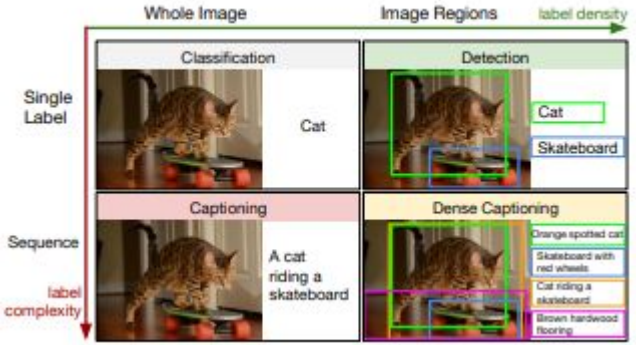
Fine-grained image captioning

Team members: Ting-Chih Chen, Xiao Guo, and Yi Lu

Introduction of the problem

- Traditional Image-to-text **fails** to capture:
  - Details and attributes
  - Relationship between objects
- Challenges:
  - Precise object and attribution recognition
  - Language generation



Johnson, Justin, Andrej Karpathy, and Li Fei-Fei. "Densecap: Fully convolutional localization networks for dense captioning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

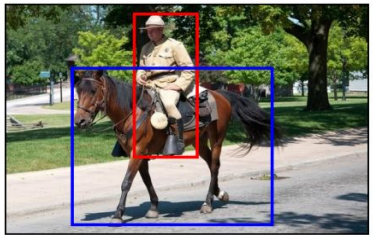


Image Sentence Captioning

A man in uniform riding a brown horse.

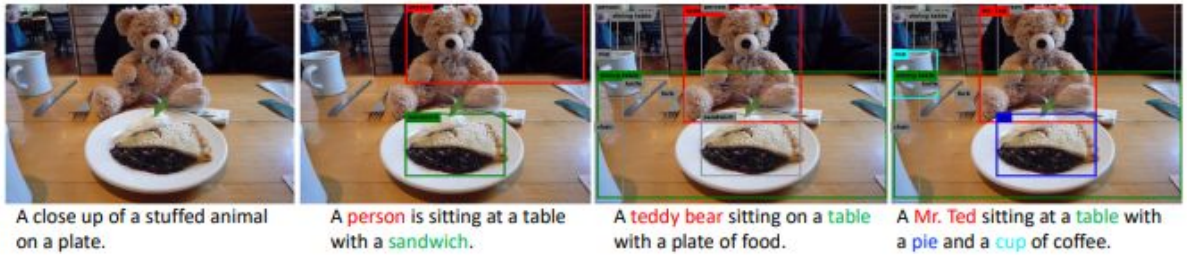
Image Paragraph Captioning

A brown horse walking on the road. A man wearing a uniform and a hat. He is riding the horse. There are some trees in the distance.

Zha, Zheng-Jun, et al. "Context-aware visual policy network for fine-grained image captioning." IEEE transactions on pattern analysis and machine intelligence 44.2 (2019): 710-722.

Related work

- Current research limitations: Indistinguishable captioning consists of most used words
- "Knowing when to look: Adaptive attention via a visual sentinel for image captioning." by Lu et al. 2017
  - Resnet for image encoder and LSTM for language model.
- "Neural baby talk." Same group, 2018.
  - filling slots in a sentence template using the detected visual objects.
- ***Fine Grained***: “Densecap: Fully convolutional localization networks for dense captioning” by Johnson et al. 2016
  - CNN + novel Dense Localization layer + RNN to generate label sequences.
- “Context-aware visual policy network for fine-grained image captioning” by Zha, Zheng-Jun, et al. 2019
  - Consider previous and current visual attention as context.



Lu, Jiasen, et al. "Neural baby talk." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018

Dataset

Flicker 30K

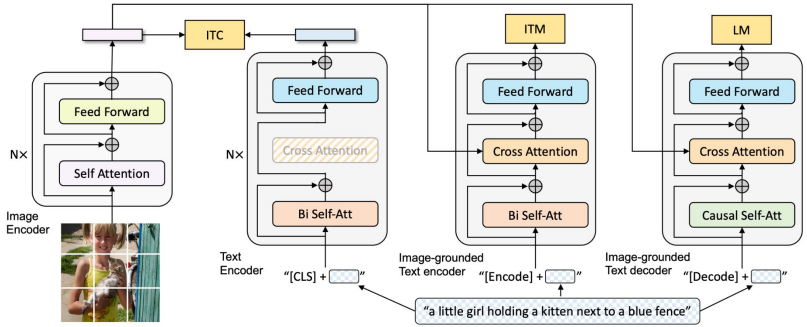
- A popular benchmark
- It contains 31,783 images collected from Flickr.
- Capture people engaged in everyday activities and events.
- Each image has 5 descriptive captions provided by human annotators.

