역전파를 이용한 개집합 도메인 적응

배경호, 이효건, 최진우
경희대학교
Bkh178@khu.ac.kr / gunsbrother@naver.com / jinwoochoi@khu.ac.kr

Open Set Video Domain Adaptation by Backpropagation

Kyungho Bae Hyogun Lee Jinwoo Choi
Kyung Hee University

요약

기존의 video domain adaptation은 closed set 환경에서 주로 연구되었다. 하지만 이는 source와 target의 label이 같다는 비현실적인 전제를 요구한다. 따라서 본 논문에서는 target의 label space가 source보다 넓은 open set video domain adaptation 문제를 다룬다. 우린 open set image domain adaptation에서 사용되는 방법들을 video로 확장 시켜 모델을 설계하고 UCF to HMDB, HMDB to UCF 와 같은 video dataset에서 실험하였다. 그 결과 source only 대비 UCF to HMDB에서 12%, HMDB to UCF 7.1% 향상된 결과를 얻었다.

1. 서론

Domain adaptation(DA)[13]은 source labeled dataset 에서 학습된 model 이 target domain 에서 잘 일반화 되지 않는 문제를 다루기 위해 최근 방대하게 연구되어- 왔다[14]. 특히 target domain 에 대해 label 이 존재하지 않는 unsupervised domain adaptation (UDA) 인 경우 막대한 target domain dataset 에 대한 labeling cost 가 필요하지 않아 효율적이다. video-based UDA 인 경우는 image 와 다르게 spatial 영역과 temporal 영역 동시에 domain discrepancy 가 발생할 수 있다[15].

따라서 frame, video, temporal relation 별로 다르게 align 시키는 방법으로 제안되었다[12,15]. 하지만 존재하는 video-based UDA 는 대부분 close-set 에서의 방법으로 open-set 상황에는 적용되지 않는다.

따라서 본 연구에서는 open-set image-based UDA 의 가장 대표적인 방법인 unknown backpropagation [16] 을 open-set video-based UDA 환경에 적용해보고 기존의 close-set UDA method 인 DANN[17] 과 비교하여 효과가 있는지 검증해본다.

2. 관련 연구

따라서 본 연구에서는 open-set image-based UDA 의 가장 대표적인 방법인 unknown backpropagation [16] 을 open-set video-based UDA 환경에 적용해보고 기존의 close-set UDA method 인 DANN[17] 과 비교하여 효과가 있는지 검증해본다.

“이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022R1F1A1070997).”

Domain shift는 data sample이 다른 distribution에서 가져올 때 발생한다. 따라서 특정 domain에서 학습한 모 델은 다른 domain에서 잘 작동하지 않는 문제가 생긴다. 이러한 문제를 해결하는 domain adaption 기술은 보통 label이 풍부한 source domain에서 label이 거의 없거나 없는 target domain에 전달 가능한 지식을 배움으로써 그 차이를 극복한다. domain adaptation은 image classification[1,2]이나 semantic segmentation[3]에서 다양하게 활용되어왔다.

Domain adaptation에는 크게 두 가지 방향으로 발전 되어왔다. 사전에 정의된 두 domain의 차이나 loss를 mi- nizing하거나 adversarial 학습을 이용하는 방법이다. Hu et al [4] 는 domain간의 distribution divergence와 marginal Fisher analysis criterion을 모두 minimizing하는 방 식을 사용했다. Long et al. [5]는 maximum mean discrepancy (MMD) metric 을 minimizing 하는 방법으로 domain invariant feature을 학습했다. GAN[6]을 활용한 방식으로는 Tzeng et al. [7]은 GAN loss를 활용한 adversarial discriminative domain adaptation (ADDA) framework를 제안했다.

현재 존재하는 Domain adaptation의 대부분은 image classification과 semantic segmentation 등 2D Image task에 집중되어있다. 그에 반에 본 연구에서는 3D video action recognition에 대해 연구한다. 초기의 video domain adaptation[8,9] 은 작은 scale 의 video data 에서만 적용 가능하여 실제적인 domain discrepancy 를 줄 이지 못하였고 biased된 결과를 학습했다. 최근의 연구에 서는 Munro et al.[11]에서는 multi-modal 환경에서 self- supervision method를 사용하여 효과적으로 video DA를 수행하였다. 또한 Choi et al.[12]는 모든 sampled된 clip들 이 relevant semantic이 아니므

