

11章 11.5.3まで

11.0 冒頭文

11.1 はじめに

前提条件1 データが少ない

そのための主な手法

前提条件2 ラベルなしデータ

それ以外の手法

11.2 データ拡張

11.2.1 基本的な考え方

前提条件 データ拡張

11.2.2 主に画像に対して有効な方法

11.2.3 主に言語に対して有効な方法

11.2.4 その他の方法

11.2.5 データ拡張の自動探索

11.3 転移学習

11.3.1 概要

11.3.2 特徴抽出器としての利用

11.3.3 ファインチューニング

11.3.4 ImageNet事前学習

11.3.5 事前学習の有効性

11.4 半教師あり学習

11.4.1 概要

11.4.2 一貫性正則化

11.4.3 疑似ラベル

11.4.4 エントロピー最小化

11.5 自己教師あり学習

11.5.1 概要

11.5.2 プレテキストタスク

11.5.3 対照表現学習

11.0 冒頭文

1. 本書で紹介している手法は、多くの訓練データが必要という制限がある
 - a. これを用意するのは難しい
 - b. 少数データで分析する方法を模索する
2. 以下の手法で、そのような制限下の分析を可能にする

- a. データ拡張
- b. 転移学習
- c. 半教師あり学習
- d. 自己教師学習
- e. マルチタスク学習
- f. ドメイン適応
- g. 少数事例学習
- h. 能動学習

11.1 はじめに

前提条件1 データが少ない

1. 手元のデータからタスク \mathcal{T} のために学習したい
 - a. 主にクラス分類を想定する
 - b. しかし、大部分はクラス分類以外でも使える
2. 手元のデータは $\mathcal{D} = \{(\mathbf{x}_n, d_n)\}_{n=1, \dots, N}$
 - a. N は小さい
 - b. だが、これを用いて学習用の準備を行う

そのための主な手法

1. データ拡張(11.2節)
 - a. 後述の手法(前述のc.からh.の手法)のもとになる
 - b. 単体でもよくつかわれる
2. 転移学習(11.3節)
 - a. 広い意味ではタスク \mathcal{T}' の学習経験を利用することを指す
 - i. 日経平均株価の予測で使ったモデルをダウ平均株価の予測に活かす、のような話？
 - b. 狭い意味だと特定の手法を指す

- i. ここでは具体的な言及なし

前提条件2 ラベルなしデータ

1. 手元のデータは $\mathcal{D}_{\text{UL}} = \{(\mathbf{x}_n)\}_{n=1, \dots, N}$
 - a. 正解ラベルがない
 - b. 半教師あり学習や自己教師あり学習は、このようなデータを対象としている

それ以外の手法

その他の手法の概要は、以下の通り。

マルチタスク学習	異なる複数タスクを1つのネットワークで学習する
ドメイン適応	訓練データとテストデータの間で統計的な分布にずれがある場合に適用する
少数事例学習	訓練データの数が極端に小さい場合を対象にした学習
能動学習	ラベルなしデータから限られたサンプルを選んで正解ラベルをつける(よいサンプルを選ぶ方法)

11.2 データ拡張

11.2.1 基本的な考え方

前提条件 データ拡張

1. $\mathcal{D} = \{(\mathbf{x}_n, d_n)\}_{n=1, \dots, N}$ をもっている
 - a. \mathbf{x}_n を適当な変換 φ で $\mathbf{x}'_n \equiv \varphi(\mathbf{x}_n)$ とする
 - b. 新たなサンプル (\mathbf{x}'_n, d_n) として利用する
2. 変換は自由度が高いが、以下の制約などがある
 - a. 変換によってラベルが変わらない
 - b. テストデータに出てきそうなデータにしないといけない
3. 数だけは際限なく増やすことができる

