# Can *language* representations emerge from a purely *image*-pretrained model?

-- An initial exploration in multimodal learning:
*Look Twice as Much as You Say: Scene Graph Contrastive Learning for Self-Supervised Image Caption Generation, CIKM'22, Chunhui Zhang, et al.*

# Look Twice as Much as You Say

- ❑ Introduction - *vision language learning*
- ❑ Method
- ❑ Experiment
- ❑ Discussion

*About me:*
Hello, my name is Chunhui Zhang, and I am currently a second-year Ph.D. student in Computer Science at Brandeis University.
My research interests and experience span a range of areas, including learning representations from diverse modalities, trustworthy machine learning, and efficient machine learning.
My prior works have been accepted to top-tier conferences, such as ICLR'23, NeurIPS'22, WWW'23, CIKM'22, etc.

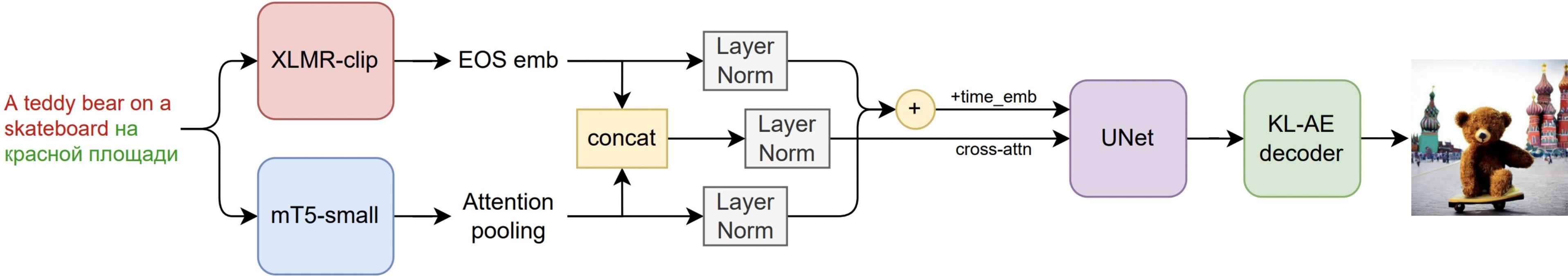
# Introduction - Vision Language Learning (modality gap between vision and language is large)

## ❑ 1. Image2Text (e.g., M^2 Transformer)

**GT:** A cat looking at his reflection in the mirror.
**Transformer:** A cat sitting in a window sill looking out.
$\mathcal{M}^2$ **Transformer:** A cat looking at its reflection in a mirror.

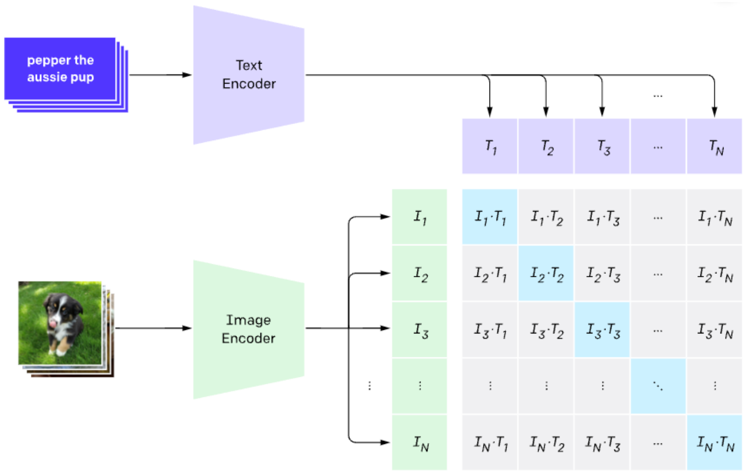
**GT:** A plate of food including eggs and toast on a table next to a stone railing.
**Transformer:** A group of food on a plate.
$\mathcal{M}^2$ **Transformer:** A plate of breakfast food with eggs and toast.