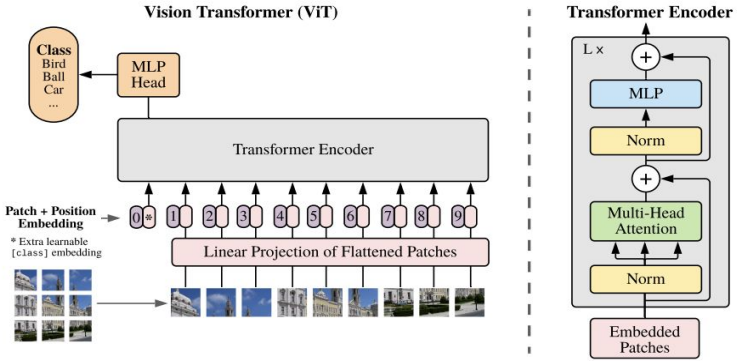
MS COCO

- The dataset consists of 328K images.
- Easily recognizable by 4-year-old kid.
- Variety of annotations.
- Versatile for object detection, captioning, stuff image segmentation and others.

Methodology - LAVIS

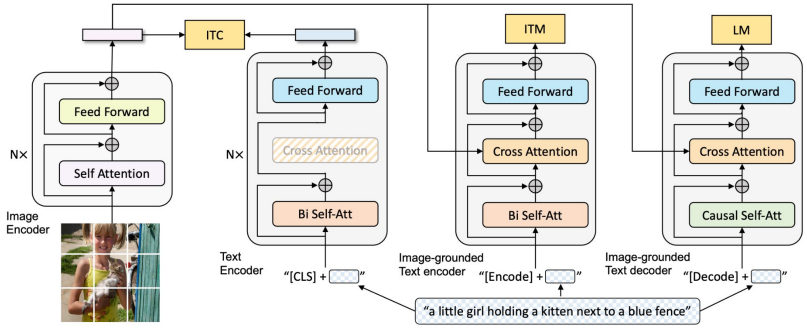
- **BLIP** model trained by LAVIS for image captioning tasks on the MSCOCO dataset
- BLIP uses **ViT** for image recognition.
- **ViT** split images into patches and flatten these patches to create lower-dimensional embeddings
- Afterwards it passes these embeddings to the transformer with **positional embeddings** for further processing

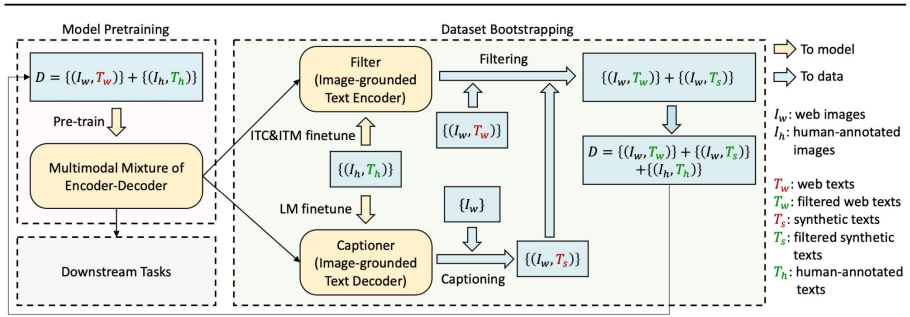




Methodology - LAVIS

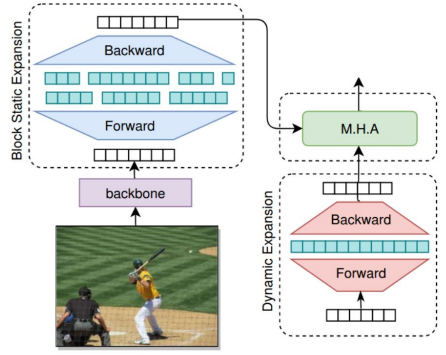
- For better captioning texts, BLIP uses **CapFilt** (caption filtering)
- Due to the high cost of labeling large scale dataset, **noise** in captioning is unavoidable.
- BLIP introduces a **captioner** to produce captions, and a **filter** to remove noisy image-text pairs

