第2回ディスカッション

論文調査

ER20038 小林亮太

担当：岩垣，張，岡本

MPRG

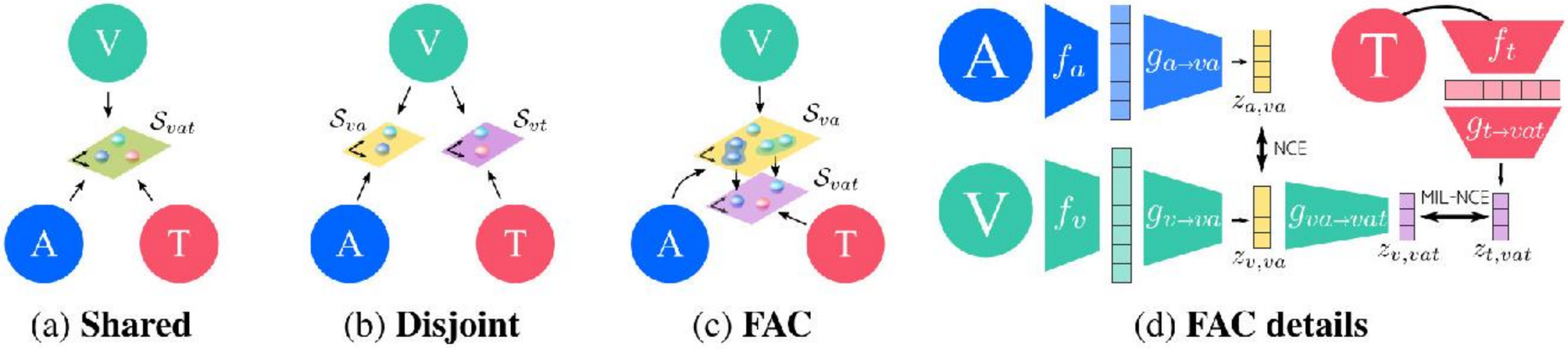
MACHINE PERCEPTION AND ROBOTICS GROUP

はじめに

- MultiModal Versatile (MMV)
  - Shared spaces
  - Disjoint spaces
  - Fine and coarse spaces
- Multimodal Contrastive Loss

MultiModal Versatile (MMV) [J.Alayrac, NeurIPS2020]

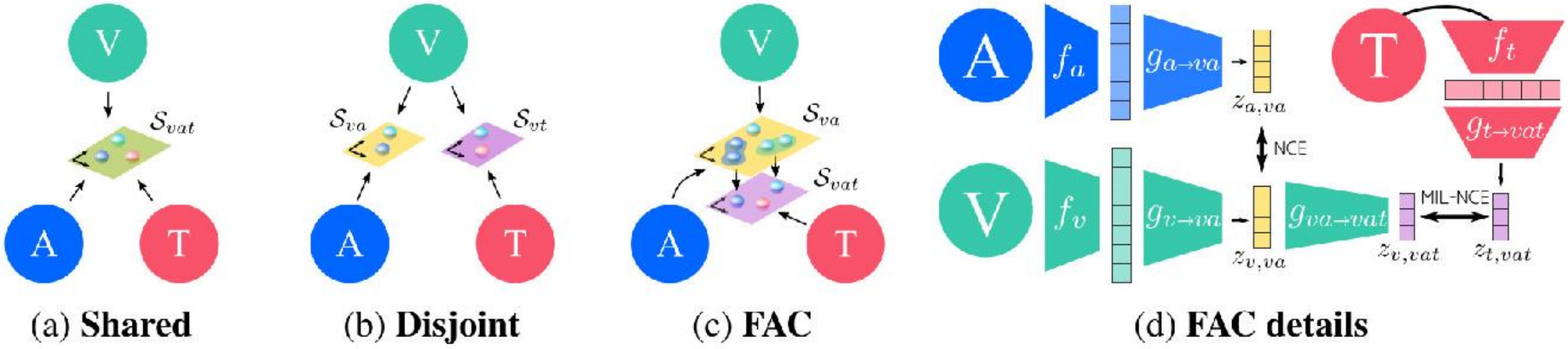
- 4つの特性を備えるように設計
  i. 3つのモダリティのいずれかを入力として受け取り可能
  ii. モダリティの特異性を重視
    - 音声とビジュアルのモダリティが言語よりも細かいことを考慮
  iii. 学習中に一緒に出現していない異なるモダリティでも簡単に比較可能
  iv. 動的なビデオまたは静的な画像としての視覚データに効率的に適用可能



