Deep Learning 輪読会 2017

第15章　表現学習

松尾研究室

リサーチエンジニア 曽根岡 侑也

構成

15章 表現学習

15.1 層ごとの貪欲教師なし事前学習

15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

15.2 転移学習とドメイン適応

15.3 半教師あり学習による原因因子のひもとき

15.4 分散表現

15.5 深さがもたらす指数関数的な増大

15.6 潜在的な原因発見のための手がかり

はじめに

- 本章の内容：表現を学習するとは何か
  - 「表現を学習する」とは
  - 「よい表現」とは
  - 表現の共有
    - マルチドメイン、マルチタスク、マルチモーダルで有用
    - 転移学習、ドメイン適応
  - 分散表現の優位性

表現学習

- **表現次第で情報処理タスクの難易度は変わる**
  - 例1）アラビア数字の計算（「CCX割るVI」 vs 「210 割る 6」 ）
  - 例2）数値リストのソート（連結リスト vs 木）
- **良い表現とは何か**
  - それに続くタスクを簡単にするもの、タスク次第
  - （例）NNの分類の場合、Softmaxに渡す表現は線形分離可能であるべき
  - 入力の情報をできるだけ保存すること vs 好ましい特性の獲得
- **教師なし学習・半教師あり学習に役立つ**
  - ラベルなしデータでよい表現を学習し、教師あり学習で使用

15.1 層ごとの貪欲教師なし事前学習

- **層ごとの貪欲教師なし事前学習**
  - 畳み込みなどの特殊な構造を使わず深層教師あり学習を可能にした

    初めての方法で**深層学習を再燃させるきっかけ**
  - あるタスク（教師無し）で学習した表現が別タスクで有用と示した実例
  - 層ごとに教師なし学習し全部の層の初期化を行ない、その後再学習させる
  - （例）深層自己符号化器、深層ボルツマンマシン

15.1 層ごとの貪欲教師なし事前学習

- **層ごとの貪欲教師なし事前学習のアルゴリズム**

Algorithm 15.1 層ごとの貪欲教師なし事前学習手続き.

教師なし特徴量学習アルゴリズムを $\mathcal{L}$ とする．$\mathcal{L}$ は訓練事例の集合を入力とし，符号化器または特徴量関数 f を返す．入力生データを $\boldsymbol{X}$ とする．$\boldsymbol{X}$ は行ごとに 1 つの事例を持つ．$f^{(1)}(\boldsymbol{X})$ は $\boldsymbol{X}$ に対する 1 段目の符号化器の出力を示す．再学習を行うには，学習器 $\mathcal{T}$ を用いる．$\mathcal{T}$ は初期関数 f と入力事例 $\boldsymbol{X}$（および対応する目標 $\boldsymbol{Y}$，ただし教師あり再学習を行う場合）を入力とし，再学習された関数を返す．ステージ数を m とする．

$f \leftarrow$ Identity function
$\tilde{\boldsymbol{X}} = \boldsymbol{X}$
for $k = 1, \ldots, m$ **do**
  $f^{(k)} = \mathcal{L}(\tilde{\boldsymbol{X}})$
  $f \leftarrow f^{(k)} \circ f$
  $\tilde{\boldsymbol{X}} \leftarrow f^{(k)}(\tilde{\boldsymbol{X}})$
end for
if *fine-tuning* ***then***
  $f \leftarrow \mathcal{T}(f, \boldsymbol{X}, \boldsymbol{Y})$
end if
Return f

