Adversarial Examples を用いたドメイン適応

ER15048 遠山 光　　指導教授：藤吉弘亘

1. はじめに

ドメイン適応は, 教師ラベルを持つデータ (Source Domain) の情報を, 教師ラベルを持たないデータ (Target Domain) へ転移することで, Target Domain の認識精度を向上させるアプローチである. ドメイン適応の問題点として, Source Domain と Target Domain のドメイン間差が大きいほど適応が難しくなり, 精度が向上しにくい. そこで, 本研究では Adversarial Examples を用いて, Source Domain に対し, Target Domain へとドメイン間の差を近づけ, ドメイン適応の精度向上を図る手法を提案する.

2. ADDA

ドメイン適応手法である Adversarial Discriminative Domain Adaptation(ADDA) [1] は, Source Domain で事前学習した Source Encoder と, Target Domain を入力とする Target Encoder の出力を Discriminator へ入力して敵対的学習を行う. 敵対的学習を行うことで, Target Encoder の分類精度が向上する. 図 1 に ADDA のネットワーク構造を示す.

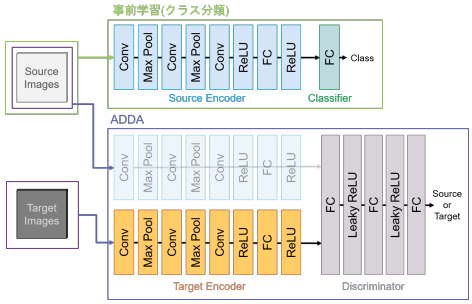


図 1：ADDA のネットワーク構造

3. Adversarial Examples

Adversarial Examples [2] は, 学習済みネットワークの勾配を利用し, 指定したクラスの微小ノイズ画像を生成するアプローチである. ノイズの生成は, 入力画像を認識した際に, クラス尤度が最も低いクラスを指定する方法や, 任意のクラスを指定する方法がある. ノイズ生成は式 (1) のように行う. x は入力画像, $\tilde{x}$ はノイズを付与した画像である.

$$\tilde{x} = x + \epsilon \mathrm{sign}(-\nabla_x J(\theta, x, t)) \quad (1)$$

ノイズ生成ネットワークのパラメータ θ は固定した状態で, 識別した際にラベルとの誤差が最小になるよう x を変化させるノイズを生成する. また, ϵ はノイズの倍率である.

4. 提案手法

本研究では, Adversarial Examples を用いることで Target Domain のクラス情報を Source Domain へとノイズとして付与することでドメイン間の差を近づけ, ドメイン適応の精度向上を図る手法を提案する. 図 2 に提案手法の構造を示す.

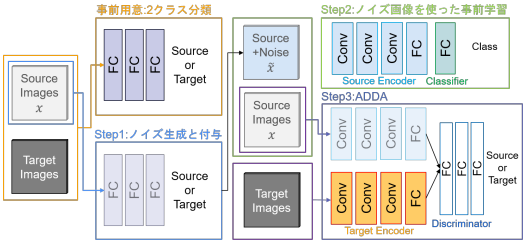


図 2：提案手法の全体構造

提案手法は以下の 3 つのステップで行う.

Step1:ノイズ生成と付与

Source Domain または Target Domain の 2 クラス分類を行うネットワークをあらかじめ学習する. そして, Source Domain の画像 x をネットワークに入力し, 2 クラス分類を行う. その結果をもとに式 (1) によりノイズを付与した画像 $\tilde{x}$ を生成する.

Step2:ノイズ画像を用いた事前学習

画像 x と画像 $\tilde{x}$ を用いて, クラス識別を行うネットワークを事前学習する. 事前学習では, Source Encdoer と Classifier の学習を行う.

Step3:ADDA によるドメイン適応

Source Encoder のパラメータを Target Encoder に転移し, Encoder と Discriminator を敵対的に学習する. Target Encoder は Discriminator に対し, 出力を Source として識別させるように学習する. また, Discriminator は入力が Source か Target かを識別できるように学習する.

5. 評価実験

本実験では, 提案手法の有効性を従来手法と比較することで評価する.

5.1. 実験概要

本実験では, 数字画像データセット SVHN を Source Domain, 手書き数字データセット MNIST を Target Domain として用いる. 分類対象は Target Domain 画像とし, 各 Encoder と Classifier を用いて分類を行う. 比較対象は, 従来手法, 提案手法の各 Source Encoder, Taregt Encoder での分類精度とする.

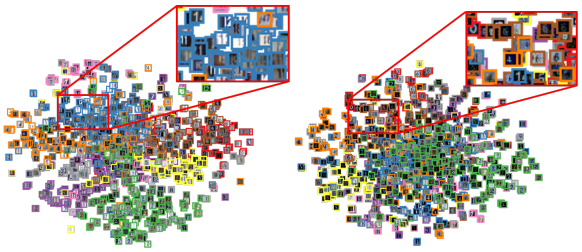
5.2. 実験結果

表 1 に評価結果を示す. 提案手法の精度は, 従来手法と同等であることが確認できる.

表 1：評価結果 [%]

	Source Encoder	Target Encoder
従来手法	64.7	80.0
提案手法	68.9	80.3

提案手法の精度が向上していない点として, ノイズにより本来のクラス情報が消失していると考えられる. そこで, t-SNE を用いて次元削減した従来手法の Source Encoder における SVHN と, ノイズを付与した SVHN の特徴ベクトルを確認する. 図 3 に特徴ベクトル分布を示す.



(a)SVHN　　(b)SVHN+Noise

図 3：Source Encoder の特徴ベクトル分布

図 3(a) では, 各クラスの特徴ベクトルがそれぞれまとまっていることが確認でき, クラス分類ができていることがわかる. 図 3(b) では, 図 3(a) に比べ各クラスが混在し, クラス分類ができていないことがわかる. この結果より, ノイズを付与することで, 画像本来のクラス情報が消失している. クラスを分離するようなノイズを生成し付与できるようにする必要がある.

6. おわりに

本研究では Adversarial Examples によるドメイン適応を提案し, 従来手法と同等の精度であった. 今後はノイズの付与条件の変更による精度変化について調査する.

参考文献

[1] E. Tzeng, *et al.*, "Adversarial Discriminative Domain Adaptation", CVPR, 2017.

[2] A. Kurakin, *et al.*, "Adversarial examples in the physical world", ICLR, 2017.