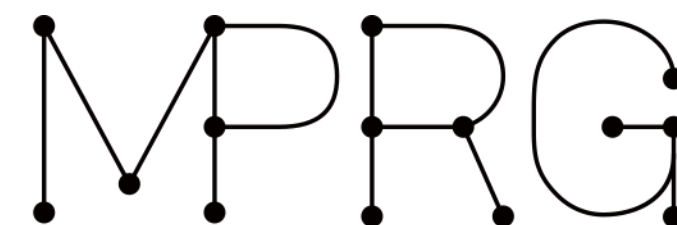


第23回ディスカッション

実験状況

ER20038 小林亮太

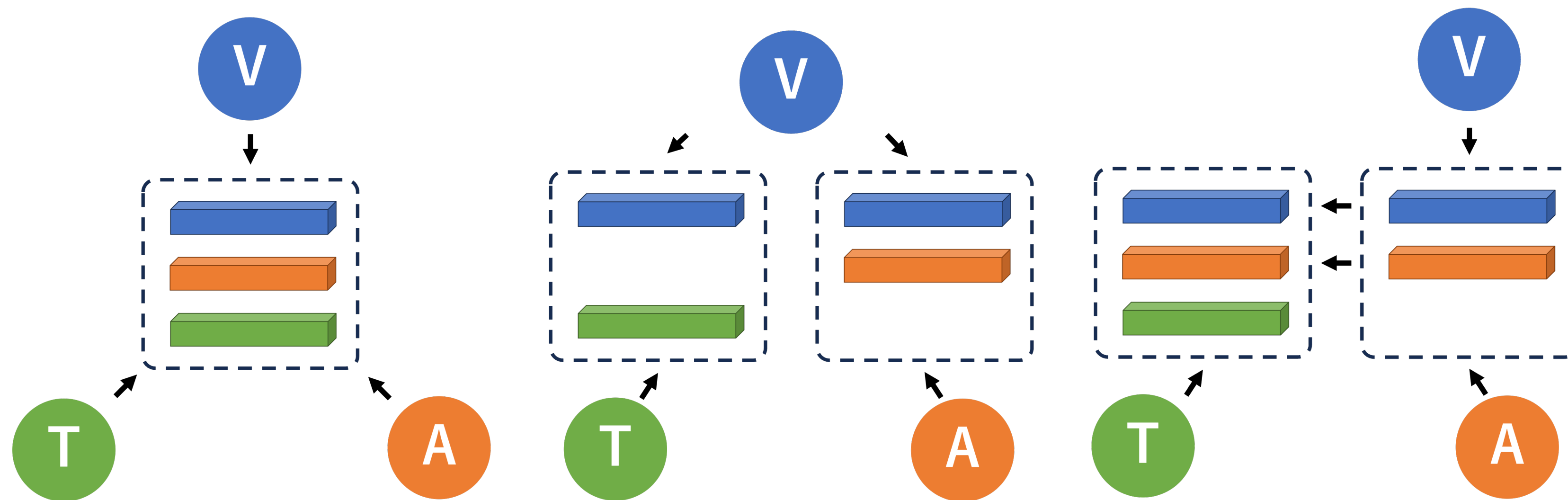
担当：鈴木雅★， 福井， 張



MACHINE PERCEPTION AND ROBOTICS GROUP

- 研究テーマ
- 実験概要
- 実験条件
- 実験状況

- 3モーダル（ビデオ，オーディオ，テキスト）のマルチモーダル自己教師あり学習
- テキストに比べビデオやオーディオにはノイズが多く存在
 - 各モーダルの組み合わせでノイズを抽出せずに学習ができる可能性
 - 近づけるモーダルの組み合わせによる学習効果への影響について調査



- 3モーダルを二段階で学習実験
 - 3パターンを実験
 - AV_T : AとVで学習した後からTを追加
 - VT_A : VとTで学習した後からAを追加
 - AT_V : AとTで学習した後からVを追加
- HowTo100Mデータセットを用いて学習
- YouCook2データセットを用いてゼロショットで評価
- テキストからビデオの検索タスクで評価
 - テキストによるビデオ内の該当箇所の検索

- Feature Extractor
 - ビデオ : ResNet152
 - オーディオ : DaveNet [D Harwath+, ECCV'18]
 - テキスト : Word2vec
- バッチサイズ : 128
- エポック数 : 30
 - 前半20, 後半10に設定
- 学習率 : 0.0001
- 最適化手法 : Adam
- GPU : A100 × 4

- AV_Tのパターンが完了
- AT_Vのパターンを実験中
 - 1段階目でLossがNaNになる問題が発生
 - Clustering Lossの計算を含む場合のみ発生
 - Clustering Lossの計算結果を割る処理を追加
 - 対応済み
- 一段階目3/20epochまで完了
 - 日曜日に完了予定
 - 完了後二段階目開始予定

- 評価タスク：テキストからビデオの検索 on YouCook2
 - R@k
 - R：Recall（再現率）
 - K：各クエリの上位k個の予測のうちの正解数の総正解数に対する割合

		R@1	R@5	R@10
論文		18.1	35.5	45.2
実験1	再現実験	14.7	34.1	44.7
実験2	AV_T			
	VT_A		実験中	
	AT_V			

- 実験 : 実行中
- 今後の予定 :
 - 他のパターンの実験
 - 卒論の修正

Multimodal Clustering Network (MCN) [B. Chen+, ICCV'21]

- ラベル付けされていないナレーション付きビデオから学習
 - テキストからビデオの検索, 時系列行動検出が可能
- テキスト, オーディオ, ビデオの3つのモーダルを使用