15.1 層ごとの貪欲教師なし事前学習：事前学習

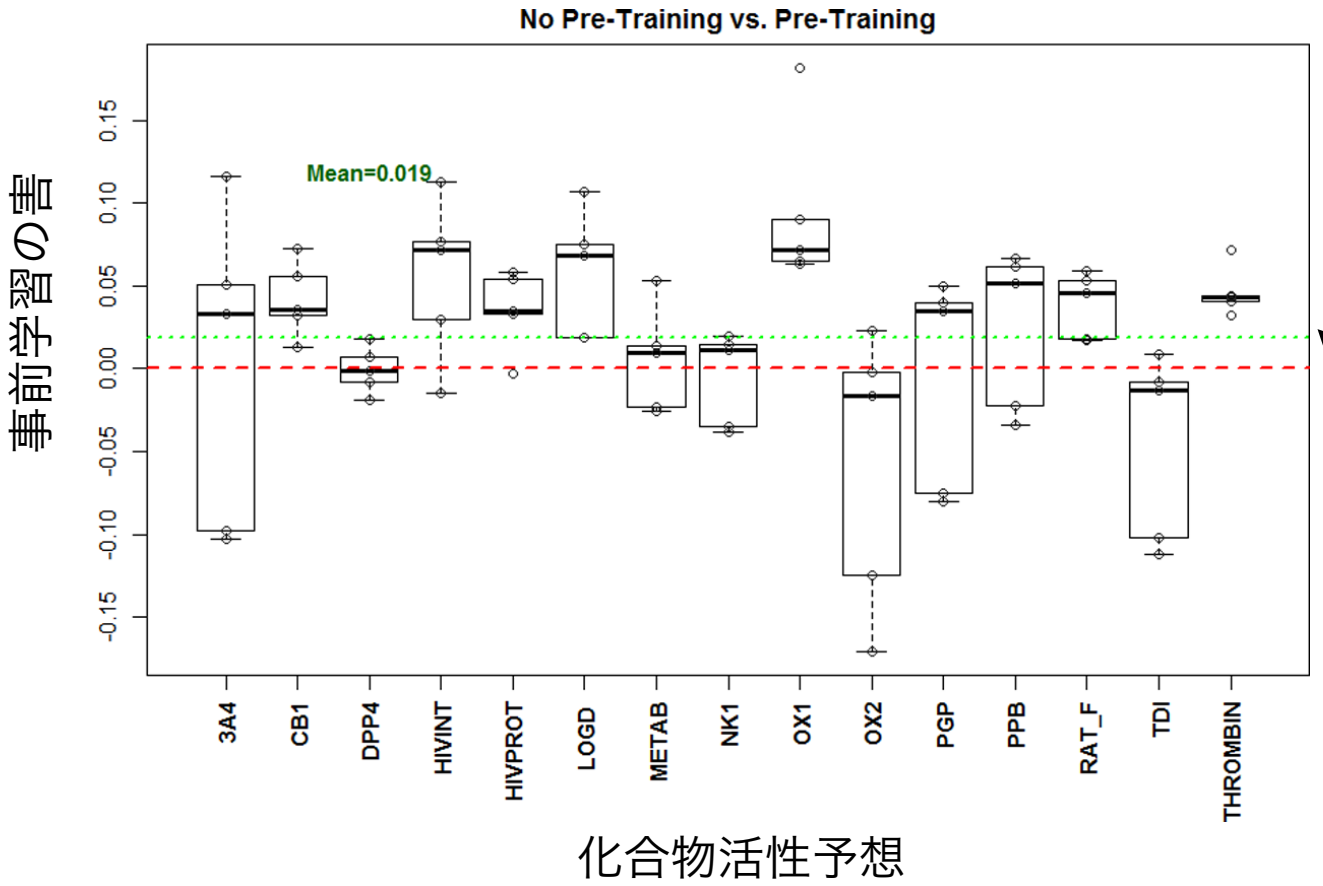
- **事前学習（Pretraining）**
  - 一度他のタスクで学習し, その後 **再学習（fine-tune）**すること
  - 事前学習と再学習の二段階を刺すことが多い
- **事前学習の主な使い方**
  - 特徴抽出器：抽出した表現の上に分類機をつけ、分類機だけを再学習
  - パラメータの初期化：モデル全体を再学習させる

    （参考：ChainThaw：層ごとに再学習）

15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- 多くの分類タスクでは事前学習によりテスト誤差を改善する