Method	UCF_HMDB U-7 -> H-14 TOP-1 acc (%)	△	UCF_HMDB H-7 -> U-14 TOP-1 acc (%)	△
Source-only	51.2 ± 0.05		57.2 ± 0.04	
ClosedSet DA (DANN w/o known/unknown classification)	54.8 ± 0.18	+ 3.6	61.5 ± 0.58	+ 4.3
OpenSet DA[1] (w/ known/unknown classification)	63.2 ± 0.32	**+ 12.0**	64.3 ± 0.39	**+ 7.1**
Target-only	86.9 ± 0.04		89.2 ± 0.02	

표 1

로 sub-optimal solution을 초래한다고 주장하였다. 따라서 attention mechanism을 통해 informative clip과 non-informative clip을 구분하여 align 하였다.

UDA 관련 연구는 대부분 close-set 으로 source와 target이 동일한 label 범위를 갖는다. 이는 비현실적이고 따라서 현실적인 문제를 다루기 위해 제시된 open-set domain adaptation (OSDA)는 source와 targe이 동일한 label space를 가지지 않는다. 현재까지 연구방향은 부분적으로 source에 align 시킨 뒤 mapping distance를 줄이는 과정을 반복[18] 하는 방법이나 사전 학습된 threshold로 unknown 과 known의 boundary를 학습하는 방법[16]으로 해결해왔다. 우리가 다루는 문제는 Video Open-set Unsupervised Domain Adaptation(VOSUDA)로 기존의 문제를 video로 확장시켜 본다.

3. 설계

labeled source video x_S 와 그에 대응하는 label y_s 라 고 하면 source domain은 $\{X_s, Y_s\}$, target dataset에는 label이 없으므로 x_t 이다. 모델의 전체 구조는 모델의 전체 구조는 x_s, x_t, 를 input으로 사용하는 feature generation network G, K + 1 classes를 분류하는 network C로 이루어져 있다. 여기서 K는 known class의 개수이고 K+1은 unknown을 예측한다.

여기서 목표는 unknown과 known을 구별하는 boundary를 generator G가 학습하는 것이다. 따라서 unknown data에 대한 정보가 필요하지만 target domain 에 대한 label 정보는 존재하지 않는다. 따라서 pseudo decision boundary를 학습하게 하여 목표를 달성한다. generator G와 classifier C 사이에는 gradient reverse layer가 존재하여 generator는 classifier를 속이기 위해 학습된다. 만약 classifier가 $p(y = K+1|x_t) = 1$ 을 학습하고 generator가 이를 속이기 위해 학습된다면 source와 target은 정확하게 align되어 unknown sample을 검출하지 못한다. 따라서 $p(y = K+1|x_t) = t$ 에서 t값을 $0 < t < 1$로 적당히 조절하여 목표를 달성한다. known class에 대해서는 cross-entropy를 활용하여 학습한다.

unknown sample에 대한 boundary를 학습 할 때는 binary cross entropy loss를 활용하여 학습한다.

$$L_s(x_s, y_s) = -\log(p(y = y_s|x_s))$$
$$p(y = y_s|x_s) = (C \circ G(x_s))_{y_s}$$

실험에는 [16]과 같게 t를 0.5로 설정한 뒤 실험한다.

$$L_{adv}(x_t) = -t log(p(y = K+1|x_t)) - (1-t)\log(1-p(y = K+1|x_t))$$

total object는 아래와 같다.

$$\min_C L_s(x_s, y_s) + L_{adv}(x_t)$$
$$\min_G L_s(x_s, y_s) - L_{adv}(x_t)$$

4. 실험

Open-set video based UDA 실험을 위해 UCF101(U) [19], and HMDB51(H) [20]를 사용했다. UCF101 은 101의 action classes와 13,320개의 clip으로 이루어져 있고 HMDB51은 51개의 action classes와 6,766개의 clip으로 이루어져 있다. [21] setting과 동일하게 7개의 공유하는 label을 만들고 각각 7개의 label을 설정하여 unknown label로 지정하였다. generator는 ImageNet-1K [22] pretrained ResNet50 [23] 을 사용한 TSM[24]를 사용하였고 classifier는 3층의 fc layer를 사용하였다.

source only는 source domain data만을 이용해 학습 한 뒤 평가한 것이고 CloseSet DA는 DANN을 적용한 뒤 평가한 것이다. 두 방법은 unknown에 대해 판별하지 못하므로 공평한 비교를 위해 threshold를 통해 unknown을 분류하였다. OpenSet DA는 위에서 설명한 [16]을 적용한 것이고 target only를 target domain data의 label을 준 뒤 학습한 값으로 upper boundary이다.

