균일한 부류 확률값 학습을 통한 도메인 일반화

윤성준 심규진 김창익
한국과학기술원
{sungjoonoy, kjshim1028, changick} @kaist.ac.kr

Domain Generalization via Class Balanced Probability Learning

Yoon, Sungjoon Shim, Kyujin Kim, Changick
Korea Advanced Institute of Science and Technology

요약

본 논문에서는, 영상 분류 문제에서 손실 값 계산 시 정답 부류를 제외한 나머지 부류에서 우세한 결괏값이 나오지 않도록 평활화하는 보조적인 손실함수를 고안한다. 합성곱 신경망 구조를 이용해 학습이 진행되면 손실함수가 작아지는 방향으로 가중치가 갱신되기 때문에, 정답을 제외한 나머지 부류들의 결괏값은 줄어든다. 하지만, 정답을 제외한 나머지 부류들 사이의 상대적인 값이 고려되지 않고 손실함수가 줄어들기 때문에 값들은 균일하지 않게 되고, 정답 부류와 유사한 특징을 가진 부류들의 값이 상대적으로 커지게 된다. 이는 정답 부류와 나머지 부류 중 가장 값이 큰 부류 사이에 공통의 특징을 공유한다고 생각할 수 있다. 정답 부류만이 가지고 있는 고유의 특징을 추출하지 못하고, 다른 부류도 가지고 있는 특징의 흔적이 남아있게 됨으로써 테스트 시 소스 도메인과 전혀 다른 도메인의 영상이 보일 때 그러한 특징이 부각 되어 부정확한 결과를 초래하게 된다. 본 논문에서는 단순한 손실함수의 추가로 도메인이 다른 환경에서 기존의 연구보다 좋은 분류 결과를 보여주는 것을 실험을 통해 확인하였다.

1. 서론

인공 신경망 구조의 눈부신 발전은 컴퓨터 비전의 다양한 문제를 해결해 왔으며 특히, 영상 분류 작업에서 놀라운 성능을 보여주었다. 그러나 테스트 영상의 환경이 학습 환경과 다를 경우 성능이 급격하게 저하되는 문제 또한 보여주었다 [1]. 예를 들어 영상 분류 문제에서, 강아지 사진을 분류할 수 있는 모델을 학습한 뒤 다른 도메인인 강아지 그림을 분류하기 위해 테스트하면, 분류 모델은 영상을 강아지로 잘 판단하지 못한다. 이러한 문제를 해결하기 위해 도메인 적응 기법 연구들이 많이 제안되었다 [2, 3, 4, 5].

도메인 적응의 핵심 생각은, 소스 도메인과 타겟 도메인이 상이해서 테스트 성능이 떨어지므로 타겟 도메인의 정답이 주어지지 않은 몇 개의 영상을 학습 시 이용하여 테스트 성능을 올리자는 것이다. 일반적으로 영상에 정답을 매기는 것은 많은 작업을 요구하기 때문에, 정답이 주어지지 않은 테스트 영상으로 타겟 도메인의 성능을 높이는 것은 굉장히 실용적이다.

하지만, 타겟 도메인의 데이터가 주어지지 않을 때는 도메인 적응법을 적용하기 어려워진다. 가령 자율주행 환경에서 모든 교통 환경에 대한 데이터를 수집하여 학습하기는 사실상 불가능하고 또 큰 비용이 들기 때문에 적합하지 않다. 그래서 최근에는 타겟 도메인 영상을 전혀 사용하지 않고 소스 도메인의 데이터만으로 훈련하고도 테스트 시 좋은 성능을 끌어낼 수 있는 도메인 일반화 기법들이 소개되고 있다 [6, 7, 8, 9].

먼저 도메인에 불변한 특징을 학습하여 도메인 일반화를 하는 방법들이 있다 [6, 9, 10]. 최대 평균 불일치라는 거리 분포를 이용하여 서로 다른 도메인의 영상 사이의 특징 분포를 일치시킴으로써 도메인 불변 특징을 학습하기도 하고 [6], 직소 퍼즐을 맞추는 자기 지도 학습 신호를 이용해 도메인의 관계없는 일반적인 특징을 추출하여 도메인 일반화의 성능을 향상하기도 한다 [9]. 또, 최근에는 합성곱 신경망 구조의 낮은 단계 특징이 영상의 도메인과 연관이 있다는 사실로부터 의사 도메인을 지정해 도메인 분류기가 도메인을 분류할 수 없도록 적대적으로 손실함수를 주어 도메인 불변 특징을 학습한 방법이 소개되었다 [10].