  一方、平均的には「害がある」or「利益がない」



事前学習しないときの
平均値

15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- **教師なし事前学習を行う動機・アイデア**

  1. NNの初期パラメータ選択が正則化の効果があるという仮説

     - 事前学習なしでは到達できない領域にモデルを初期化している

  2. 入力分布の学習が出力への写像の学習を手助けしうる

     - 教師なしタスクで有用な特徴量は教師ありでも有用

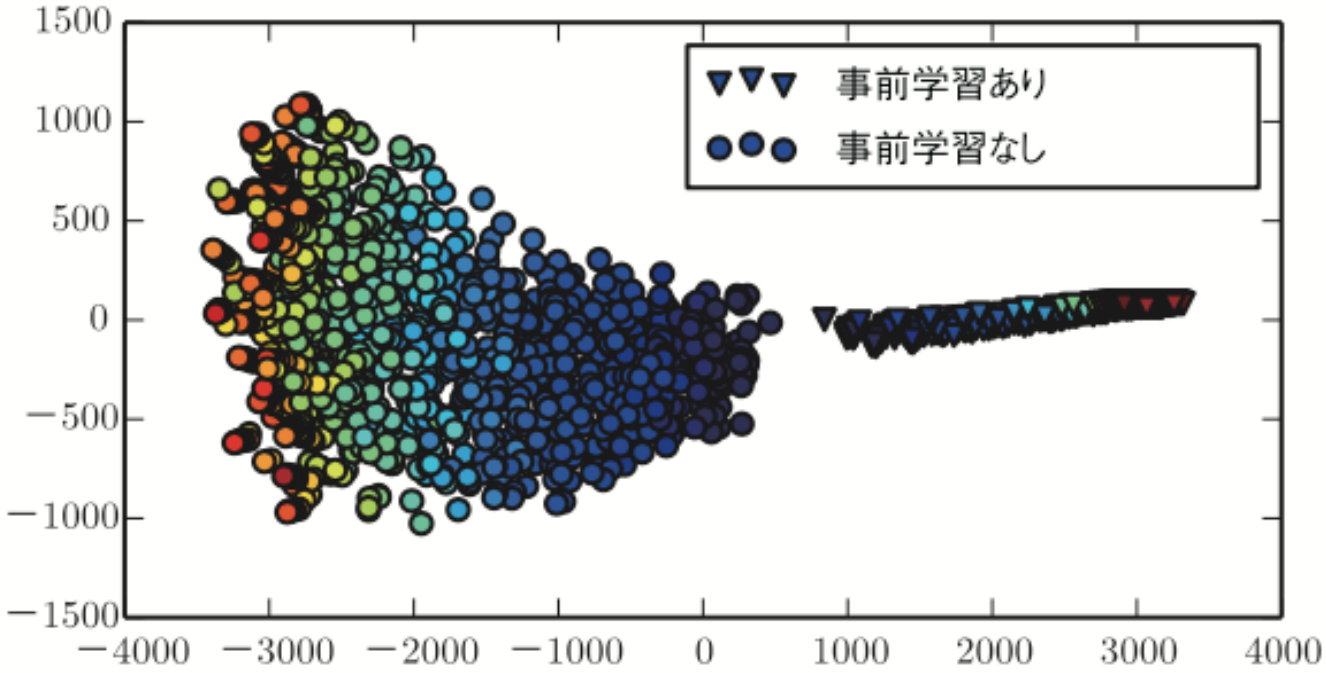
15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- **表現学習の観点で有効なとき**
  - 自然言語のような**初期表現が不十分なものに有効**
  - 画像のような既にリッチなベクトル空間ではそこまで有用ではない
- **正則化の観点で有効なとき**
  - **ラベルなしデータは大量にあるが、ラベルありデータが少ないとき**
- **その他の有効なとき**
  - 学習される関数が極めて複雑な場合
    - 通常の正則化が単純な関数を学習させようとしているのに対して、

      学習タスクに有用な特徴量を発見させる方向に働いている

15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- **教師なし事前学習は正則化だけでなく訓練誤差も改良する**
- 到達不可能な領域に事前学習が導いていることを示す実験 [Erhan, (2010)]
  - 事前学習をすると一貫した特定の関数へ収束する
  - **事前学習により、推定過程の分散を低下 + 過適合のリスクを低下**



15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- **事前学習が最もよく働くのはいつか** [Erhan 2010]
  - より深いネットワークではテスト誤差の平均と分散を低下させた

    ※ この実験は、ReLU・ドロップアウト・バッチ正規化以前の話
- **教師なし事前学習がどのように正規化として機能するのか**
  - 観測データを生成する潜在因子に関する特徴量の発見の手助けをしている

15.1.1 教師なし事前学習はいつ、なぜうまく働くのか

- **教師なし事前学習のデメリット**
  - 正則化の強度を調整するわかりやすい方法がない
  - 事前学習時のハイパーパラメータ調整が難しい
    - 2段階でのステップをふむため、事前学習のハイパーパラメータを再学習時の結果をみて変更する必要がある
- **現在のユースケース**
  - 自然言語処理の単語埋め込み以外ほとんど使わない

    （one-hot, 大量のラベルなしデータがあるという理由）
  - **教師あり事前学習**はよく使われており主流のものは公開されている

15.2 転移学習とドメイン適応

- ある設定（分布P1）で学んだことを別の設定（分布P2）の汎化能力の向上を試みるアイデアがある
  - 同一の表現が双方の問題設定で有用というアイデア
- **転移学習**
  - 異なる2つ以上のタスクで学習
  - P1とP2の変化を説明する因子が近いという仮定
  - あるタスクで学習したあとに近いタスクで再学習する
  - 例1）犬猫分類　→　アリハチ分類（エッジなどの下位概念を共有）
  - 例2）音声認識（人毎に変更、上位層の言語モデルを共有）

15.2 転移学習とドメイン適応

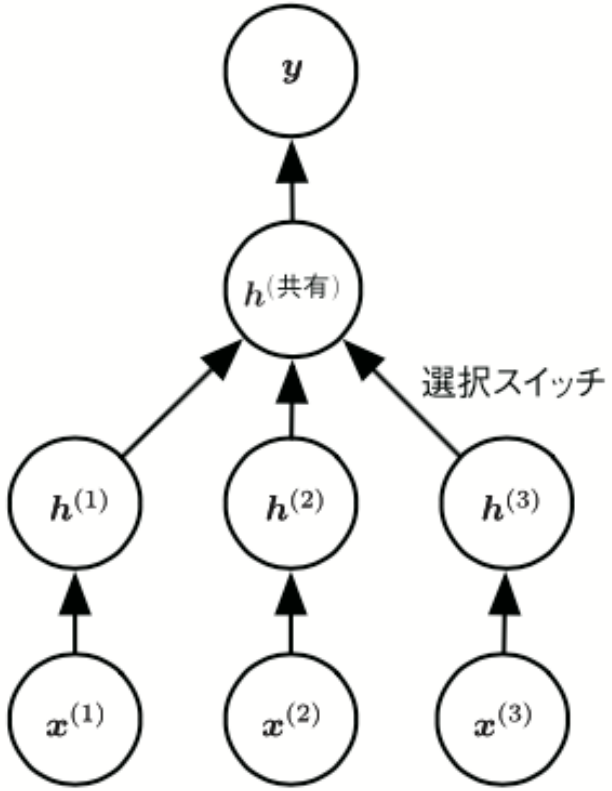
- **ドメイン適応**
  - タスクは同じであるが、入力分布が異なる問題で学習
  - 例）音楽レビューの感情分析　→　家電レビューの感情分析