(a) **Shared** (b) **Disjoint** (c) **FAC** (d) **FAC details**

MultiModal Versatile (MMV) [J.Alayrac, 2020]

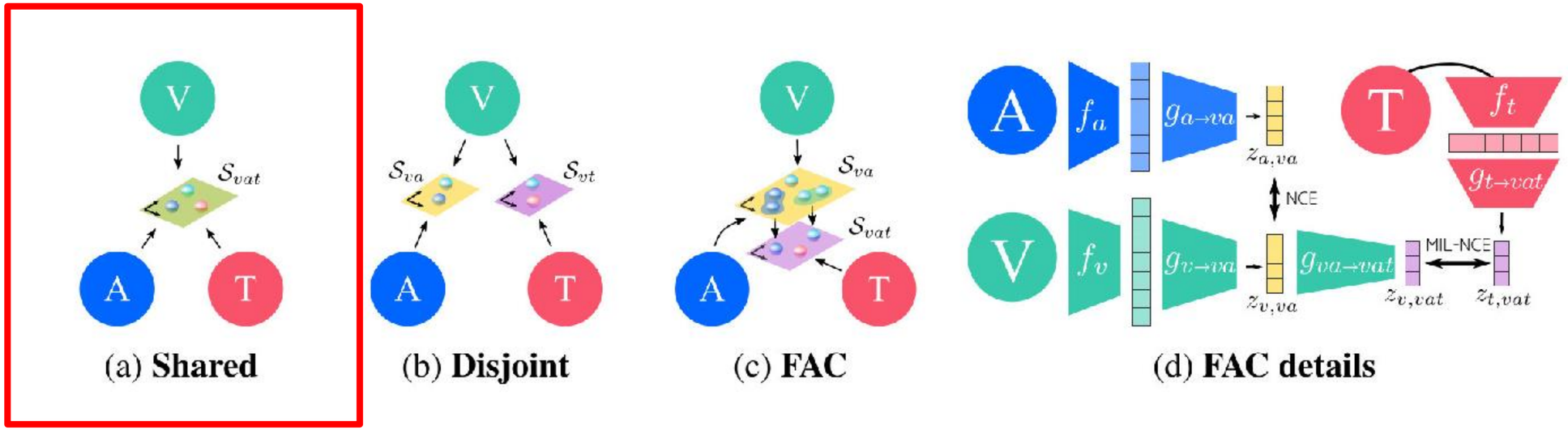
- 3つのモダリティを使用
  - テキスト，オーディオ，ビジュアル
  - テキストはオーディオに自動音声認識を用いて作成
- モダリティの埋め込みに3つの選択肢を検討
  - Shared spaces
  - Disjoint spaces
  - Fine and coarse spaces