## ❑ 2. Text2Image (e.g., Diffusion Model)

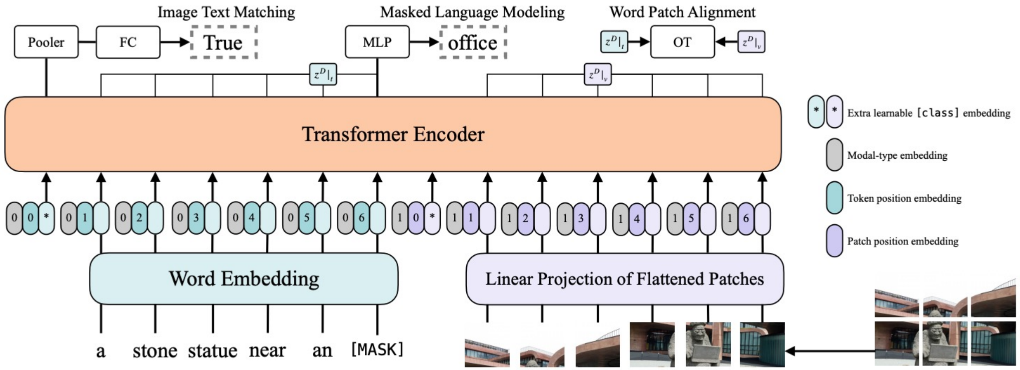


# Introduction - Vision Language Learning (modality gap between vision and language is large)

## ❑ 3. Models for both modalities (e.g., CLIP, ViLT)




(a) CLIP




(b) ViLT

# Introduction - Vision Language Learning (modality gap between vision and language is large)

Summary:

- ❑ Method 1 requires ***ground-truth text*** for supervised loss (e.g., CE loss)
- ❑ Method 2 requires ***ground-truth image*** for supervised loss (e.g., MSE loss)
- ❑ Method 3 requires ***well-paired image-text input*** for (un)supervised loss

All three of the above popular cross-modal learning paradigms have **limitations** on the training data, to mitigate the large modality **gap** between vision and language.

To tackle these limitations, we start with a question in image caption generation:

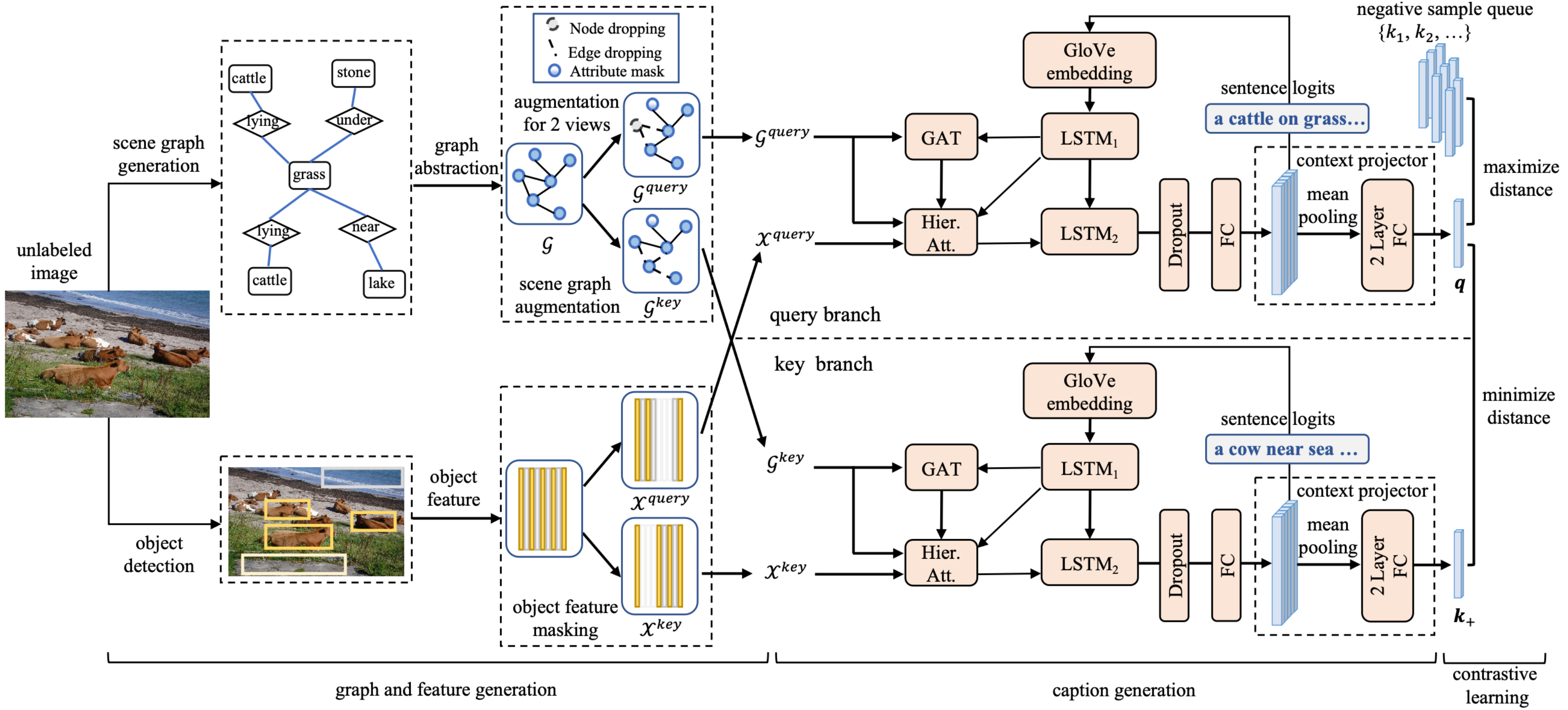
Can language representations emerge from a purely image-pretrained model?