실험은 총 3번 진행한 뒤 평균을 내었다. 실험 결과는 표 1 에서 볼 수 있다. 실험 결과를 보면 domain gap 과 unknown sample로 인해 source only와 target only의 차 이가 각각 35.7%, 32%로 크게 발생함을 알 수 있다. DANN을 적용하였을 때 각각 3.6%, 4.3%의 향상이 있었지 만 DANN은 ClosedSet DA이므로 큰 향상은 이루지 못했 다.

하지만 OpenSet DA를 적용하였을 때 각각 12%, 7.1% 의 향상으로 이는 [16]의 method가 openset video based UDA에도 잘 동작함을 실험적으로 알 수 있다.

5. 결론

본 연구에서는 비교적 연구가 적은 open set video based UDA 환경에서의 image based method를 직접 적 용해 봄으로써 효과가 있음을 알아보았다. 현재 대부분의 video based UDA는 frame, video feature별로 다르게 attention하는 방식이 주를 이루고 있는데 향후 연구로 frame feature별로 unknown 에 대해 학습하는 모델 개발 로 진행 할 것이다.

참고문헌

[1] Timnit Gebru, Judy Hoffman, and Fei Fei Li. Fine-grained recognition in the wild: A multi-task domain adaptation approach. In ICCV, pages 1358-1367, 2017.

[2] Piotr Koniusz, Yusuf Tas, and Fatih Porikli. Domain adaptation by mixture of alignments of second- or higher-order scatter tensors. In CVPR, 2017.

[3] Zhang Yang, Philip David, and Boqing Gong. Curriculum domain adaptation for semantic segmentation of urban scenes. In ICCV, 2017

[4] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Deep transfer metric learning. In CVPR, pages 325-333, 2015.

[5] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I.Jordan. Learning transferable features with deep adaptation networks. In ICML, pages 97-105, 2015.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, BingXu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In NIPS, pages2672-2680. 2014.

[7] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain

adaptation .CVPR, pages 2962-2971, 2017

[8] Waqas Sultani and Imran Saleemi. Human Action Recognition across Datasets by Foreground-Weighted Histogram Decomposition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[9] Arshad Jamal, Vinay P Namboodiri, Dipti Deodhare, and KS Venkatesh. Deep Domain Adaptation in Action Space. In The British Machine Vision Conference (BMVC), 2018

[10] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A

comprehensive study of deep video action recognition. arXiv preprint arXiv: 2012.06567, 2020.

[11] Jonathan Munro and Dima Damen. Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

[12] Jinwoo Choi, Gaurav Sharma, Samuel Schulter, and Jia-Bin Huang. Shuffle and attend: Video domain adaptation. In European Conference on Computer Vision, pages 678-695. Springer, 2020

[13] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering (TKDE), 22(10):1345-1359, 2010.

[14] Gabriela Csurka. A comprehensive survey on

domain adaptation for visual applications. In Domain Adaptation in Computer Vision Applications, pages 1-35. Springer, 2017.

[15] Min-Hung Chen, Zsolt Kira, Ghassan AlRegib, Jaekwon Yoo, Ruxin Chen, and Jian Zheng. Temporal attentive alignment for large-scale video domain adaptation. In International Conference on Computer Vision (ICCV), October 2019

[16] Saito, K.; Yamamoto, S.; Ushiku, Y.; and Harada, T. 2018b. Open set domain adaptation by backpropagation. In ECCV, 153-168.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franc ¸ois Laviolette,

Mario Marchand, and Victor Lempitsky. Domain adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1-35, 2016.

[18] Pau Panareda Busto and Juergen Gall, "Open set domain adaptation," in ICCV, 2017.

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak

Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," arXiv preprint arXiv:1212.0402, 2012.

[20] Hildegard Kuehne, Hueihan Jhuang, Est´ıbaliz Garrote, Tomaso Poggio, and Thomas Serre, "Hmdb: a large video database for human motion recognition," in

ICCV, 2011.

[21] Xu, Y.; Yang, J.; Cao, H.; Li, Q.; Mao, K.; and Chen, Z. 2021a. Partial Video Domain Adaptation with Partial Adversarial Temporal Attentive Network. arXiv preprint arXiv:2107.04941.

[22] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in CVPR, 2009.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in CVPR, 2016.

[24] Ji Lin, Chuang Gan, and Song Han. Temporal shift module for efficient video understanding. CoRR, abs/1811.08383, 2018. 8