    （語彙や文体が違うため、再学習が必要）
- **コンセプトドリフト**
  - 時間経過でデータ分布が変化することを考慮した転移学習

15.2 転移学習とドメイン適応

- **マルチタスク学習**
  - 復数のタスクを同時に学習し共通の表現を一部で利用する
- **マルチタスク学習や転移学習の構造例（右図）**
  - 出力変数yはすべてのタスクで共通
  - 3つのタスクに対する入力を

    共通の特徴量に変換するよう学習



15.2 転移学習とドメイン適応

- **ゼロショット学習：ラベルあり事例はなし**
  - 入力x, 出力y, タスクを記述する確率変数Tでモデル化 $p(\boldsymbol{y} \mid \boldsymbol{x}, T)$
  - 例）大量のテキストで学習し、物体認識する

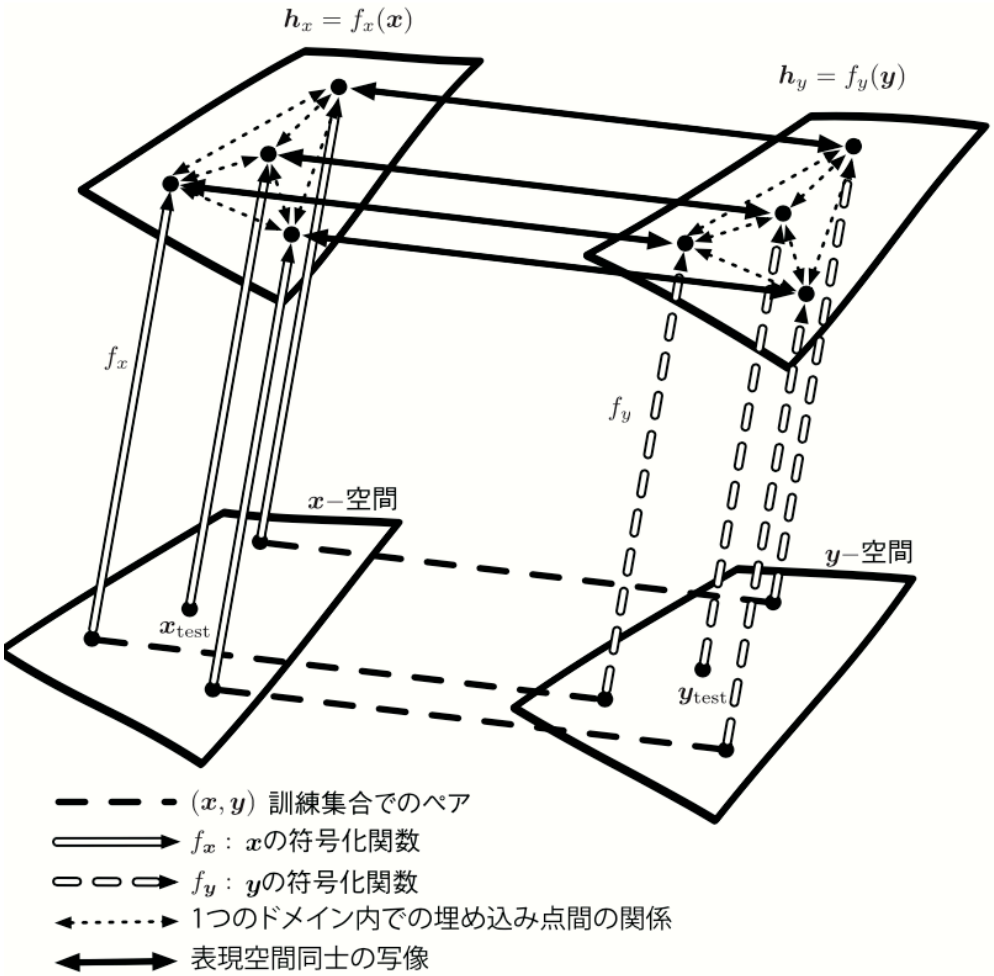
    猫の特徴のテキストから画像に猫がいるかを判定
  - 例2） 機械翻訳

    単一言語コーパスからそれぞれの言語の分散表現を学習し結びつける
- **ワンショット学習：ラベルありが1つだけ与えられる**
  - ラベル無しで潜在的なクラスを分けるような表現を学習
  - 1つのラベルありデータがあれば特徴空間上で周りにあるデータの

