東京大学 THE UNIVERSITY OF TOKYO

JST ACT-I「情報と未来」先端研究フォーラム

OMRON

研究課題：多様なデータへのキャプションを自動で生成する技術の創出

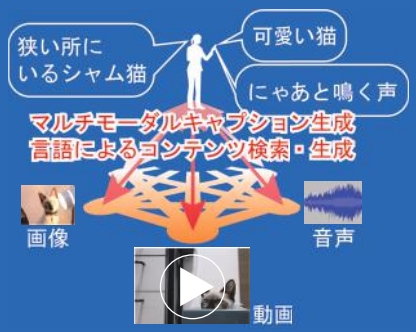
OMRON SINIC X 牛久祥孝

あなたの見たものを言葉にする技術

未来ビジョン

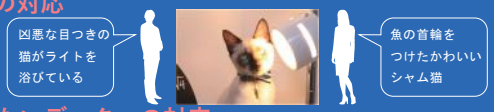
- 言語をクエリとしたメディアの**検索・生成**
- 視聴に時間のかかるメディアを**キャプション生成**して表示

多様なデータへのキャプション生成



未来ビジョン達成のために必須＋未達の4要求機能

(i) 個人の属性や傾向への対応
(ii) 詳細な表現への対応



(iii)教師キャプションがないデータへの対応
(iv)ストーリーの多様性への対応

個人の属性や傾向への対応

個人へのキャプション生成転移へ向けた現実的なドメイン適応

ドメイン適応：教師付きデータ（ソースドメイン）から教師が無い/欠けたデータ（ターゲットドメイン）への転移

画像からの物体検出をイラストへ適応した例 [Saito+, 2019]

ただし通常のドメイン適応は**ラベルノイズが無く、ソースとターゲットでクラスが一致している**
➡ラベルノイズがあり、ソースとターゲットで潜在的にクラスが一致していないドメイン適応

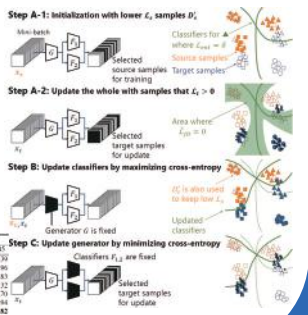
2つの識別器によるNoisy Universal Domain Adaptation

A) ノイズ除去ステップ：
1. ソースデータでの学習
   識別器間の意見の食い違いと推定の曖昧さ＋損失が大きい＝**ノイズ**
   ➡上記のようなサンプルを除外した上でモデルの初期値を学習
2. ターゲットデータでの学習
   - **2つの識別器の食い違い**または**識別器での曖昧さが一定以下**
     ➡ソースにも存在するクラス
   - **一定以上**➡ターゲットにしか存在しないクラス

B) クロスエントロピー最大化ステップ：
2つの識別器の食い違いを最大化するように識別器を更新

C) クロスエントロピー最小化ステップ：
2つの識別器の食い違いを最小化するように分布（特徴量）を更新

上記Step BとStep Cを**ミニバッチごとに繰り返す**



実験結果
提案手法が最もノイズ(S/Pxx)に強いことを複数データで確認

CVPR 2021 投稿中

詳細な表現への対応

人がすばやく見分けられる差分に注目した記述生成

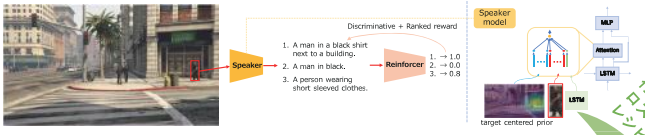
正しく差分を記述していても**見つけられない**

差分の記述が過剰だと**読解に時間がかかる**

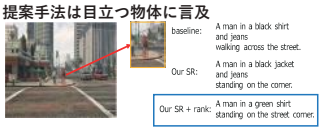
CGデータセットの収集＋公開

ランク学習と視覚的顕著性＋アテンションによる生成モデル

行って帰ってくるロスによる詳細表現



提案手法は目立つ物体に言及



単語ごとに注目する情報を可視化

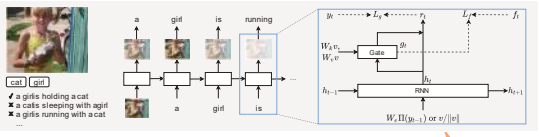
ICCV 2019 採録！

教師キャプションがないデータへの対応

疑似キャプションを用いた教師なしキャプション生成

画像と文がバラバラのデータ➡含まれる物体が同じ画像と文を疑似的にペア化・学習

- ゲートの活用
  - ✓ 画像を用いた単語推定
  - ✓ 生成中の文を用いた単語推定
- ゲートの疑似教師
  - ✓ ゲートの教師情報は存在しない
  - ✓ だが疑似ペアリングの根拠になった物体は画像を用いるべき
    ➡ゲートの疑似教師情報