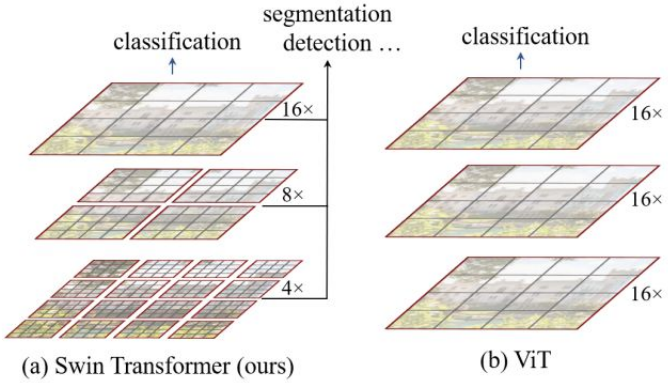




Methodology - ExpansionNet_v2

- ExpansionNet_v2 has a **SOTA** in MSCOCO 2014
- ExpansionNet_v2 changes the **sequence length**, dynamically or statically converts sequence input into sequences of different lengths
- Swin Transformer builds **hierarchical feature maps** by merging image patches and has linear computation complexity because self-attention only focuses on the local windows
- Swin Transformer is good for **image classification** and **dense recognition** tasks





Methodology - ExpansionNet_v2

- Forward expansion(**Encoder**)
- Backward expansion(**Decoder**)
- This model is **not limited by the length of the input**
- The red one is designed to **preserve the autoregressive property**
- The blue one can be operated in both **bidirectional processing and decoding stage**

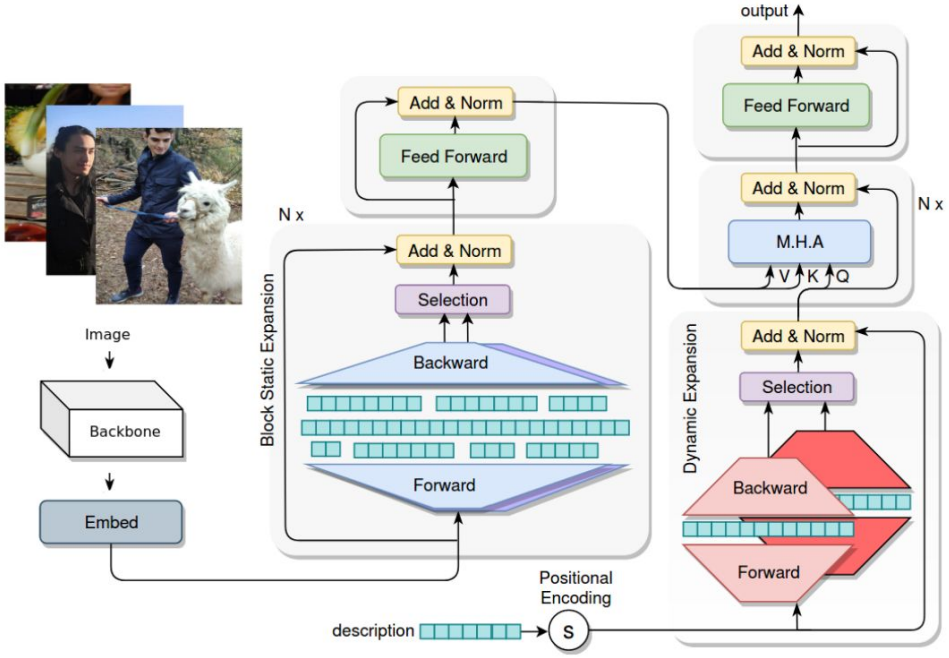


Figure 2: ExpansionNet v2 architecture.

Results & Performance

E: A sail boat is sitting in the ocean.
L: A sailboat in the middle of the ocean

E: A group of men walking down a city street.
L: A person holding a key to the eiffel tower'

E: A crowd of people standing in front of a tower
L: A person holding a key to the eiffel tower

Results & Performance

	BLeU-1	BLeU-2	BLeU-3	BLeU-4	Spice	Meteor	Rouge	Cider
Expansion Net_v2	**0.6776**	0.4848	0.3303	0.2236	0.1499	0.2192	0.4606	0.3834
Lavis	0.6386	0.4302	0.2922	0.1966	0.1565	0.2201	0.3887	0.4290

Table 1. Comparison of two model over different evaluation metrics

Conclusion(Analysis)

- Backbone: Swin transformer > ViT
- BLIP have caption filtering to denoise generated captions
- Future work:
  - How to know the captioning is fine-grained or not (new metric)
  - Continue working on improving the architecture of the ExpansionNet_v2
  - Investigate other works that strives to provide more attributes of the images

Thanks & Questions ?