(a) **Shared** (b) **Disjoint** (c) **FAC** (d) **FAC details**

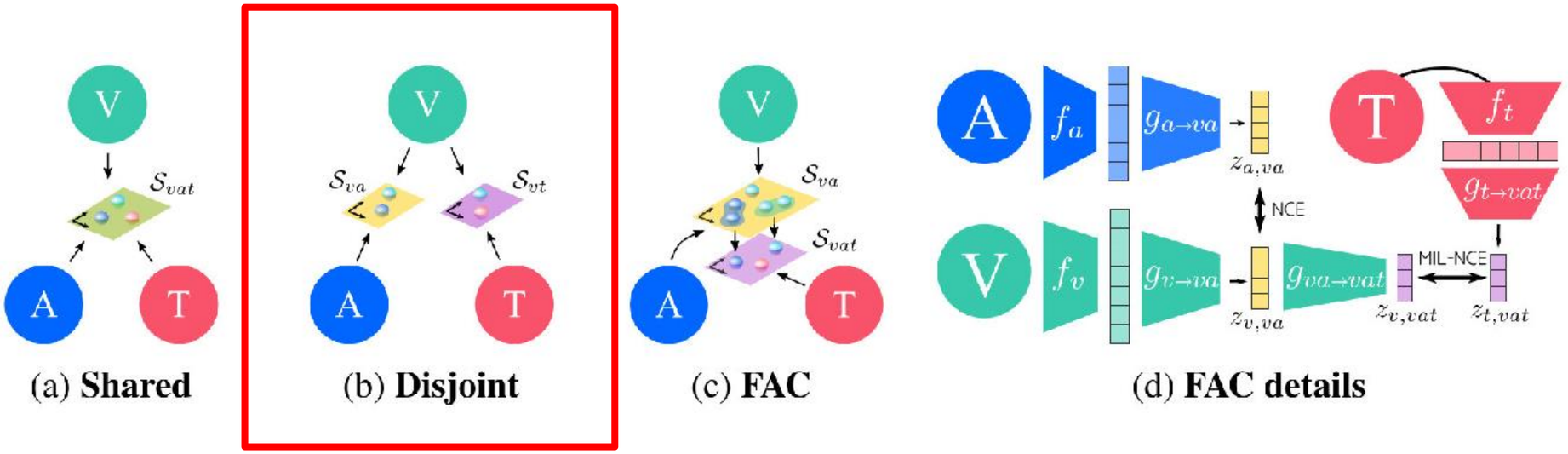
Shared spaces

- 全てモダリティが単一の空間S_{vat}を共有
  - 全てのモダリティ間で直接比較可
- 全てのモダリティが均一の粒度を持つことを暗黙のうちに仮定
  - 粒度：入力データの分割の程度
    - モダリティごとに異なりオーディオ，ビジュアルと比較してテキストの粒度は低い傾向
  - モダリティの特異性を無視
  - 特性 ii の欠如



Disjoint spaces

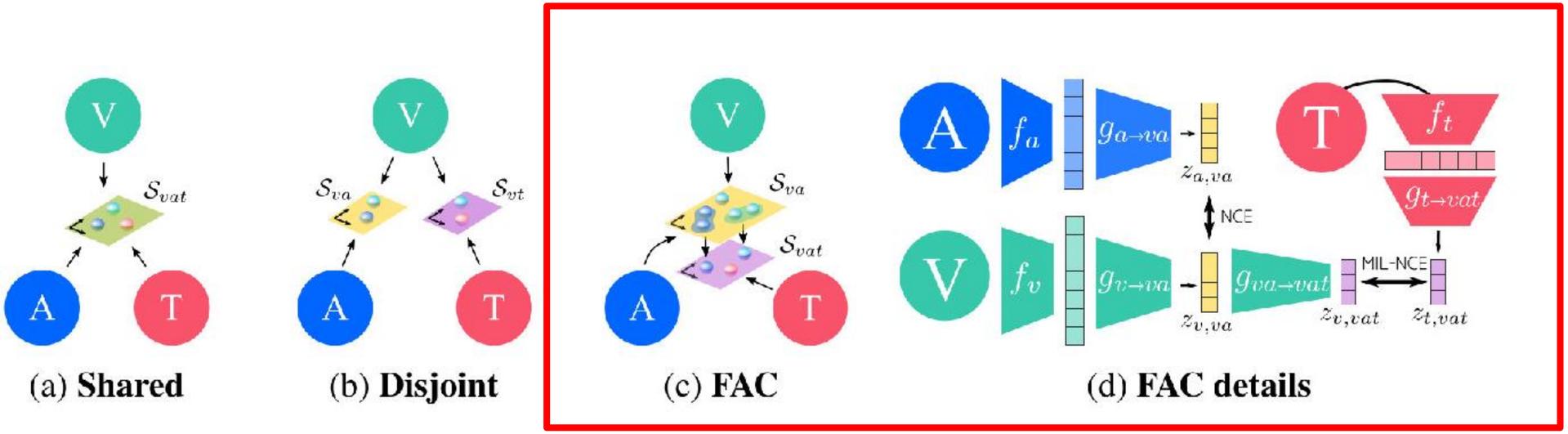
- ビジュアル-オーディオ，ビジュアル-テキストの空間S_{va}，S_{vt}が別々に存在
  - Shared spcaesよりもモダリティの特異性を重視
- 空間間の探査が不可能
  - テキストからオーディオへの検索が不可能
  - 特性iiiの欠如



