Image Captioning via Vision and Language Transformers

Luke Davidson - davidson.lu@northeastern.edu
Kishore Pagidi - pagidi.k@northeastern.edu
Nand Dave - nand.da@northeastern.edu

April 20, 2023

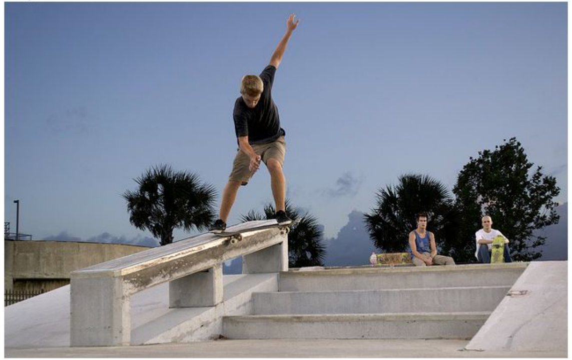
Motivation

What is Image Captioning?

- The combination of computer vision and natural language processing techniques to automatically generate accurate captions of unobserved input images

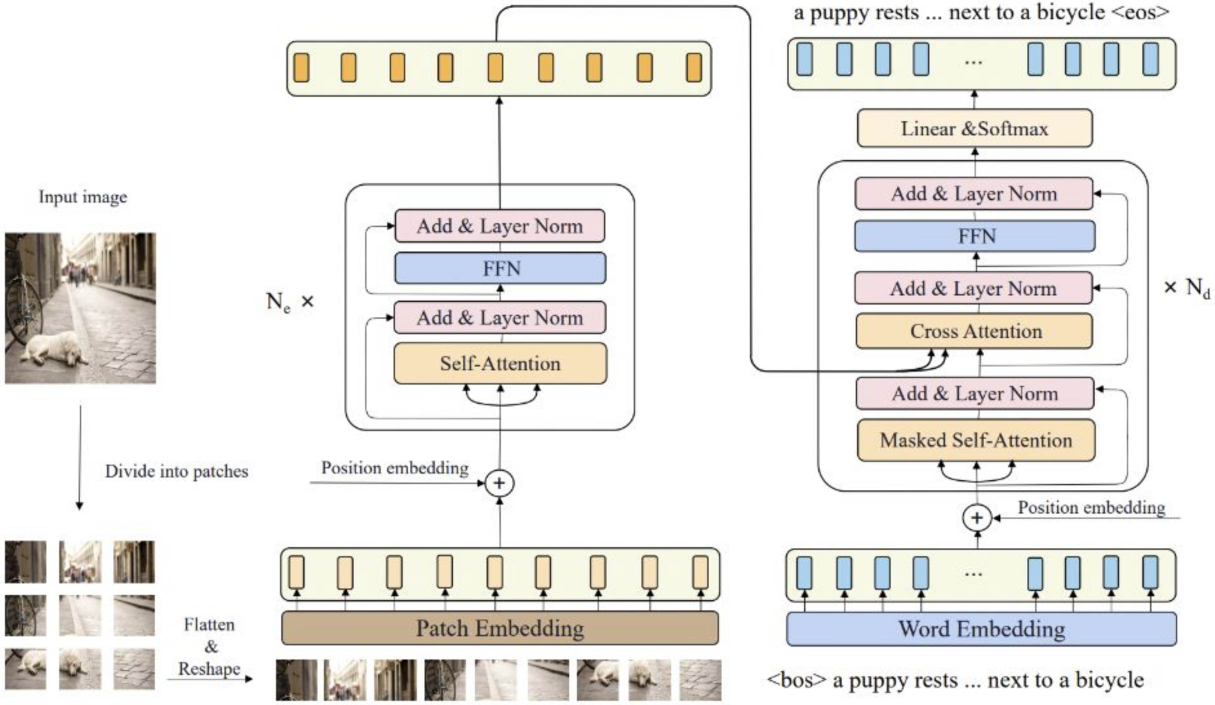
Use Cases

1. Marketing and media
    a. Automated captioning of social media posts or campaigns
2. Production
    a. Recommendation systems for image and video editing
3. Assistance Services
    a. Aiding visually impaired individuals in describing a live view or still image of something



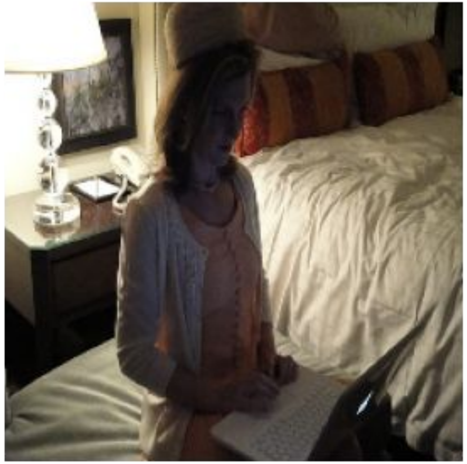
the young boys are practicing tricks on their skate boards.
a boy doing a trick on a skateboard on a rail.
this is a skateboarder doing a dangerous trick.
a boy on a skateboard going down a handrail at a set of stairs outside.
a boy skateboards down a hand rail while two others watch.

Approach

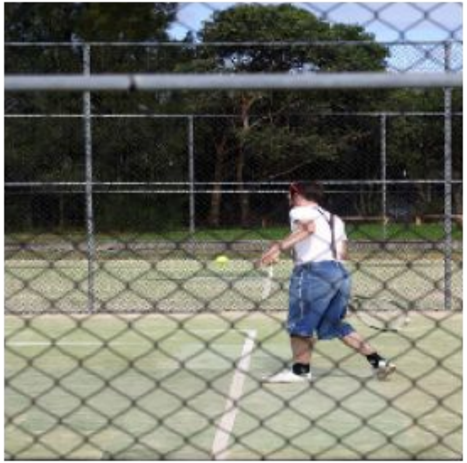


Input image
Divide into patches
Flatten
&
Reshape
Patch Embedding
Position embedding
N_e ×
Self-Attention
Add & Layer Norm
FFN
Add & Layer Norm
a puppy rests ... next to a bicycle <eos>
Linear &Softmax
Add & Layer Norm
FFN
Add & Layer Norm
Cross Attention
Add & Layer Norm
Masked Self-Attention
× N_d
Position embedding
Word Embedding
<bos> a puppy rests ... next to a bicycle

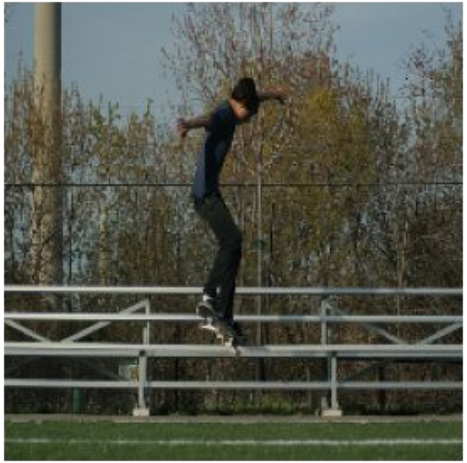
Results



A man is sitting on a bed with a laptop on the bed



A man in white shirt holding a bat and a ball



A man jumping over a skate board on a ramp

Conclusion

Takeaways

- Demonstrated the effectiveness of Transformer-based models for image captioning tasks.

Achievements

- Successfully generated coherent and contextually relevant captions for a wide range of images.

Future Work

- Evaluate and optimize the model's performance in real-world applications, such as accessibility tools and content generation.