    ラベルを推論できる

15.2 転移学習とドメイン適応

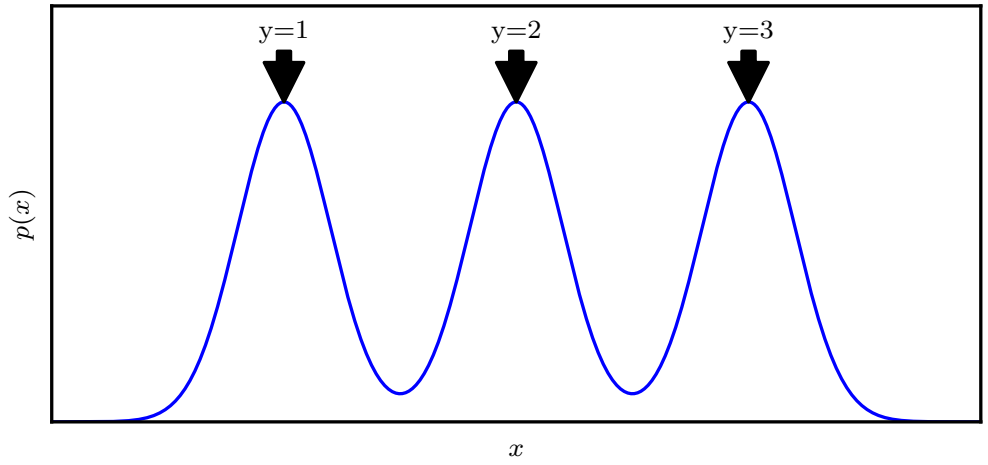
- **マルチモーダル学習**
  - あるモダリティの表現と他のモダリティの表現、事例のペアから表現の関係を捉える



15.3 半教師あり学習による原因因子のひもとき

- **良い表現とは何か**
  - データの潜在的原因に対応しており、

    各特徴量が異なる潜在的な原因に対応しているため、互いに解きほぐせる
  - モデリングしやすい表現であることが多い
- **半教師あり学習**
  - 観測データからxの表現として潜在的原因を学習しyに結びつける



半教師あり学習が成功する例：混合分布

15.3 半教師あり学習による原因因子のひもとき

- 観測 x の原因となる潜在因子hを見つけることができ、

  ラベルyがその潜在因子と結びつきであれば、半教師あり学習が有効

$$p(\mathbf{h}, \mathbf{x}) = p(\mathbf{x} \mid \mathbf{h}) p(\mathbf{h}). \qquad p(\boldsymbol{x}) = \mathbb{E}_{\mathbf{h}} p(\boldsymbol{x} \mid \boldsymbol{h}).$$

  - p(y l x)は周辺分布p(x)と密接に結びついているため、

    p(x)の知識は手助けになるはずである

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{y}) p(\mathbf{y})}{p(\mathbf{x})}.$$

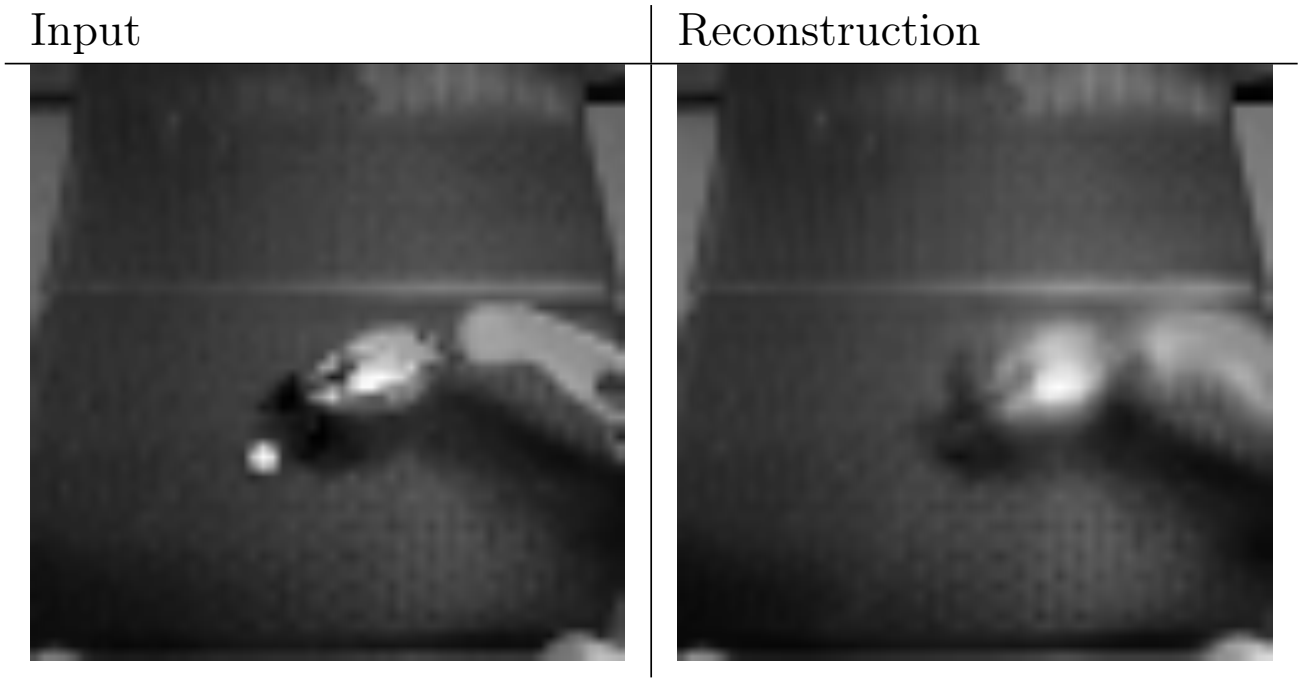
- **課題：**
  - **観測結果の大半が多くの潜在的原因によって生成されている**
  - **全ての潜在的原因を捉えて、紐解くことはほぼ不可能**
- **重要な研究テーマ**
  - 「**それぞれの状況で何を符号化するべきか**」を決定すること