(a) **Shared** (b) **Disjoint** (c) **FAC** (d) **FAC details**

Fine and coarse spaces (FAC)

- 2つの空間S_{va}, S_{vat}を作成
  - S_{va}：オーディオとビジュアルの空間
  - S_{vat}： S_{va}を投影してテキストと比較する空間
- ネットワークの特性すべてをカバー



Multimodal Contrastive Loss

- 映像xが与えられたときに損失を最小化

$$\mathcal{L}(x) = \lambda_{va} NCE(x_v, x_a) + \lambda_{vt} MIL\text{-}NCE(x_v, x_t)$$

$$NCE(x_v, x_a) = -\log\left(\frac{\exp(z_{v,va} \cdot z_{a,va}/\tau)}{\exp(z_{v,va} \cdot z_{a,va}/\tau) + \sum_{z' \sim \mathcal{N}(x)} \exp(z'_{v,va} \cdot z'_{a,va}/\tau)}\right)$$

$$MIL\text{-}NCE(x_v, x_t) = -\log\left(\frac{\sum_{z \in P(x)} \exp(z_{v,vat} \cdot z_{t,vat}/\tau)}{\sum_{z \in P(x)} \exp(z_{v,vat} \cdot z_{t,vat}/\tau) + \sum_{z' \sim \mathcal{N}(x)} \exp(z'_{v,vat} \cdot z'_{t,vat}/\tau)}\right)$$

Multimodal Contrastive Loss

- NCE：Noise Contrastive Estimation
  - 真のペアであるか推定するために偽のペアと比較をして確率を出力

$$NCE(x_v, x_a) = -\log\left(\frac{\exp(z_{v,va} \cdot z_{a,va}/\tau)}{\exp(z_{v,va} \cdot z_{a,va}/\tau) + \sum_{z' \sim \mathcal{N}(x)} \exp(z'_{v,va} \cdot z'_{a,va}/\tau)}\right)$$

- $MIL\text{-}NCE$
  - MIL：Multiple Instance Learning
  - 複数のインスタンスに対するNCEの処理

τ ：温度パラメータ
$\mathcal{N}(x)$ ：偽のペアの集合
$P(x)$ ：真のペアの集合
$z()$ ：ネットワークの出力
$z'()$ ：偽のペア

$$MIL\text{-}NCE(x_v, x_t) = -\log\left(\frac{\sum_{z \in P(x)} \exp(z_{v,vat} \cdot z_{t,vat}/\tau)}{\sum_{z \in P(x)} \exp(z_{v,vat} \cdot z_{t,vat}/\tau) + \sum_{z' \sim \mathcal{N}(x)} \exp(z'_{v,vat} \cdot z'_{t,vat}/\tau)}\right)$$

おわりに

- MultiModal Versatile (MMV)
  - Shared spaces
  - Disjoint spaces
  - Fine and coarse spaces
- Multimodal Contrastive Loss

- 今後の予定
  - パソコンの環境構築
  - 前回と今回で調査した2つの論文の再現実験