두 번째로는 학습 데이터를 증대하는 방법들이 있다 [11, 12]. 학습 데이터의 범위가 넓어지면 타겟 도메인과 유사한 데이터를 학습할 가능성이 커지기 때문에 일반화 성능이 개선된다. 먼저, 도메인 분류 신호를 기반으로 입력 데이터를 교란하여 데이터를 증대함으로써 도메인 일반화에서 좋은 성능을 보인 연구가 발표된 바 있다 [11]. 최근 연구에서는 데이터 생성기를 사용하여 영상을 합성하여 소스 도메인의 범위를 확장하기도 하였다 [12].

마지막으로 전통적인 일반화 기법을 이용하여 학습 데이터에 과적합 되는 것을 방지하여 일반화하는 방법이 있다 [13, 14]. 손실함수가 작아지는 방향으로만 단순하게 학습을 진행하면 특정 가중치 값들이 커지면서 학습성능이 악화할 수 있다. 가중치 감쇠는 학습된 모델의 복잡도를 줄이기 위해서 학습 중 가중치가 너무 큰 값을 가지지 않도록, 가중치가 큰 경우에 대한 페널티 항목을 손실함수에 추가하여 모델을 일반화한다 [13]. 드롭아웃은 훈련 단계 동안 신경망의 뉴런을 무작위로 '0'으로 설정함으로써 해당 뉴런의 기능을 제거하여 전체 네트워크가 작은 하위 네트워크로 조합되는 효과를 통해 모델 일반화 효과를 준다 [14].

본 논문에서는 도메인에 불변한 특징을 학습하는 것과 가깝지만, 도메인에 불변한 특징이 아닌, 부류에 불변한 특징을 더 잘 학습하도록 하

는 손실함수를 새롭게 제안한다. 영상 분류 문제에서는, 모델이 각 부류의 고유한 특징을 추출하는 것이 중요하다. 만약 각 부류의 고유한 핵심 특징을 추출할 수 있다면, 영상이 훈련 시 보지 않은 다른 도메인이더라도 해당 부류의 핵심 특징은 포함될 가능성이 크기 때문에 정확한 추론을 할 수 있을 것이다. 부류의 핵심 특징이란 다른 모든 부류와 공통으로 공유하는 특징을 제외하고는 나머지 부류에 대해서는 나타나지 않는 특징이다. 하지만 일반적으로 분류에 사용하는 교차 엔트로피 손실함수는 정답 부류에 대한 지도만이 존재하므로 훈련을 마친 후에 정답이 아닌 부류들의 확률값 분산이 크다. 즉, 정답 부류가 어떤 특정 부류와는 전체 부류의 공통 특징이 아닌 특징을 공유하고 있다는 것이다. 따라서, 정답이 아닌 부류들에 대한 결괏값이 고르게 나타나도록 추가적인 손실을 주어 모델이 정답 부류의 핵심 특징을 학습할 수 있도록 한다. 그럼으로써 모델은 부류에 대해 불변한 특징들을 추출할 수 있고 도메인 일반화에 좋은 성능을 가지게 된다.

2. 본론

인공 신경망 구조는 훈련 데이터가 정답일 가능성을 손실함수로 계산하여 이 손실함수를 최적화하는 방향으로 매개변수 θ를 다음과 같이 학습한다.

$$\hat{\theta} = \arg\min_{\theta} \sum_{<x,y>} l(f(x;\theta), y) \quad (1)$$

이때 정답일 가능성을 측정하는 손실함수로 교차 엔트로피를 널리 사용한다. 교차 엔트로피를 최소화하는 것은 정답 부류의 가능성을 최대화하는 것과 같기 때문이다. 이때 손실함수 l로 사용하는 교차 엔트로피 H는 다음과 같이 정의된다.

$$\begin{aligned} l &= H(f(x;\theta), y) \\ &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} 1_{[y_{i,j}=1]} \log\left(f(x_i;\theta)_j\right) \end{aligned} \quad (2)$$

x_i는 i 번째 학습 영상, y_i는 i 번째 학습 영상의 정답 one-hot vector, B는 배치 크기, C는 부류의 개수, 함수 f는 합성곱 신경망을 나타내는 함수이다. f의 출력값은 softmax가 취해진 부류의 확률값이다. 교차 엔트로피는 정답 부류의 확률값이 1이 되도록 한다. 수식 2에서 볼 수 있듯 정답이 아닌 부류의 확률값은 손실함수 계산 시 전혀 고려되지 않기 때문에 기울기 갱신에도 사용되지 않으며, 정답이 아닌 부류들 사이의 상대적인 확률값 또한 고려되지 않는다.

