[Week 11] Image Captioning

SWE3032-41 Artificial Intelligence Project

Mann Soo Hong[1], Soon Cheol Noh[2]
msjo91@skku.edu[1], yhnuhb27@skku.edu[2]
SKKU Information & Intelligence Lab

03/05/2021

Schedule Spring 2021

	Contents	Date
9	CNN, Mini-batch, Transforms, Training Flow	04/19~04/25
10	RNN, NLP, ~~Attention~~	04/26~05/02
11	**Image Captioning**	05/03~05/09
12	Image Segmentation, Object Detection, **Final Project Announcement**	05/10~05/16
13	Generative Models	05/17~05/23
14	Final Project QnA, ETC. (TBD)	05/24~05/30
15	Final Project QnA, **Final Project Deadline**	05/31~06/04

So Far

- Learned some basics of deep learning (from our professor)
- Tried out tutorial plotting, logging, image augmentation, etc.
- Practiced 101 computer vision and natural language processing with CNN and RNN

Computer Vision

- **Computer vision (CV)** is a scientific field that studies how machines can understand concepts from digital images or videos

- We want computers to be able to **watch & learn**

- Object/event/motion/… detection, video tracking, facial/character/… recognition, semantic segmentation, motion/pose/… estimation, image restoration, scene reconstruction, 3D modelling, style transfer, etc.

Natural Language Processing

- **Natural Language Processing (NLP)** is a field that studies how machines can interact with people via human language

- We want computers to **communicate with us in our own language**

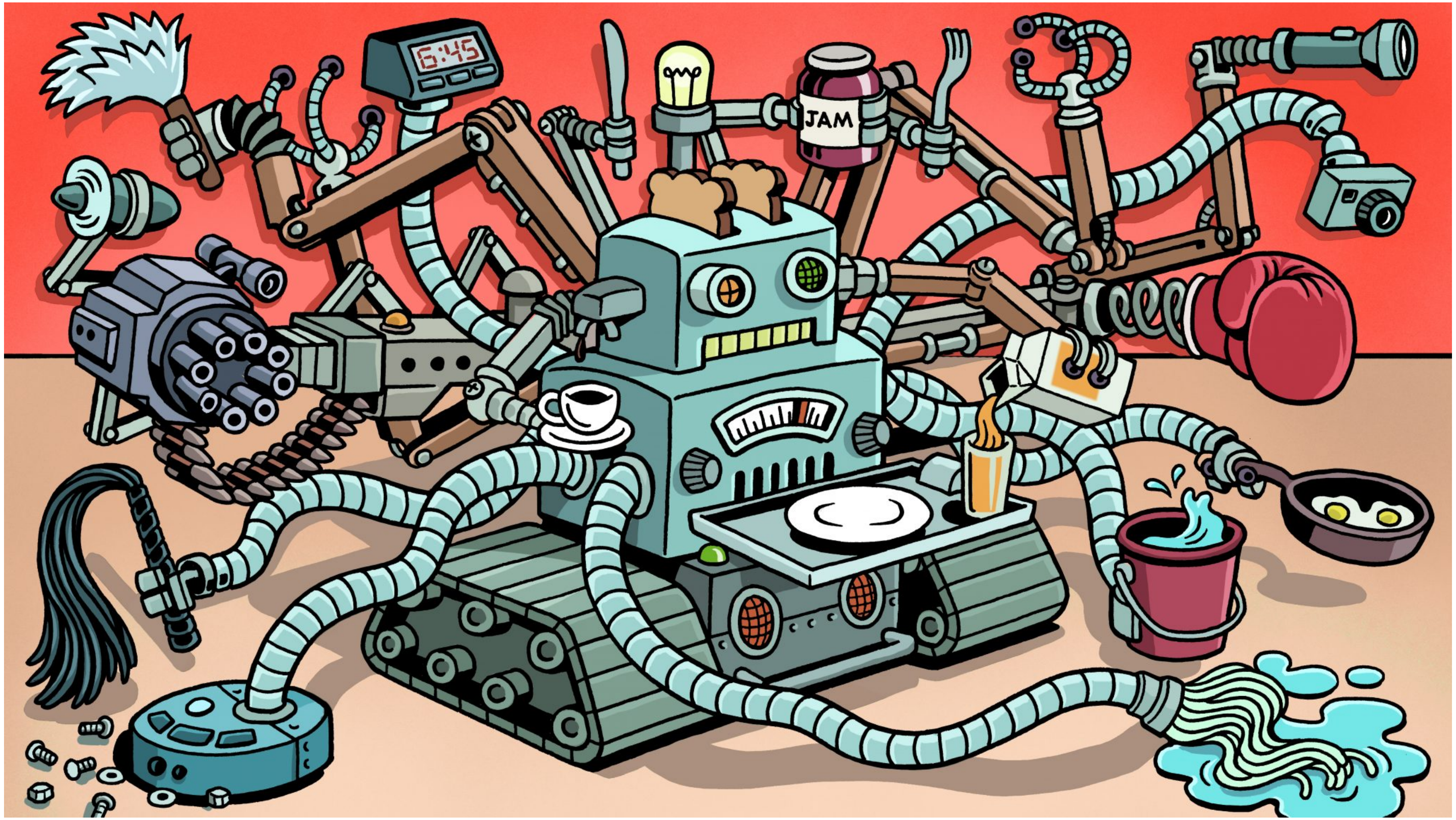
- Speech recognition, text generation, text-to-speech, spam/sarcasm/… detection, sentiment analysis, summarization, machine translation, question answering, etc.