    1. 最も適切な変化の要因群を捉えるように教師あり + 教師なしを利用
    2. 教師なしのみであれば大きな表現を使う

15.3 半教師あり学習による原因因子のひもとき

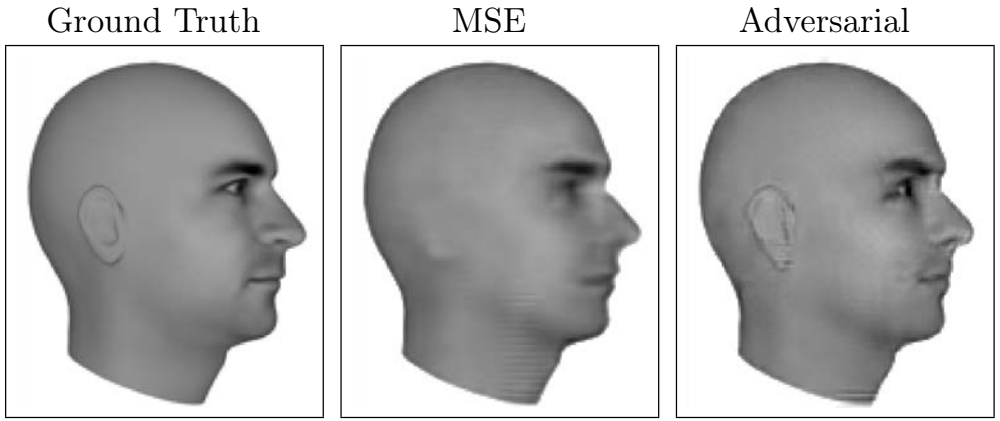
- **どの潜在的原因を顕著とするか**
  - 最適化する誤差の設定 ≒ どの原因が顕著と考えるかの設定
  - 画像の場合、平均二乗誤差だと小さいものが消える

    ⇒ 卓球の玉を動かすロボティクスタスクだと問題に



15.3 半教師あり学習による原因因子のひもとき

- **GANの登場**
  - 分類器を騙すように画像を生成モデルを学習する方法
  - **顕著とするものを決める方法すら学習する**
    - 分類機が識別できるものは全て顕著とされる



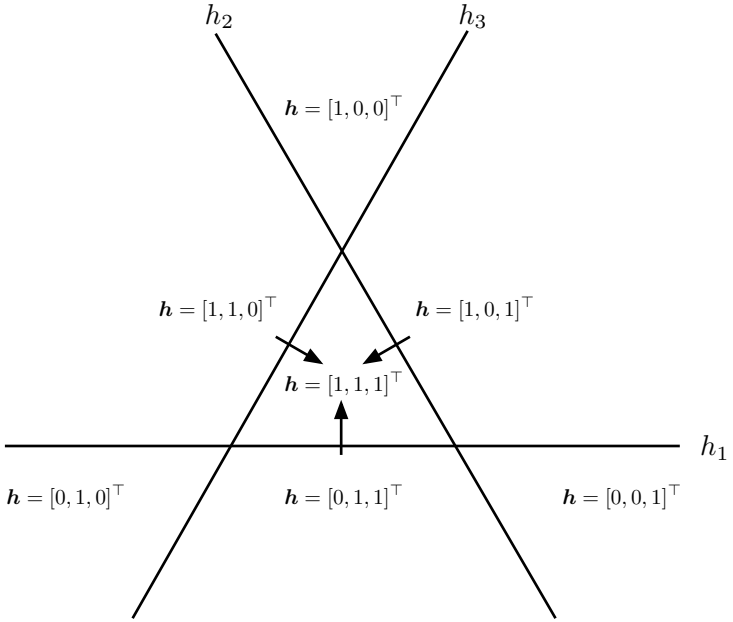
- **潜在的原因因子の学習のメリット**
  - 因果関係は普遍的なため、時間やドメインなどに対してロバストになる