그림 1은 교차 엔트로피 손실함수를 사용하여 PACS 데이터 세트를 학습한 분류 모델에서, 정답 부류 "개"를 제외한 나머지 부류들의 교차 엔토르피 값을 그래프로 나타낸 것이다. 정답을 제외한 나머지 부류들의 결괏값은 고르게 분포되지 않는 것을 확인할 수 있다. 이는 곧 정답이 아닌 부류 중에서, 정답과 더 많은 특징을 공유하고 있는 부류들이 있다고 생각할 수 있으며 해당 특징이 타겟 도메인에서 두드러지어 오류가 발생할 수 있다. 만일 합성곱 신경망 구조가 정답 부류 고유의 특징을 추출한다면, 나머지 부류들의 결괏값은 0에 가까운 값을 가지면서도 서로 비슷한 값을 가질 것이다. 나아가 부류 불변한 특징을 추출 함으로써 다른 도메인에서도 그러한 특징을 잡아내어 더 좋은 분류를 할 수 있을 것이다. 우리는 이러한 동기를 가지고 다음과 같은 손실함수 l'을 추가로 설계하였다.

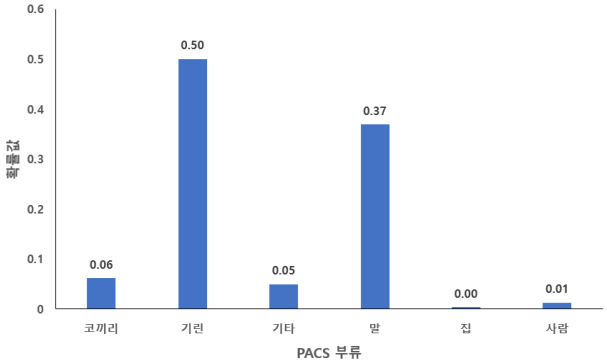

그림 1. 정답 부류(개)를 제외한 나머지 부류들의 확률값 분포

$$\begin{aligned} l' &= H(f(x;\theta)', (1/(C-1))^{C-1}) \\ &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{C} 1_{[y_{i,j}\neq 1]} \frac{1}{C-1} \log\left(f(x_i;\theta)'_j\right) \end{aligned} \quad (3)$$

$f(x;\theta)'$은 합성곱 신경망을 통과한 부류 결괏값들 중 정답 부류를 제외한 나머지 값들의 확률값이다. 이 확률값들이 고르게 분포하도록 one-hot vector가 아닌, $1/(C-1)$의 균등한 확률값으로 정답을 주어 교차 엔트로피를 계산한다. 추가적인 손실함수를 통해 나머지 부류들의 확률값은 평활화된다.

이제 정답 부류를 맞추기 위해 정답 부류의 확률값을 1이 되도록 학습하는 원래의 손실함수와 정답을 제외한 나머지 부류들의 확률값을 평활화하기 위한 두 번째 손실함수를 λ를 곱한 적절한 비율로 더하여 최종 손실함수 l_{total}을 다음과 같이 계산한다.

$$l_{total} = l + \lambda l' \quad (4)$$

3. 실험 및 결과

분류 작업에서의 도메인 일반화 성능을 평가하기 위해 일반적으로 널리 사용되는 데이터 세트 PACS [15] 을 이용해 모델을 평가하였다. PACS는 7개의 부류와 4개의 도메인 (사진, 예술 그림, 만화 및 스케치)을 다룬다. 우리는 3개의 도메인을 학습 데이터 세트로, 나머지 하나의 도메인을 테스트 데이터 세트로 고려하여 모델을 훈련하였다.

기본적으로 우리는 JiGen [9] 에서 사용하는 하이퍼 파라미터와 이미지 전처리를 사용하였다. 모델은 Resnet-18 [16] 을 사용하였으며 모멘텀 0.9, 가중치 감소 5e-4로 설정된 확률적 경사 하강법(SGD)를 사용하였다. 미니배치 크기는 128으로 설정하였고 30 epoch 동안 모델을 훈련하였다. 초기 학습률을 0.004을 시작으로 훈련 epoch의 80% 후에 0.1 배로 감소시켰다. 영상 전처리로 영상을 기존 크기의 0.8배 비율로 무작위하게 잘랐으며 0.5의 확률로 가로로 뒤집었다. 밝기, 대비, 채도,

Dataset	PACS (Resnet-18) (%)				
Domain	Art paint	Cartoon	Sketch	Photo	Avg.
MASF [17]	80.29	77.17	71.69	94.99	81.03
Epi-FCR [8]	82.10	77.00	73.00	93.90	81.50
JiGen [9]	79.42	75.25	71.35	96.03	80.51
MetaReg [7]	**83.70**	77.20	70.30	95.50	81.70
MMLD [10]	81.28	77.16	72.29	**96.09**	81.83
Ours	82.35	**77.67**	**72.83**	96.05	**82.22**

표 1. PACS 데이터 세트에 대한 도메인 일반화 성능 결과

Dataset	PACS (Resnet-18) (%)				
Domain	Art paint	Cartoon	Sketch	Photo	Avg.
Baseline ($\lambda=0$)	80.36	75.77	71.07	95.28	80.62
Ours ($\lambda=0.5$)	**82.35**	77.67	**72.83**	**96.05**	**82.22**
Ours ($\lambda=1$)	81.78	**78.52**	72.54	95.68	82.13

표 2. 제안모델의 $\lambda \in \{0.5, 1\}$에 대한 Baseline과의 성능 비교

색조를 랜덤하게 변경하였고, ImageNet의 이미지 통계를 사용하여 학습 데이터와 테스트 데이터를 정규화하였다. 테스트 정확도는, 유효성 검사에서 최고의 성능을 달성할 때의 값으로 사용하였다.