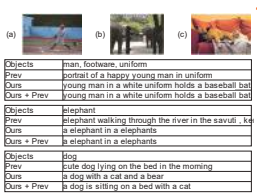
定量的な実験結果

- 先行研究[Feng+, 2019][Laina+, 2019]と比較➡提案手法はよりシンプルかつ高精度
  BLEU-x, METEOR, ROUGE-L, CIDEr, SPICE…文生成の評価尺度　値が高い方が正解文に近い

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[Feng+, 2019]	41.0	22.5	11.2	5.0	12.4	28.7	28.6	8.1
Ours	**49.5 ± 0.7**	**27.3 ± 1.2**	**13.1 ± 0.8**	**6.3 ± 0.5**	**14.0 ± 0.1**	**34.5 ± 0.3**	**31.9 ± 1.0**	**8.6 ± 0.2**
[Laina+, 2019]				6.5	12.9	35.1	22.7	
Ours	50.4 ± 1.5	29.5 ± 0.8	14.4 ± 0.5	7.6 ± 0.4	13.5 ± 0.3	37.3 ± 0.2	31.8 ± 0.7	8.6 ± 0.1

定性的な実験結果

- ペアの教師情報がなくてもキャプション生成できる
- 先行研究[Feng+, 2019]と組合せると精度up

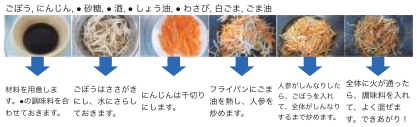


EACL 2021 採録！

疑似材料ツリーを用いた教師付きレシピ生成へ転用

ストーリーの多様性への対応

構造に配慮した写真列からの手順書生成

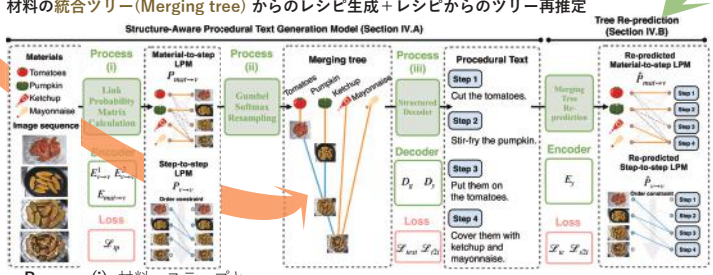


動画＝主要な静止画の系列とみて材料＋静止画系列➡レシピ生成

画像列・レシピ、材料リストからなるデータセットvSIMMRの構築・公開

Datasets	Recipes?	Ingredients?	Structure?	Visual data	#Recipes
Breakfast [6]				Video	N/A
EPIC-Kitchens [7]				Video	N/A
YouCook2 [8]	✓			Video	89
RecipeQA [9]	✓			Image sequence	19,779
Story boarding [1]	✓			Image sequence	16,405
Cookpad Image Dataset [10]	✓	✓		Image sequence	1,715,595
Recipe Flow Graph [5]	✓	✓	✓	N/A	266
Action Graph [4]	✓	✓	✓	N/A	133
SIMMR [3]	✓	✓	✓	N/A	260
MM-ReS [11]	✓		✓	Image sequence	9,850
vSIMMR (ours)	✓	✓	✓	Image sequence	2,103

材料の統合ツリー(Merging tree) からのレシピ生成＋レシピからのツリー再推定



- **Process (i):** 材料→ステップとステップ→ステップの結合予測
- **Process (ii):** Gumbel softmaxを用いた**微分可能ツリー生成**の半教師あり学習＋**疑似ツリー利用**
- **Process (iii):** ツリーLSTMから各ステップのレシピを生成
- **Tree Re-prediction:** レシピから上記2種の結合を**再予測**

vSIMMRを用いた実験結果
Ablation studyから材料ツリーとレシピの生成結果の正確さを確認

IEEE Access 採録！

まとめ

- **複数の要求機能を満たすキャプション生成技術**
  - ✓ **行って帰ってくるロス**による詳細表現
  - ✓ **疑似教師情報**による教師データ不足への対応
  - ✓ **材料統合ツリー**を用いた作業動画のストーリー理解
- **今後の展望**
  - ✓ コロナ禍で遅滞した個人の属性や傾向を含む総合データセットの整備（クックパッド社との共同研究）
  - 個人への対応を含めた総合評価
  - 画像キャプション生成＝異なるモダリティ間の変換
    ➡今後はより多様なモダリティ間のクロスモーダル変換へ