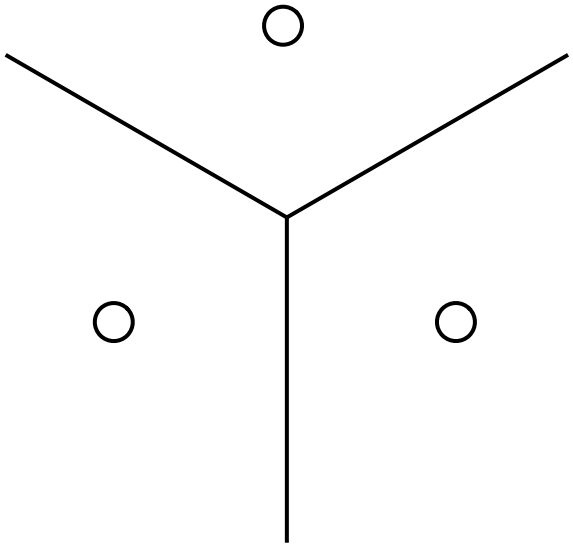
15.4 分散表現

- **分散表現：互いに分離可能な要素により構成された表現**
  - k個の値を持つn個の特徴量を利用することで k^n の概念を表現できる
  - データを説明する潜在因子を隠れ層が学習できるという仮説に基づく

15.4 分散表現

- **分散表現：n次元のバイナリ素性ベクトル（左）**
  - 2^n の状態を取ることができる、一般的に n^d を表すことができる
- **シンボリック表現（onehot表現）：各要素が各シンボルを表す（右）**
  - n個の異なる状態のみが起こる相互排他的な表現





15.4 分散表現

- 非分散表現に基づく学習アルゴリズム例
  - **k平均法を含むクラスタリング手法**：1つのクラスタに割り当てられる
  - **k近傍法アルゴリズム：**

    少数のテンプレートやプロトタイプ事例が関連付けられる
  - **決定木**：1つの葉ノードのみが活性化する
  - **混合ガウスや混合エキスパート**：
  - **ガウシアンカーネルを用いたカーネルマシン：**