표 1은 PACS 데이터 세트에 관하여 다른 모델들과 우리가 제안한 모델과의 비교결과이다. 각 열의 도메인들은 타겟 도메인이고 이를 제외한 나머지 도메인으로 모델을 학습하였다. 결과는 다섯 번 시도한 값들의 평균이다. 실험적으로 우리 최종 모델은 λ를 0.5로 설정 하였다. 표에서 볼 수 있듯이, 우리는 단순한 손실함수 추가만으로 이전의 방법들보다 더 좋은 성능을 달성함을 확인할 수 있었고 평균값이 최대 1.71%의 좋은 성능을 보였다.

표 2는 우리가 제안한 손실함수 l'을 사용하지 않은 Baseline과 제안한 손실함수를 사용한 결과를 비교한 것이다. λ 값은 0.5와 1일 때 비교하였으며 제안하는 손실함수를 사용할 경우 $\lambda=0.5, 1$일 때 모두 Baseline보다 성능이 개선되었다. 또한, $\lambda=0.5$일 때 최고 성능을 보였으며 Baseline보다 평균 1.60% 더 성능이 좋았다. λ값에 따른 실험 결과를 통해 정답 부류의 확률값 오류를 줄이기 위한 기본 손실함수 l과 제안된 손실함수 l'의 적절한 비율이 중요함을 알 수 있다.

4. 결론

우리는 정답 부류가 아닌 나머지 부류들의 확률값을 평활화함으로써 모델이 도메인 일반화를 이전 연구보다 효과적으로 달성할 수 있음을 보여주었다. 교차 엔트로피 손실함수는 정답 부류만이 가지고 있는 고유의 특징을 추출하지 못하고 다른 부류 또한 정답 부류의 특징 흔적이 남아있게 한다. 새롭게 제안하는 추가 손실함수를 사용함으로써 테스트 시 소스 도메인과 전혀 다른 도메인의 영상이 주어질 때 그러한 특징이 부각 되어 부정확한 결과를 초래하는 것을 완화할 수 있었다. 본 방법은 추가 모델 구조 없이 영상 분류를 위한 모든 합성곱 신경망 구조에 직접 적용할 수 있는 간단한 손실함수이며 실험을 통해 이전의 방법들보다 효과적으로 도메인 일반화할 수 있음을 보여주었다.

참고문헌

[1] Torralba, A., and Efros, A. A. Unbiased look at dataset bias. In CVPR, 2011.

[2] Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. Unsupervised pixel-level domain adaptation with generative adversarial networks. In CVPR, 2017.

[3] Chen, Y.-C., Lin, Y.-Y., Yang, M.-H., and Huang, J.-B. Crdoco: Pixel-level domain transfer with cross-domain consistency. In CVPR, 2019.

[4] Long, M.:,Zhu, H., Wang, J., and Jordan, M. I. Unsupervised domain adaptation with residual transfer networks. In NeurIPS, 2016.

[5] Ganin, Y., and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In ICML, 2015.

[6] Muandet, K., Balduzzi, D., and Sch¨olkopf, B. Domain generalization via invariant feature representation. In ICML, 2013.

[7] Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using meta-regularization. In NeurIPS, 2018.

[8] Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.Z., and Hospedales, T.M. Episodic Training for Domain Generalization. In ICCV, 2019.

[9] Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. Domain generalization by solving jigsaw puzzles. In CVPR, 2019.

[10] Matsuura, T., and Harada, T. Domain Generalization Using a Mixture of Multiple Latent Domains. In AAAI, 2020.

[11] Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S. Generalizing across domains via cross-gradient training. In ICLR, 2018.

[12] Zhou, K., Yang, Y., Hospedales, T., and Xiang, T. Learning to Generate Novel Domains for Domain Generalization. In ECCV, 2020.

[13] Nowlan, S.J., and Hinton, G.E. Simplifying neural networks by soft weight-sharing. In Neural computation, 1992.

[14] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. In the journal of machine learning research, 2014.

[15] Li, D., Yang, Y., Song, Y.Z., and Hospedales, T.M. Deeper, broader and artier domain generalization. In ICCV, 2017.

[16] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In CVPR, 2016.

[17] Dou, Q., Castro, D.C., Kamnitsas, K., and Glocker, B. Domain generalization via model-agnostic learning of semantic features. In NeurIPS, 2019.