Controllable Invariance through Adversarial Feature Learning

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, Graham Neubig

Carnegie Mellon University
Language Technologies Institute

NIPS 2017

Outline

Introduction

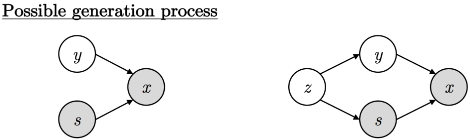
Carnegie Mellon University

- Representations with invariance properties are often desired
  - Spatial invariance: CNN
  - Temporal invariance: RNN
- This work: a generic framework to induce invariance to a specific factor/attribute of data
  - Image classifications: classifying people's identities invariant to lighting conditions
  - Multi-lingual machine translation (fr-en, de-en): translation invariant to source language for sentences with the same meaning
  - Fairness classifications: predicting credit and saving conditions invariant to the age, gender and race of a person

Problem formulation

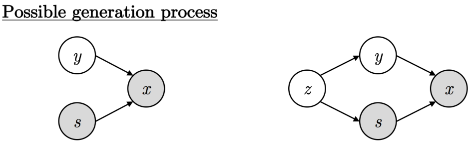
Task:

- Given input x (images, sentences or features), attribute s (can be discrete, continuous or structured) of x
- Predict target y
- Prior belief: Prediction **should** be invariant to s
- e.g., predicting identities of a person in a image. s is the lighting condition
- Two possible data generation processes:



Discriminative model

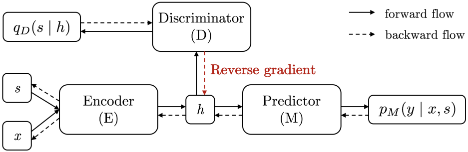
- y and s are not independent given x although they can be marginally independent (Explaining-away)
- $p(y \mid x, s)$ is more accurate than $p(y \mid x)$, i.e., knowing s helps in inferring y.
  - "brighten" the representation if it knows the original picture is dark



- Encoder E: obtain the invariant representation $h = E(x, s)$. (s is used as the input of the encoder)
- Predictor M: Outputs $q_M(y \mid h)$ (predict y based on h)

Enforcing Invariance

- h is invariant to s means that $\nexists f : f(h) = s$
- Employ a Discriminator D to model f: Outputs $q_D(s \mid h)$ (predict s based on h)
- An adversarial game to enforce invariance:
  - Discriminator tries to detect s from the representation
  - Encoder learns to conceal it



Two objective

- Standard MLE loss: $\min_{E,M} -\log q_M(y \mid h = E(x, s))$
- Adversarial loss to ensure invariance: $\min_{E} \max_{D} \gamma \log q_D(s \mid h = E(x, s))$

Theoretical Analysis

- Overall objective:

$$\min_{E,M} \max_{D} J(E, M, D)$$

where $J(E, M, D)$ is

$$\mathbb{E}_{x,s,y\sim p(x,s,y)} \left[\gamma \log q_D(s \mid h = E(x, s)) - \log q_M(y \mid h = E(x, s))\right]$$

- Definition: $\tilde{p}(h, s, y) = \int_x p(x, s, y) p_E(h \mid x, s) dx$
- Claim 1: Given an encoder, the optimal discriminator and optimal predictor:
  - $q_D^*(s \mid h) = \tilde{p}(s \mid h)$ and $q_M^*(y \mid h) = \tilde{p}(y \mid h)$
  - Note that q_D and q_M are functions of E
- Claim 2: The optimal encoder is defined by:

$$E^* = \arg\min_E J(E) = \boxed{-\gamma H(\tilde{q}_E(s \mid h))} + \boxed{H(\tilde{q}_E(y \mid h))}$$

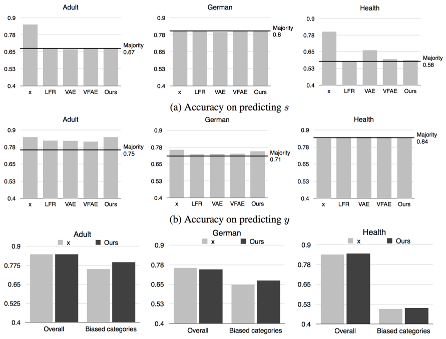
- [Red] maximizing the <u>uncertainty</u> of inferring s based on h
- [Green] increasing the <u>certainty</u> of predicting y based on h

Equilibriums Analysis

- The equilibrium of the minimax game is defined by $\min_E -\gamma H(\tilde{q}(s \mid h)) + H(\tilde{q}(y \mid h))$
- Win-win equilibrium:
  - s and y are marginally independent
  - Two entropy terms reach the optimum at the same time
  - e.g., removing the lighting conditions in image classifications results in better generalization
- Competing equilibrium:
  - s and y are NOT marginally independent
  - The optimal of the two entropies cannot be achieved simultaneously
  - Filtering out s from h does harm the prediction of y
  - e.g., removing bias in fairness classifications hurts the overall performance

Experiments: Fairness Classifications

- Task: Predict savings, credit and health condition based on features of a person. s can be gender or age
- E, M, D are all DNN



(a) Accuracy on predicting s

(b) Accuracy on predicting y

(c) Overall performance and performance on biased categories

Figure 1: Fair representations should lead to low accuracy on predicting factor s and high accuracy on predicting y.

Experiments: Multi-lingual Machine Translation

Carnegie Mellon University

- Task: Translation from German (de) and French (fr) to English. s indicates the source language (an one-hot vector)
- E, M, D are all LSTM
- Separate encoders for different languages (Recall that $h = E(x, s)$).
  - Sharing encoder does not work
  - DNN based discriminator (even with attention) does not work
  - Lesson: It is important for E, M, D to have enough capacity to achieve the equilibrium

Model	test (fr-en)	test (de-en)
Bilingual Enc-Dec [Bahdanau et al., 2015]	35.2	27.3
Multi-lingual Enc-Dec [Johnson et al., 2016]	35.5	27.7
Our model	**36.1**	**28.1**
w.o. discriminator	35.3	27.6
w.o. separate encoders	35.4	27.7

Table 1: BLEU score on IWSLT 2015. The ablation study of "w.o. discriminator" shows the improvement is not due to more parameters

Experiments: Image Classification

Carnegie Mellon University

- Task: classifying identities. s is the lighting condition
- E, M, D are DNN

Method	Accuracy of classifying factor s	Accuracy of classifying target y
Logistic regression	0.96	0.78
NN + MMD [Li et al., 2014]	-	0.82
VFAE [Louizos et al., 2016]	**0.57**	0.85
Ours	**0.57**	**0.89**

Table 2: Results on Extended Yale B dataset

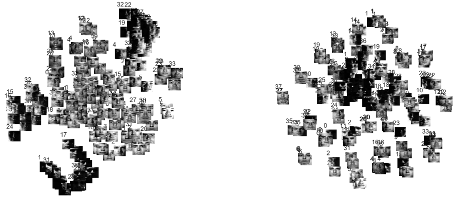


Figure 2: t-SNE visualizations of original pictures and learned representations. The original picture is clustered by lighting conditions. The learned representation is clustered by identities