11.2.2 主に画像に対して有効な方法

- そもそもデータ拡張は画像データに対して有用
- 画像データを変化させても正解ラベルが変化しないような変換をする必要がある
- 図11.1にあるような変換を加える
- 図11.2にあるよう、正解ラベルが変わってしまいそうな加工等はよくない
 - また、タスクによって有効な加工が違う点にも注意
 - カットアウト、ランダム消去は物体認識に効果が高い
- 音声でも同様の有効な加工が存在する

11.2.3 主に言語に対して有効な方法

- 言語の加工は、意味が同じ言い換えを作る方法が考えられる
- 具体的には以下の方法がある
 - 文中の単語を類義語に入れ替える
 - 機械翻訳で別言語にしたものを再翻訳する
- 文の埋め込みベクトルにノイズを足す方法も考えられる
- 多様な言い換えができるかどうか重要

11.2.4 その他の方法

- クラス分類を対象とするmixupという方法がある
- 2つのサンプルを足し合わせる方法
- 特徴量、正解ラベルをそれぞれ適当な加重和で表現する
 - (\mathbf{x}_i, d_i) と (\mathbf{x}_j, d_j) について適当な λ で以下のようにサンプルを作る
 - $(\lambda \mathbf{x}_i + (1 - \lambda) \mathbf{x}_j, \lambda d_i + (1 - \lambda) d_j)$
- λ は適当なベータ分布からサンプルしたもの

11.2.5 データ拡張の自動探索

- 変換の候補が多数あるとき、その中から選択するのは重要
- バリデーションデータへの予測精度が最大化されるように変換パラメータを選択するような手法がある

- パラメータ空間の探索方法もよく提案されている

11.3 転移学習

11.3.1 概要

1. $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1, \dots, N}$ と \mathcal{D}' をもっている
 - a. 前者については量が十分でなく、後者のデータは十分にある
 - i. 新規に上場した銘柄の株価を既存の銘柄の株価から推定するような話？
 - b. \mathcal{D}' で学習したタスク \mathcal{T}' と \mathcal{D} で学習したタスク \mathcal{T} には類似性が必要
 - c. 出力の空間は違っていてもかまわないが、特徴量・入力は同じ必要がある
2. 以下のような方法で転移学習を行う
 - a. あるネットワークで \mathcal{D}' を使いタスク \mathcal{T}' の学習をし、その結果 \mathbf{f}' を利用する
 - i. このタスク \mathcal{T}' の学習を \mathcal{T} の事前学習と呼ぶ
 - ii. 逆に、タスク \mathcal{T} は $\mathcal{T}\mathcal{T}'$ を下流タスクと呼ぶ
 - iii. タスク \mathcal{T}' を十分学習できたものとする
 - b. 結果 \mathbf{f}' を下流タスクのための特徴抽出器として使うことができる
 - c. 結果 \mathbf{f}' の構造、一部重みをコピーして別のネットワークを作る方法もある
 - i. この方法はファインチューニングと呼ばれる
3. 難しい点もある
 - a. タスクの類似度を測るのは難しい
 - b. 事前に転移学習の成否を予測するのも難しい

11.3.2 特徴抽出器としての利用

1. \mathbf{f}' の中間層を1つ選び、入力に対する中間出力とする
 - a. 入力 \mathbf{x}_i に対して選んだ層の出力 $\mathbf{z}(\mathbf{x}_i)$ を中間出力とする
 - b. $\mathcal{D} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1, \dots, N}$ から $\tilde{\mathcal{D}} = \{(\mathbf{z}(\mathbf{x}_n), \mathbf{y}_n)\}_{n=1, \dots, N}$ をつくる
 - c. 新たに作ったデータがよいものであれば、学習アルゴリズムはシンプルでも性能がでると想定できる

d. 層の選び方、
e.
f.
g.
h.
i.
j.
k.
l.
m.
n.
o.
p.
q.
r.
s.
t.
u.
v.
w.
x.
y.
z.
aa.
ab.
ac.
ad.
ae.

af.
ag.
ah.
ai.
aj.
ak.
al.
am.
an.
ao.
ap.
aq.
ar.
as.
at.
au.
av.
aw.
ax.
ay.
az.
ba.
bb.
bc.
bd.
be.
bf.
bg.

bh.
bi.
bj.
bk.
bl.
bm.
bn.
bo.
bp.
bq.
br.
bs.
bt.
bu.
bv.
bw.
bx.
by.
bz.
ca.
cb.
cc.
cd.
ce.
cf.
cg.
ch.
ci.

