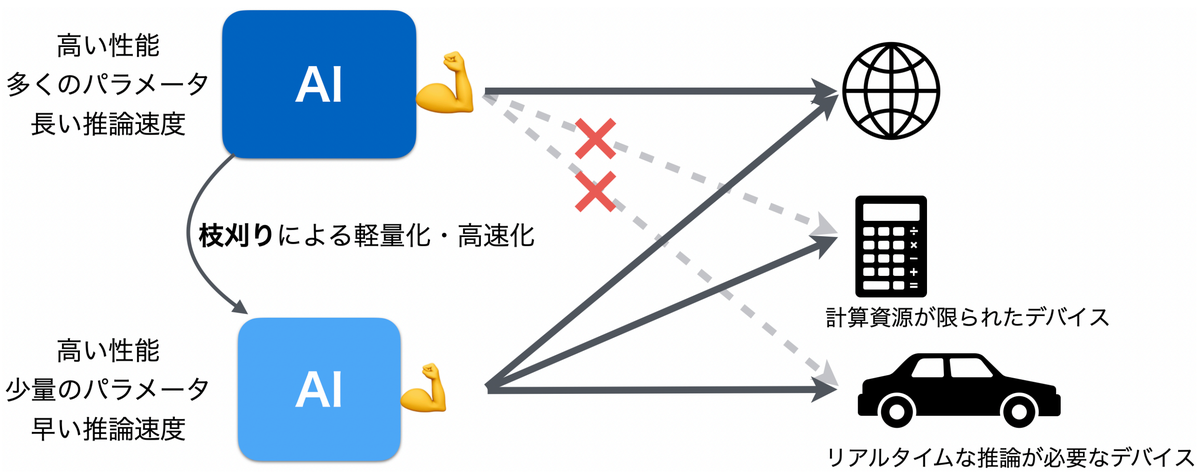
パネル 3：枝刈り

小林亮太，西川実希

中部大学　MPRG：機械知覚 & ロボティクスグループ

枝刈りの目的・社会的貢献

- 現状の課題：
  AI モデルの高性能化により，モデルのパラメータ数が増加
- 高性能モデルは：
  大量のメモリと電力を消費 ⚡
  推論速度が遅い 🕐
- 結果として，計算資源の限られたデバイスやリアルタイムな推論が求められる環境では活用が困難
- 枝刈りによるモデル圧縮
  1. 冗長なパラメータを削除し，モデルを圧縮
  2. 高い性能を維持しつつ，モデルサイズ削減・推論時間を短縮



