PARTIALLY-SUPERVISED NOVEL OBJECT CAPTIONING USING CONTEXT FROM PAIRED DATA

Paper ID: 0649

Shashank Bujimalla[1], Mahesh Subedar[2], Omesh Tickoo[2]

Intel Corporation[1], Intel Labs[2], USA

BMVC 2022

Our Approach: PS-NOC

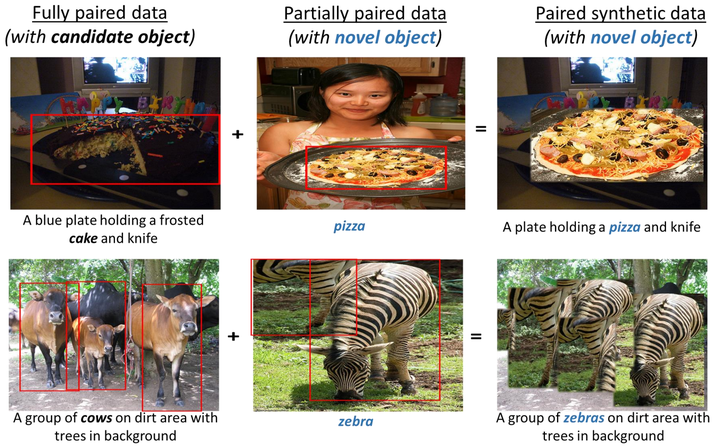
Problem Statement:

- Caption images with novel objects that lack fully paired image-captions in the training dataset.
- Use object detection datasets (partially paired data) to generate captions that correctly include these novel objects.

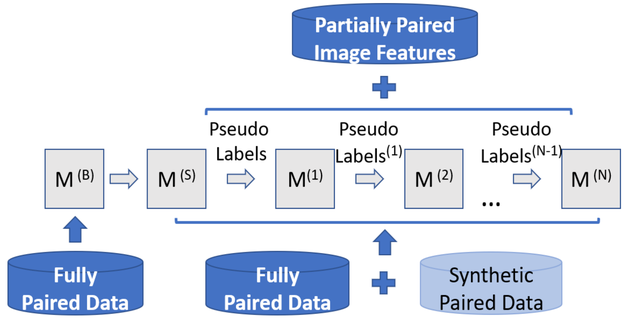
PS-NOC: Partially-Supervised Novel Object Captioning

- Agnostic to model architecture
- Uses partially paired data
  - Generate paired synthetic data before training
  - Generate pseudo-label data during training
- Three-step training process
  - Use fully-paired data, synthetic data, pseudo-label data
  - Novel techniques: SCST-F1, Pseudo-labeling for NOC

Synthetic Data Generation



Training Technique



Training Steps:

- Step I: Train using fully-paired data → $M^{(B)}$ (Cross entropy + SCST-F1 loss)
- Step II: Fine-tune using synthetic data → $M^{(S)}$ (Cross entropy + SCST-F1 loss)
- Step III: Generate pseudo-labels on partially-paired data and fine-tune → $M^{(1)}$, $M^{(2)}$, …, $M^{(N)}$ (SCST-F1 loss)

Results

PS-NOC provides the highest scores for both out-of-domain CIDEr and F1-scores compared to SOTA.

Approach	C-RL	Out-of-domain S	M	C	F1	CF1	CF1.5	In-domain S	M	C
PS-NOC (Sol-1)	Yes	**19.7**	**27.2**	**101.5**	**86.1**	**93.2**	**96.2**	19.2	26.9	110.1
PS-NOC (Sol-2)	Yes	**20.8**	**28.0**	**103.8**	**85.9**	**94.0**	**97.6**	20.5	27.7	110.9
PS3 [4]	No	17.9	25.4	94.5	63	75.6	81.9	19.0	25.9	101.1
FDM (no CBS) [8]	No	19.4	25.9	84.8	64.7	73.4	77.4	20.2	27.2	109.7
FDM (CBS) [8]	No	19.6	25.6	85.3	85.7	85.5	85.4	19.7	26.2	105.5
NBT (CBS) [19]	No	17.4	24.1	86.0	70.3	77.4	80.5	18.0	25.0	92.1
Reg. Sel. [7]	No	18.3	24.9	78.2	75.0	76.6	77.2	19.2	26.2	97.0
Reg. Sel. (DGBS) [7]	Yes	19.4	26.3	88.5	75.1	81.3	83.9	21.0	27.9	115.3
ANOC [9]	Yes	18.2	25.2	94.7	64.3	76.6	82.7	-	-	-
ECOL-R (CBS) [26]	Yes	19.1	25.7	99.1	71.8	83.3	88.7	20.8	26.8	112.6

Examples of captions generated using our approach

Baseline: A group of animals that are standing in the grass. **Ours**: Group of **zebras** are standing in a field.

Baseline: A blue and white car on a city street. **Ours**: A **bus** driving down a city street with cars.

Baseline: A woman laying on a bed with a laptop. **Ours**: A woman sitting on a **couch** using a laptop computer.

Baseline: A man is cutting a large pie on a plate. **Ours**: A **pizza** on a plate with a knife.

Baseline: A little girl standing on top of a lush green field. **Ours**: A little girl holding a tennis **racket** on a field.

Baseline: A cat is sitting in a blue bowl. **Ours**: A cat sitting in a **suitcase** on a floor.

Training technique	Out-of-domain S	M	C	F1	In-domain S	M	C
I	19.6	28.0	69.7	0.0	19.4	27.9	108.0
I + CBSInf	18.1	26.1	76.6	56.2	17.6	25.8	88.8
I + II	19.8	28.2	89.0	62.1	19.3	27.6	103.4
I + II + III	19.9	28.4	96.3	75.8	19.5	27.8	105.9
I + II(SCST)	20.5	27.8	98.6	70.8	20.2	27.6	113.4
I + II(SCST) + III	20.2	28.3	99.8	72.4	19.9	27.9	108.2
I + II(SCST) + III(SCST)	20.1	27.8	101.0	78.8	19.6	27.1	111.0
I + II(SCST-F1)	20.6	28.1	99.2	76.4	19.9	27.6	111.3
I + II(SCST-F1) + III	20.2	28.5	102.2	75.7	20.0	28.2	108.2
I + II(SCST-F1) + III(SCST-F1) Sol-1	19.7	27.2	101.5	86.1	19.2	26.9	110.1
I + II(SCST-F1) + III(SCST-F1) Sol-2	20.8	28.0	103.8	85.9	20.5	27.7	110.9

Ablation Studies: Results demonstrate the benefits of using i. Synthetic data, ii. Pseudo-labeling, iii. SCST, iv. SCST-F1, and v. Our overall approach

Conclusions

- **PS-NOC gives improvements over baseline and previous works**
  - Held-out MS-COCO: Out-of-domain CIDEr 103.8, F1-score 85.9
- **PS-NOC uses fully paired data and partially paired data effectively**
  - Paired synthetic data generation: Generic and not restricted to region-based captioning models
  - Three-step training process: Effective and includes novel techniques