- cj.
- ck.
- cl.
- cm.
- cn.
- co.
- cp.
- cq. 複数の
- cr. (\mathbf{x}_i, d_i) を入力として選んだ中間層の出力を利用して新たなデータ $(\mathbf{z}(\mathbf{x}_i), d_i)$ を作って学習する
- cs. シンプルな予測器でも、選択した中間層がよければ性能がでると想定できる
- ct. 選択する中間層、および複数の層の出力を組み合わせるバリエーションに自由度がある
 - i. 選び方は、経験的な検証に依存する

11.3.3 ファインチューニング

1. \mathbf{f}' の一部をコピーして、新たなネットワーク \mathbf{f} をつくる
 - a. \mathcal{T}' と \mathcal{T} が近ければネットワークの変化はごくわずかだと想定され、それが名前の由来となっている
 - b. 単純に出力層以外をコピーする方法などが典型
 - i. 出力層のみ、クラス数の増減などの影響から作り直す必要がある
 - ii. コピーした層に追加で中間層を追加、あるいは部分的にコピーから削減するなどをしてよい
 - c. コピー、加工して作ったネットワーク \mathbf{f} を \mathcal{D} を用いて学習する
 - i. 追加した層の重みのみランダムで、そのほかはコピー元の値で初期化する
 - ii. この状態の \mathbf{f} を事前学習済みであると呼ぶ
 - d. 学習するパラメータ数は調整できる
 - i. 訓練は一部の層の重みだけ更新することもできる

- ii. 少数データによる学習となるので、すべて更新する方法が裁量ではないと思われる
 - 1. 過剰適合のおそれがあるため
- iii. 典型的には、出力層に近い一部の層のみを訓練する傾向にある
 - 1. 入力に近い層よりも、出力に近い層のほうが具体的なタスク依存度が高いとされているため

11.3.4 ImageNet事前学習

- 1. 物体認識をするための大きなデータセットを用いた事前学習
 - a. 学習結果のネットワークを前節、前々節の方法で学習できる
 - b. 手持ちの訓練データが小さくても、学習に利用できる
 - i. 物体認識から、タスクとしては遠い単眼深度推定までそれが可能

11.3.5 事前学習の有効性

- 1. 事前学習に3つの有効性が示唆されている
 - a. 予測精度の向上
 - b. 学習の収束速度の向上
 - c. 推論時の頑健性の向上
- 2. 予測精度の向上について
 - a. 事前学習を通じて獲得した特徴抽出の能力が、下流タスクでも再利用されるという理解
 - b. 下流タスクのデータ量が十分な場合にはそれほどの効果がないことも報告されている(179より)
- 3. 学習の収束速度の向上
 - a. 少ない重みの更新で精度が向上すること
 - i. 下流タスクのデータセットが大きくても期待できる能力とされる(180、236より)
 - ii. 事前学習がランダムな初期値よりも統計的によい初期値に重みを重みに与えているためという理解
- 4. 推論時の頑健性の向上

- a. 以下について性能が向上することを指す(181、182より)
 - i. GAN
 - ii. 分布外入力 of 検出
 - iii. 確信度の校正精度
 - b. 事前学習なしだと最低限の特徴空間しか得られないからとされている
5. 残された課題
- a. 以下の説明されていない部分がある(180、326より)
 - i. 統計的な性質が異なる、特徴の再利用がほぼないタスク間でも性能が向上する
 - 1. 事前学習用のデータセットが大きいから平気...というような理解