枝刈りによって小型・低消費電力，リアルタイムが必要なデバイスへの AI 導入が可能に

枝刈りとは

▪ 非構造枝刈り（Unstructured Pruning）

- パラメータを要素単位で削除する手法

メリット

- 高い圧縮率を実現可能
- 元の性能を維持しやすい

デメリット

- 専用のハードウェアでないと推論速度の向上が限定的

▪ 構造化枝刈り（Structured Pruning）

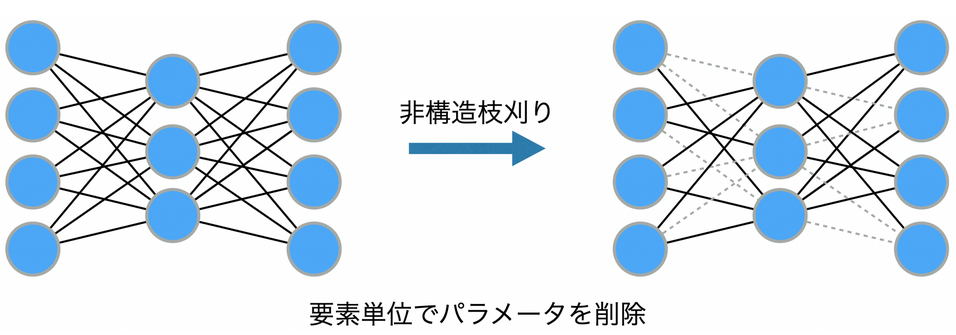
- パラメータを構造単位で削除する方ほ

メリット

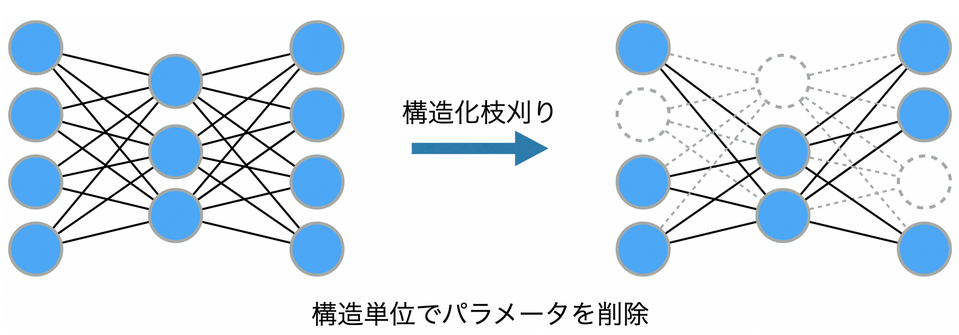
- 標準的なハードウェアで高速推論

デメリット

- 圧縮率に制限がある
- 性能劣化が大きい可能性



要素単位でパラメータを削除



構造単位でパラメータを削除

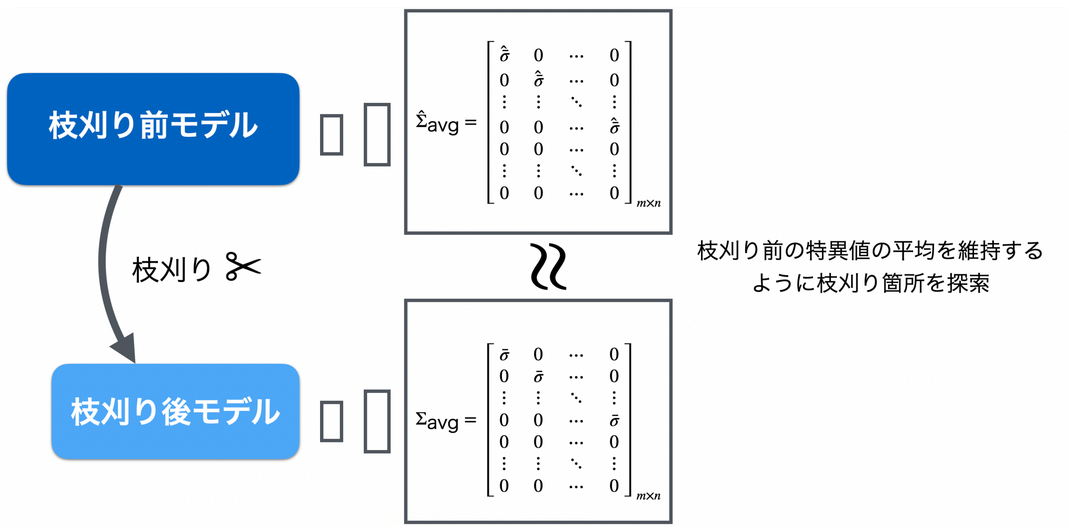
MPRG の枝刈り手法

▪ 事前学習で得た知識を維持する非構造枝刈り

- 事前学習で得た知識をどのように評価するのか
  特異値分解（SVD）を用いて得られる特異値の平均値を事前学習で得た知識として評価
- SVD の定義: 任意の $m \times n$ 行列 $\boldsymbol{A}$ を以下のように分解

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T = \sigma_1 \cdot \circ + \cdots + \sigma_r \cdot \circ$$

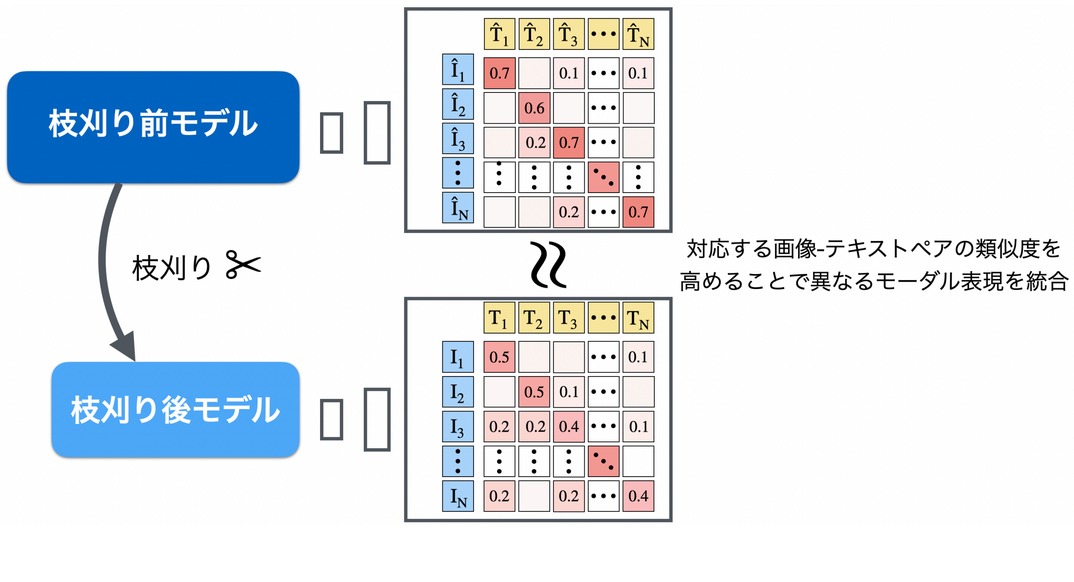
  ここで，$\boldsymbol{U}$ $(m \times m)$ と $\boldsymbol{V}$ $(n \times n)$ は直交行列 $\boldsymbol{\Sigma}$ $(m \times n)$ が特異値 $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$ を対角成分に持つ広義対角行列
- **特異値の意義**: σ_i は対応する特異ベクトルが捉える「方向」の重要度やエネルギーの大きさを示す



枝刈り前の特異値の平均を維持するように枝刈り箇所を探索

▪ VLM におけるモーダル間の知識蒸留による構造化枝刈り

- Vision-Language Model（VLM）
  画像とテキストを入力とする AI モデル
- 知識蒸留
  大規模で性能の良いモデルの知識を小規模なモデルへ転移させるアプローチ
- 手法の概要
  画像と言語を統合させる役割を持つ表現を，知識蒸留により維持しつつ枝刈り箇所を探索



対応する画像-テキストペアの類似度を高めることで異なるモーダル表現を統合