# Look Twice as Much as You Say

- ❑ Introduction - *vision language learning*
- ❑ Method - Scene Graph Contrastive Learning
- ❑ Experiment
- ❑ Discussion

# Method – Scene Graph Contrastive Learning

- Contrastive training with purely visual Inputs, to learn (pseudo) caption sentence generation



# Look Twice as Much as You Say

- ❑ Introduction - *vision language learning*
- ❑ Method - Scene Graph Contrastive Learning
- ❑ Experiment - effectiveness of emerged language representations in image-pretrained model on captioning
- ❑ Discussion

# Experiment – Effectiveness of emerged language representations

- Performance: finetuning with very few ground-truth captions

image:

**1% labels are used:**

| | | | | | |
|---|---|---|---|---|---|
| $M^2$-T: | A sheep standing in a a a. | A man girl a a a a a. | A girl tennis a a tennis tennis | A elephant of in a a a. | A cow standing standing a a a. |
| SGAE: | A sheep of standing a a a. | A man is a a a a a. | A man girl a a a a a. | A man is a a a a a. | A sheep of in a a a a a. |
| VSUA: | A sheep of in a a a. | A man standing a a a a. | A girl girl a a a a a. | A people of a a a a a a. | A cow cow cow a a a a a. |
| C-GAT: | A group of a a a a. | A man is a a a a a. | A man girl a a a a a. | A street of a a a a. | A sheep of a a a. |
| **SGCL:** | **A couple of sheep standing in the grass in a field.** | **A group of people playing a frisbee standing by the sea.** | **A woman is holding a tennis racket on a tennis ball.** | **A group of people standing in a street with a building.** | **A herd of cows walking down a road in the grass.** |

**30% labels are used:**

| | | | | | |
|---|---|---|---|---|---|
| $M^2$-T: | A group of sheep standing in a fenced area. | Two men playing frisbee on a dirt field. | A woman is holding a tennis racket in her hand. | A man riding an elephant in front of a building. | A group of cows standing next to each other on a field. |
| SGAE: | A white sheep is standing in the grass. | A group of men playing a game of frisbee. | A man hitting a tennis ball on a tennis court. | A group of people standing next to an elephant. | A cow standing on top of a lush green field. |
| VSUA: | A couple of sheep standing next to each other. | A man holding a frisbee in his hand. | Two men playing frisbee on a dirt field. | An elephant standing in front of a building. | A group of cows are standing in the grass. |
| C-GAT: | A group of sheep grazing in a grassy field. | A man holding a frisbee in his hand. | A woman is playing tennis on the court. | A man riding on the back of an elephant. | A brown cow standing next to a brown cow. |
| **SGCL:** | **A couple of sheep standing on a lush green field near a fence.** | **A man is jumping in the air to catch a frisbee on a sea beach.** | **A woman is trying to hitting a tennis ball on a tennis court.** | **An elephant walking down a street with people in the background.** | **A herd of black cows standing next to each other on a lush green field.** |