    復数の値により表現されるが独立に制御するのは容易ではない
  - **n-gramによる言語/翻訳モデル：**

    文脈(シンボルの連なり)の集合は接尾辞の木構造に応じて仕切られる

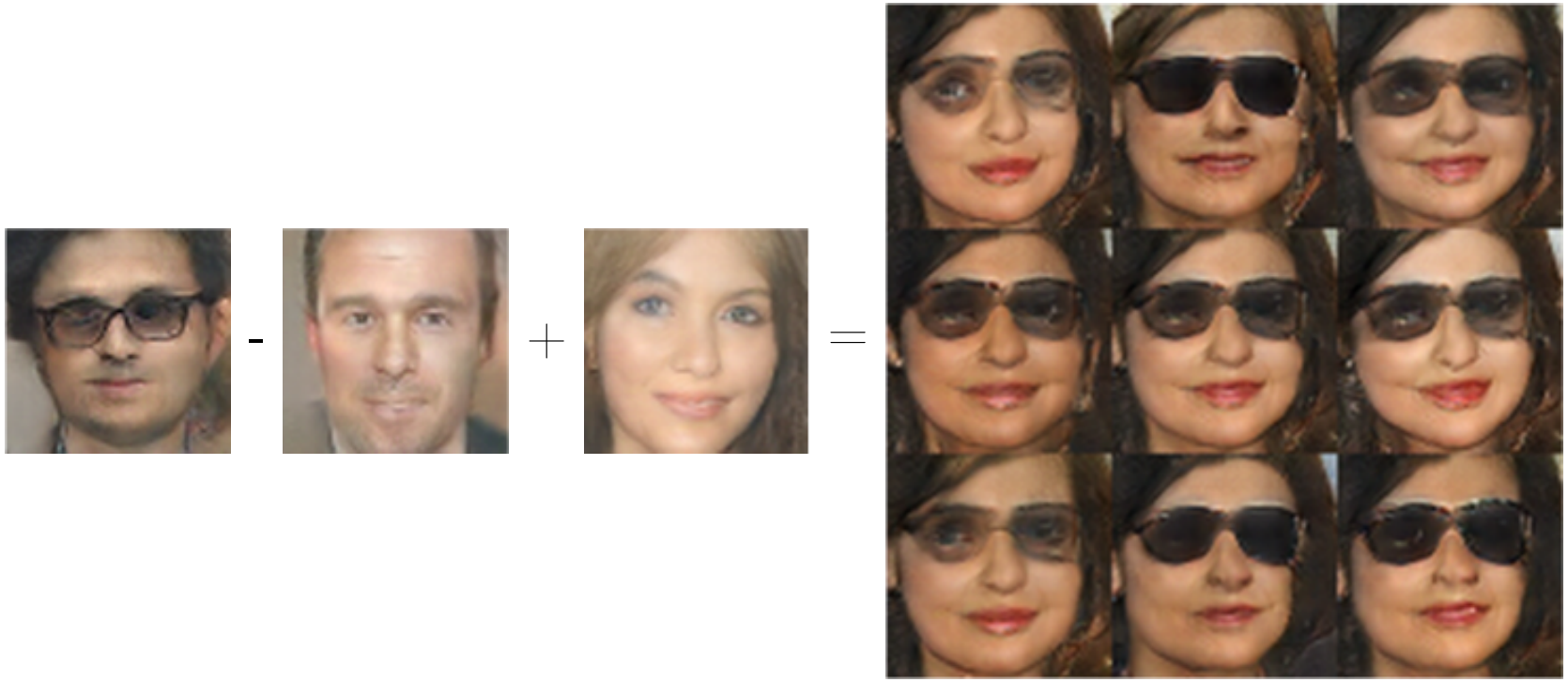
15.4 分散表現

- 分散表現によるメリット
  - **異なる概念感で共通の属性による汎化**（ex：猫と犬）
  - **類似度の測定が可能**
  - 複雑な構造が少数のパラメータによって表現できる時が有効
- 統計的な観点による比較：$O(n^d)$ の領域を分解する際
  - 非分散学習：識別可能な領域数と同じ事例が最低でも必要　　$O(n^d)$
  - 分散表現：$O(nd)$　のパラメータ数で分割可能

    ⇒ 少ないパラメータのモデルで良いため汎化に必要なデータは少ない

15.4 分散表現

- **GANを使った分散表現**
  - 性別やサングラスなどの表現を分離して獲得できている



15.5 深さがもたらす指数関数的な増大

- **原因因子はかなり高次なため、深層分散表現が必要**
- **莫大なユニット数があれば隠れ層を持つネットワークは万能近似器**
  - 多種類の関数を近似
  - 不十分な深さの場合、ユニット数は入力規模に対して指数関数的になる
  - **基本的に十分な深さがあると指数関数的なメリットがある**
- **例）積和ネットワーク（SPN）**
  - 確率変数の集合に対する確率分布を計算する ために多項式回路を利用
  - 有限の深さで確率分布を表せるが、表現力が制限される可能性がある
  - 深くなるほど、指数関数的な表現力を持つ

15.6 潜在的な原因発見のための手がかり

- **まとめ**
  - よい表現は、原因因子をひもとく表現である
  - 教師あり学習は、直接的に変動の因子を少なくとも1つ紐解いている
  - 表現学習は特設的ではない手がかりを利用している
- **よりよい表現を獲得するために正則化戦略が必要**
  - 人間が解けるタスクに類似した様々なAIタスクに適用可能な

    一般的な正則化戦略を見つけるのが深層学習の目的の1つ

15.6 潜在的な原因発見のための手がかり

- **滑らかさ：**僅かな変動εに対して　$f(\boldsymbol{x}+\epsilon \boldsymbol{d}) \approx f(\boldsymbol{x})$が成立するという仮定.

  これにより、入力の近傍店を汎化できる

- **線形性**：幾つかの変数の間の関係が線形であるという仮定.

- **複数の説明因子**：データが複数の潜在的な説明因子により生成されており,

  因子の各状態がわかれば大半のタスクは解けるという仮定.

- **原因因子**：モデルは変動因子hを観測データ x の原因として扱うように構築し

  その逆は成り立たない

- **深さあるいは説明因子の階層性** ：

  上位の抽象的な概念は階層構造をなす単純な概念により定義できるという仮定

15.6 潜在的な原因発見のための手がかり

- **タスク間の共有因子：**

  タスク間で共有の原因を持ちそうな時、因子hを共通に使える部分がある

- **多様体：**確率の大部分は集中しており、集中領域は局所的につながっている

- **自然なクラスタリング：**入力空間において結合された多様体は

  それぞれ 1 つのクラスに割り当てられると仮定.

- **時間的・空間的なコヒーレンス：**

  1. 最も重要な説明因子は時間をかけてゆっくり変化

  2. 真の潜在的な説明因子を予測する方が生の観測値を予測するよりも簡単であることを仮定

- **スパース性：**大半の特徴量は，大半の入力を説明するのに関連していない

- **因子の依存関係の簡潔さ：**

  よい高次の表現では因子は互いに単純な依存関係で結び付いている

参考文献

- Deep Learning
  - Ian Goodfellow, Yoshua Bengio, Aaron Courville
  - 日本語版
    - https://www.amazon.co.jp/%E6%B7%B1%E5%B1%A4%E5%AD%A6%E7%BF%92-Ian-Goodfellow/dp/4048930621