11.4 半教師あり学習

11.4.1 概要

1. $\mathcal{D}_L = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1, \dots, N}$ と $\mathcal{D}_{UL} = \{(\mathbf{x}'_n)\}_{n=1, \dots, N'}$ をもっている
 - a. 前者についてはラベルがあるが少数で、後者についてはラベルがないが多い、といった状態
 - i. アノテーションコストが高いが、データ自体は結構ある...といった状態
 - b. \mathcal{D}_L は普通の教師あり学習で、それ以外はいろいろ損失関数などを準備して訓練に利用する流れ
 - i. 各手法の特徴はラベルなしデータの利用方法にある
 - c. 最小化したいのは2つの損失関数 $E'(\mathbf{w}; \mathcal{D}_{UL})$ と $E(\mathbf{w}; \mathcal{D}_L)$ の和
 - i. 後者は普通の教師あり学習として考える
 - ii. 前者の特徴づけの異なる3つの手法を紹介する
 1. 一貫性正則化
 2. 疑似ラベル
 - a. 教師データの話に言及していて損失関数そのものの取り扱いについてポイントがあるタイプの手法ではない気がする
 - b. kaggle でよくみかけた手法、として認識している

3. エントロピー最小化

11.4.2 一致性正則化

1. $E'(\mathbf{w}; \mathcal{D}_{UL})$ を以下のように特徴づける場合

$$\lambda \sum_{x_i \in \mathcal{D}_{UL}} \delta(y(x_i; w), y(x'_i; w))$$

2. 式の特徴は、以下の通り

- a. 各記号の特徴

- i. λ は教師あり学習部分に対してどれだけ別の損失関数を重要視するか？を表現している
- ii. $\delta(y(x_i; w), y(x'_i; w))$ 部分は、2つの値の差を表現するものとなっている
 1. y は入力に対する予測結果
 2. x と x' はデータ拡張などによって前者から後者を作成する
 3. 予測結果が変動によってあまり変わらないように学習を行う

- b. データの作成方法について

- i. 敵対的事例からくる仮想敵対的学習
 1. 予測結果を大きく変える「小さな変動」を逆算して作る
- ii. ネットワークを確率的に変動させたとき、同じ入力に対する予測が不変になることを要請する(δ で定義していた差の評価をより厳しくする...?)
- iii. 学習中のネットワークの重みの時間変化を変動の源とするような方法(temporal ensemble)
- iv. ネットワークのパラメータの時間方向の移動平均を求めて、それを正解ラベルの作成に利用する方法(mean teacher)

11.4.3 疑似ラベル

1. 2つのネットワーク S, T を利用した方法

- a. 教師ネットワーク T 、生徒ネットワーク S とする
- b. 2つのネットワークが同一の場合というのもあり、その場合をself-trainingと呼ぶ

2. 知識蒸留とつながりがある方法

- a. ラベルなしのデータを使ってラベル付きデータの学習結果を転移する知識蒸留といえる

3. $\mathcal{D}_{UL} = \{\mathbf{x}_n\}_{n=1}^N$ を以下のように用いる

- a. ネットワーク T に \mathbf{x}_n 入力してラベルの予測 y_n を得る
- b. ラベル付きということにしたデータ $\mathcal{D}'_{UL} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ を得る

4. $E'(\mathbf{w}; \mathcal{D}_{UL})$ を以下のように特徴づける

$$\lambda \sum_{x_i \in \mathcal{D}_{UL}} \delta(y(x_i; \mathbf{w}), y_i)$$

5. 式の特徴は、以下の通り

a. 各記号の特徴

- i. λ は教師あり学習部分に対してどれだけ別の損失関数を重要視するか？を表現していて何か特殊な理由等ない限りは1となる想定
- ii. $\delta(y(x_i; \mathbf{w}), y_i)$ 部分は、2つの値の差を表現するものとなっている
 - 1. y はネットワーク S への入力に対する予測結果
 - 2. y_i はネットワーク T による疑似ラベル
 - 3. 予測の差をとっている
- iii. 教師あり学習部分を S の学習に使わない場合というものもある(損失は上の式だけ)