# Experiment – Effectiveness of emerged language representations

- Comparisons with popular baselines in caption metrics

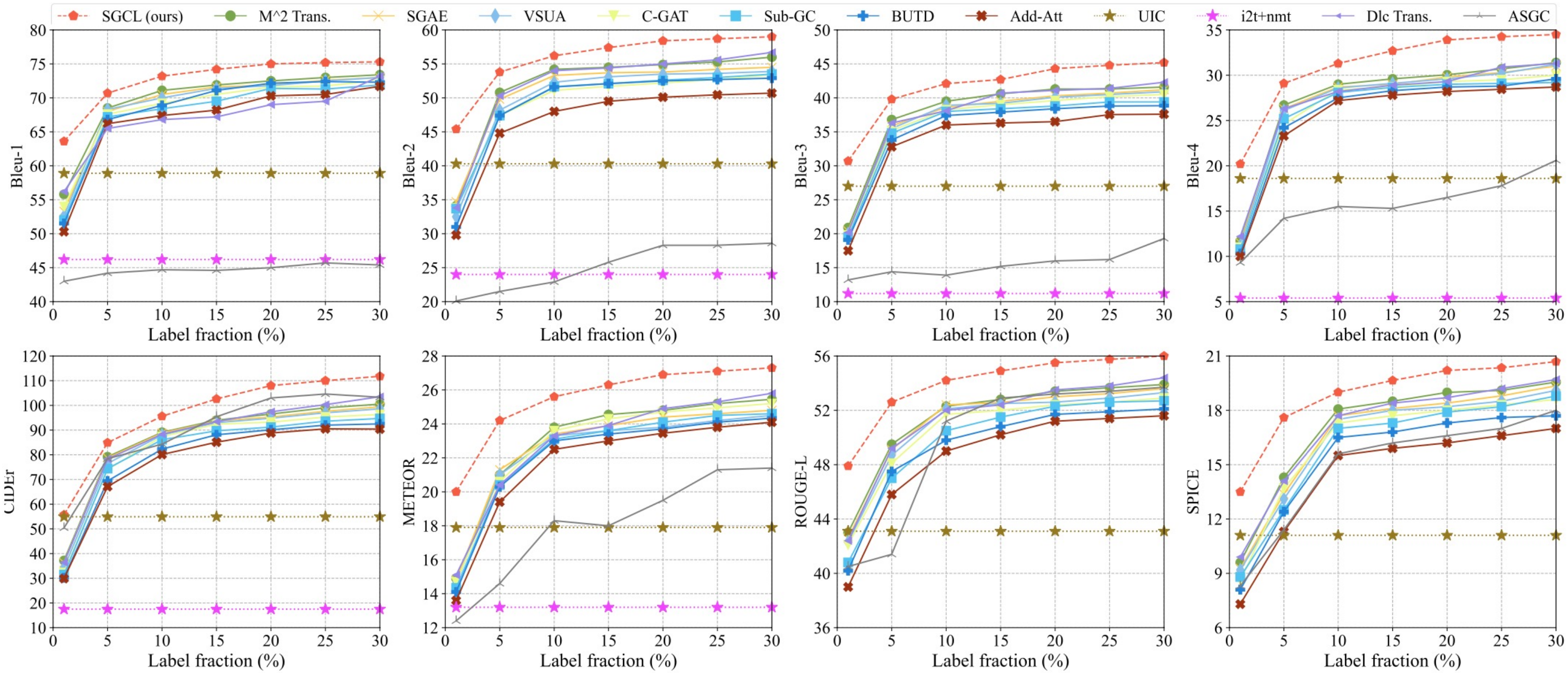



**Figure 3: Performances of all models with limited labels (Note that ROUGE-L and SPICE of i2t+nmt are not shown due to missing values in the original work).**

# Experiment – Effectiveness of emerged language representations

- Ablation study: effectiveness of scene graph augmentation in contrastive learning

**Table 1: Performances of different model variants with various graph augmentation strategies (Note: N - node dropping, E - edge dropping, A - node attribute masking, O - object feature masking).**

