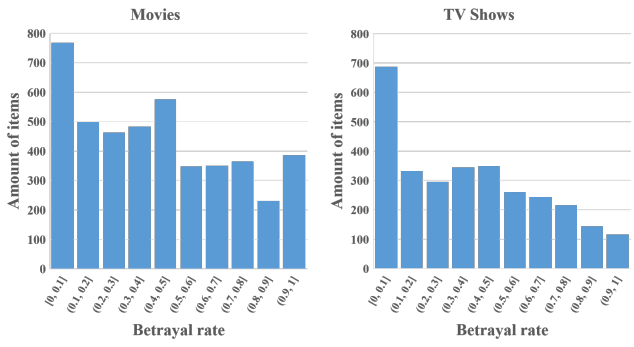

Fig. 3. The distribution of Betrayal Rate γ on TV shows and Movies tracks.

Models	Recall@100
Semi-Siamese+HN	0.168
Semi-Siamese+HS	0.199
Semi-Siamese+HS+Constraint	0.237

TABLE III
ABLATION STUDY ON TV-SHOWS

2) Results and analysis: Tab. I and II show the Recall rate of the proposed method and the baseline method, respectively. The detailed architecture of the model is illustrated in Fig 2-(a). All models in Tab. I and II are trained with the proposed Semi-Siamese network and sampled by applying the Harmonic Sampling strategy. As shown in Tab. I and II that the proposed method is clearly outperforming by a large margin on the Recall rate when compared with the baseline method whether using the C3D feature and the Inception feature separately or in combination. The clear Recall rate improvements demonstrate the effectiveness of the proposed method.

3) Ablation analysis: We conduct different experiments on the Track-show training set to study how each component help to improve the performance and the results are reported in Tab. III. The HN refers to the general Hard Negative sampling and the HS refers to the proposed Harmonic Sampling. The constrain means the added feature penalty. When compared with the general Hard Negative sampling strategy, the proposed Harmonic Sampling strategy gains 3.1% improvement.

V. CONCLUSION

This paper addressed the "cold-start" problem and make a more precise prediction for the video relevance when a new video is coming. The Semi-Siamese network is proposed to train the Siamese network from two independent stage, fitting the asymmetric data which widely existed in the dataset and the real-world. Moreover, we designed a Harmonic Sampling strategy for better forming the training pairs and making full use of the rank information. Experiments over the Contend-Based Video Relevance Prediction dataset show that the proposed method achieve superior performance when compared with the baseline method. In future work, we will focus on analyzing the temporal feature of the video data and designing a more precise video relevance predictor.

ACKNOWLEDGMENT

This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, the National Natural Science Foundation of China under contract No. U1611461, No. 61825101, No. 61425025.

REFERENCES

[1] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. Ieee, 2008, pp. 263–272.

[2] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[3] Y. Zheng, B. Tang, W. Ding, and H. Zhou, "A neural autoregressive approach to collaborative filtering," *arXiv preprint arXiv:1605.09477*, 2016.

[4] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1235–1244.

[5] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and evaluating choices in a virtual community of use," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1995, pp. 194–201.

[6] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens: an open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. ACM, 1994, pp. 175–186.

[7] U. Shardanand and P. Maes, "Social information filtering: algorithms for automating word of mouth," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1995, pp. 210–217.

[8] Y. Deldjoo, M. Elahi, P. Cremonesi, F. Garzotto, P. Piazzolla, and M. Quadrana, "Content-based video recommendation system based on stylistic visual features," *Journal on Data Semantics*, vol. 5, no. 2, pp. 99–113, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[10] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, "Advances in knowledge discovery and data mining," 1996.

[12] J. Wei, J. He, K. Chen, Y. Zhou, and Z. Tang, "Collaborative filtering and deep learning based recommendation system for cold start items," *Expert Systems with Applications*, vol. 69, pp. 29–39, 2017.

[13] A. Sepliarskaia, J. Kiseleva, F. Radlinski, and M. de Rijke, "Preference elicitation as an optimization problem," in *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 2018, pp. 172–180.

[14] D. Billsus and M. J. Pazzani, "A hybrid user model for news story classification," in *UM99 User Modeling*. Springer, 1999, pp. 99–108.

[15] D. Kelly and J. Teevan, "Implicit feedback for inferring user preference: a bibliography," in *Acm Sigir Forum*, vol. 37, no. 2. ACM, 2003, pp. 18–28.

[16] C. Musto, F. Narducci, P. Lops, G. Semeraro, M. De Gemmis, M. Barbieri, J. Korst, V. Pronk, and R. Clout, "Enhanced semantic tv-show representation for personalized electronic program guides," in *International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2012, pp. 188–199.

[17] M. Liu, X. Xie, and H. Zhou, "Content-based video relevance prediction challenge: Data, protocol, and baseline," *arXiv preprint arXiv:1806.00737*, 2018.

[18] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[19] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 539–546.

[20] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *null*. IEEE, 2006, pp. 1735–1742.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[22] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[24] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. Ieee, 2009, pp. 248–255.