6. 疑似ラベルの作り方について

- a. 決定的な方法はない
- b. 予測結果のクラス分布を疑似ラベルとする
- c. 予測結果のクラスラベルを疑似ラベルとする
- d. 予測クラス分布に従ってクラスをサンプリングする
- e. 疑似ラベル、というより自己訓練法の有効性が確認されている
- f. しかし、どの方法が良いのかははっきりしていない
- g. 転移学習を上回る場合があり、それはタスクのちがいというものが疑似ラベル法には存在しないからと理解されている

11.4.4 エントロピー最小化

1. $E'(\mathbf{w}; \mathcal{D}_{UL})$ を以下のように特徴づける場合

$$-\sum_{k=1}^K y_k \log(y_k)$$

2. ネットワークが出力するクラススコアのエン트로ピーが小さくするような方法
 - a. ラベル付きでないデータもなんらかのクラスには属するだろう、という期待による

11.5 自己教師あり学習

11.5.1 概要

1. 事前学習用のラベル付きデータがない場合の方法
 - a. これは、転移学習の前提を満たさないことを意味する
 - b. しかし、正解ラベルのない大量のデータ、少数のラベル付きデータならあるという想定
 - c. クラスタリングのように、なんらかのラベルをつかってそれを当てはめることができるデータをもっている状態
 - i. このなんらかのラベルをつかってそれを当てはめること、を事前学習とした転移学習のイメージ
 - ii. 具体的には以下のようなタスクがあり、このような単独では意味をなさないタスクをプレテキストタスクと呼ぶ
 1. データのうち一部単語をカットアウトして穴埋めするタスクなど
 2. 画像に幾何学変換をくわえて別の画像と見分けるタスクなど
 - d. 教師なし学習と同じデータの状態を想定
 - i. しかし疑似ラベルをつけて行うプレテキストタスクに対する教師あり学習ではある
 - ii. 手持ちのデータに対する学習まで含むと半教師あり学習といえる
 - iii. word2vecのようなものを想定していると思うが、目的とするタスクに正解ラベルが必要ない場合もありそれも含めて自己教師学習と呼ぶことが

ある

1. しかし、この設定の問題はここでは扱わない

11.5.2 プレテキストタスク

1. 自己教師学習はプレテキストタスクの教師あり学習
 - a. そのタスク T' はラベルを自動で付与できるようにでなければならない
 - b. 目的に対して有用な事前学習のタスクを選ばなければならない
 - c. 具体的に、自然言語分野では以下のようなものがある
 - i. 完全な文から一部を取り除きその一部を文脈から予測させるタスク
 - ii. 2つの文を入力として、その文が実際に連続しているかを予測させるタスク
 - d. 音声や画像でも同じようなことは可能
 - i. 穴埋め、色情報の予測等、ラベル付与が簡単な内装が基本
 - ii. 画像では距離計量学習に近い話題もある

11.5.3 対照表現学習

1. 画像分類のためのプレテキストタスクにはいくつかの方法が知られている
 - a. 画像の一部をマスクして周囲の画像からその部分を穴埋めするタスク (inpainting)
 - b. カラー画像から色情報を除去してグレースケール画像を作り元の色情報を予測(colorization)
 - c. 画像を矩形複数に分割し、順序をランダムに入れ替えたものを入力としてその正しい順序を予測させる(jigsaw puzzle)
 - d. 画像分類に対して上の3つの方法はそこまで効果がないが、データ拡張で得た画像と元画像の同一性を予測するタスクなどは効果がある
2. contrastive representation learningの概要は以下のとおり
 - a. 画像分類のプレテキストタスクに紐づいた話題
 - b. 対照損失というものをを用いる方法
 - c. 特徴表現空間を得ることが期待される