| Label | N | E | A | O | B-1 | B-2 | B-3 | B-4 | C. | M. | R.-L | S. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1% | | | | | 61.8 | 43.8 | 28.9 | 18.2 | 47.7 | 18.5 | 46.3 | 11.9 |
| | | | | ✓ | 62.5 | 44.6 | 30.0 | 19.1 | 49.2 | 19.9 | 47.0 | 13.1 |
| | ✓ | | | ✓ | 62.5 | 44.5 | 29.0 | 18.5 | 52.9 | 19.1 | 47.3 | 13.2 |
| | | ✓ | | ✓ | 63.1 | 44.3 | 28.8 | 18.6 | 52.2 | 19.3 | 47.2 | 13.0 |
| | | | ✓ | ✓ | 63.0 | 45.1 | 29.9 | 19.3 | 53.3 | 19.6 | 47.5 | 13.3 |
| | ✓ | ✓ | ✓ | ✓ | **63.6** | **45.4** | **30.7** | **20.2** | **55.0** | **20.0** | **47.9** | **13.5** |
| 5% | | | | | 69.4 | 51.7 | 36.6 | 26.2 | 75.9 | 22.2 | 49.4 | 16.3 |
| | | | | ✓ | 70.3 | 53.0 | 38.6 | 27.9 | 79.4 | 23.9 | 51.9 | 17.3 |
| | ✓ | | | ✓ | 62.5 | 52.8 | 38.9 | 28.5 | 81.4 | 19.1 | 51.9 | 17.2 |
| | | ✓ | | ✓ | 63.1 | 53.5 | 38.7 | 28.6 | 82.2 | 19.3 | 52.0 | 17.1 |
| | | | ✓ | ✓ | 70.3 | 53.3 | 39.2 | 28.1 | 82.3 | 24.1 | 52.2 | 17.4 |
| | ✓ | ✓ | ✓ | ✓ | **70.7** | **53.8** | **39.8** | **29.1** | **84.9** | **24.2** | **52.6** | **17.6** |

# Experiment – Effectiveness of emerged language representations

- Ablation study: effectiveness of big dropout rate in contrastive learning

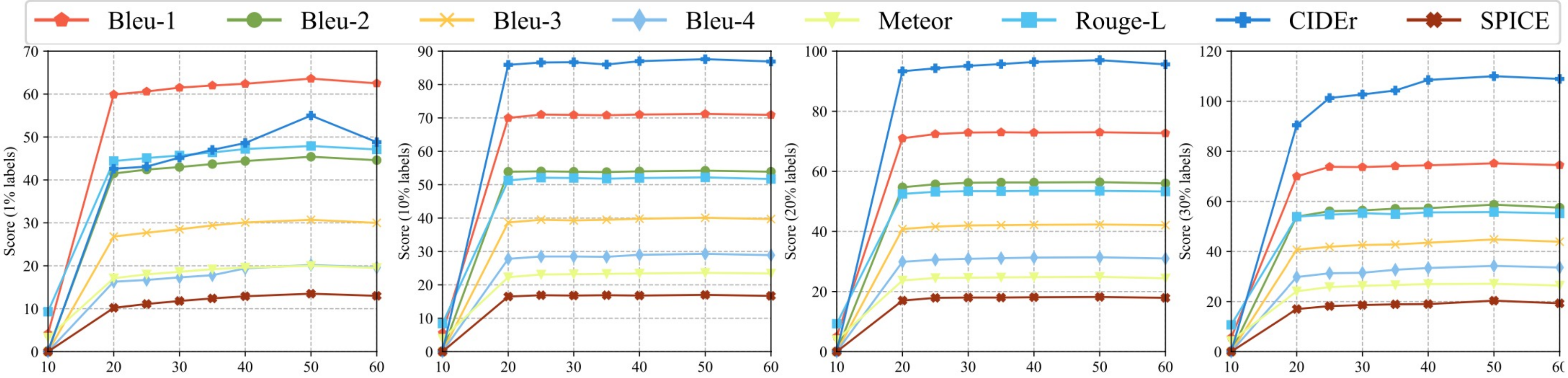



**Figure 4: Impact of dropout rate at the output layer on model performance.**

# Look Twice as Much as You Say

- ❑ Introduction - *vision language learning*
- ❑ Method - Scene Graph Contrastive Learning
- ❑ Experiment - effectiveness of emerged language representations in image-pretrained model on captioning
- ❑ Discussion - conclusion and hypothesis

# Discussion

## Conclusion

- Language representations ***can*** emerge from a purely image-pretrained model, even on small model in this work (i.e., LSTM_1 -> Attention -> GAT -> Attention -> LSTM_2).
- Although the image captioning model only uses **pseudo** sentence logits in the calculation of contrastive loss, the pre-trained model weights provide an effective initialization (i.e., 1% label finetuning brings impressive caption performance).

## Hypothesis *- for potential future exploration*

- LSTM brings architecture prior related to NLP: it projects the image representation to the language dimension in a sequential manner like sentence.
- GAT introduces semantic representations: it extracts scene graph representation as one of the inputs to LSTM, which encodes object attributes and relations in images and enhance the semantic/language representations learned by LSTM.

# Thank you!

# Q & A