Why?

Why do you think your people made me?
We made you cause we could.

Scott, R. (Director). (2012). *Prometheus* [Film]. Scott Free Productions, Brandywine Productions & Dune Entertainment

Really Why?



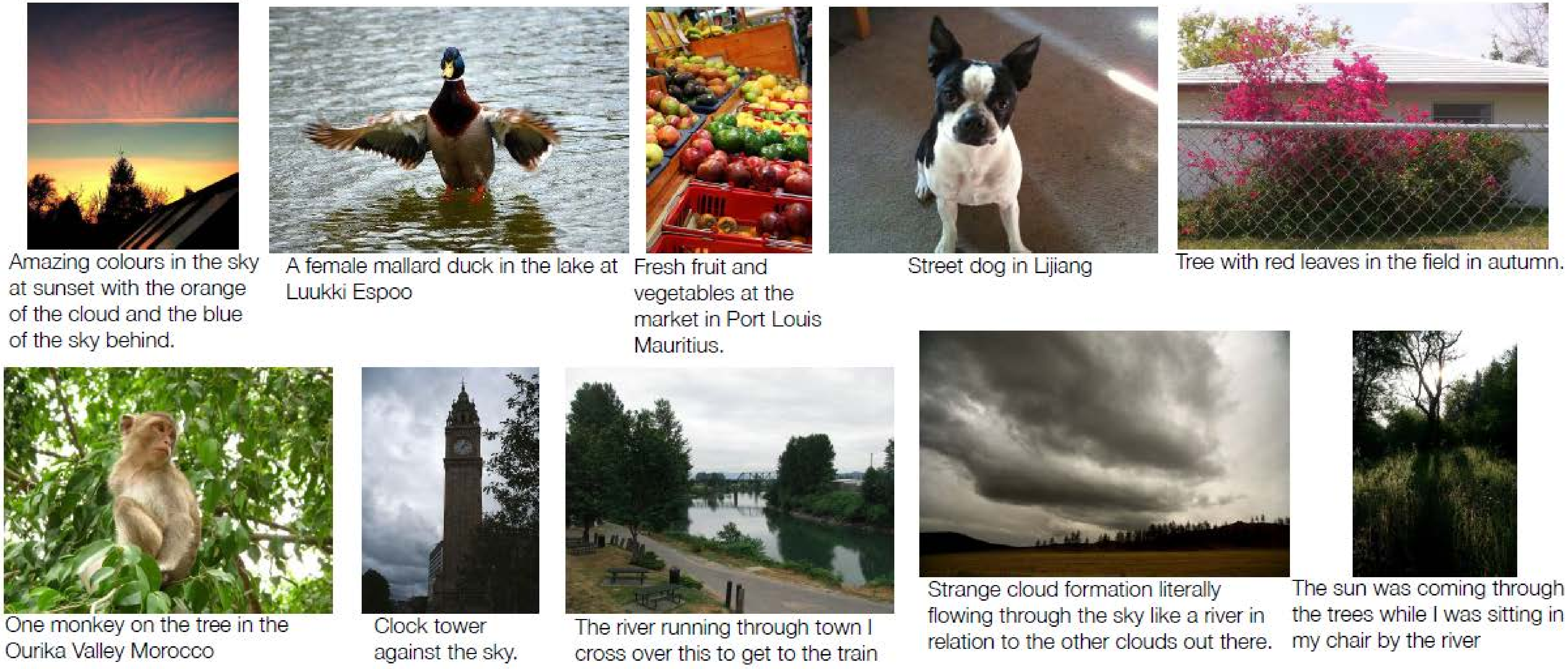
Chowdhury, A. (2017, 27 February) *10 Areas where Artificial Intelligence is going to impact our lives in Future* [Opinions]. Retrieved from https://analyticsindiamag.com/10-areas-artificial-intelligence-going-impact-lives-future/

The Future?

내 일자리 뺏으려거든 세금이나 내라 이 로봇놈들아
무능한 주제에... 미쳤습니까 휴먼?

(Got it from an acquaintance, kept it because I like it, and could not find the source)

Image Captioning: Im2Text



Amazing colours in the sky at sunset with the orange of the cloud and the blue of the sky behind.

A female mallard duck in the lake at Luukki Espoo

Fresh fruit and vegetables at the market in Port Louis Mauritius.

Street dog in Lijiang

Tree with red leaves in the field in autumn.

One monkey on the tree in the Ourika Valley Morocco

Clock tower against the sky.

The river running through town I cross over this to get to the train

Strange cloud formation literally flowing through the sky like a river in relation to the other clouds out there.

The sun was coming through the trees while I was sitting in my chair by the river

Figure 4: **Results:** Some good captions selected by our system for query images.

V. Ordonez, et al. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Proceedings of the International Conference on Neural Information Processing Systems*.

Image Captioning: DenseCap

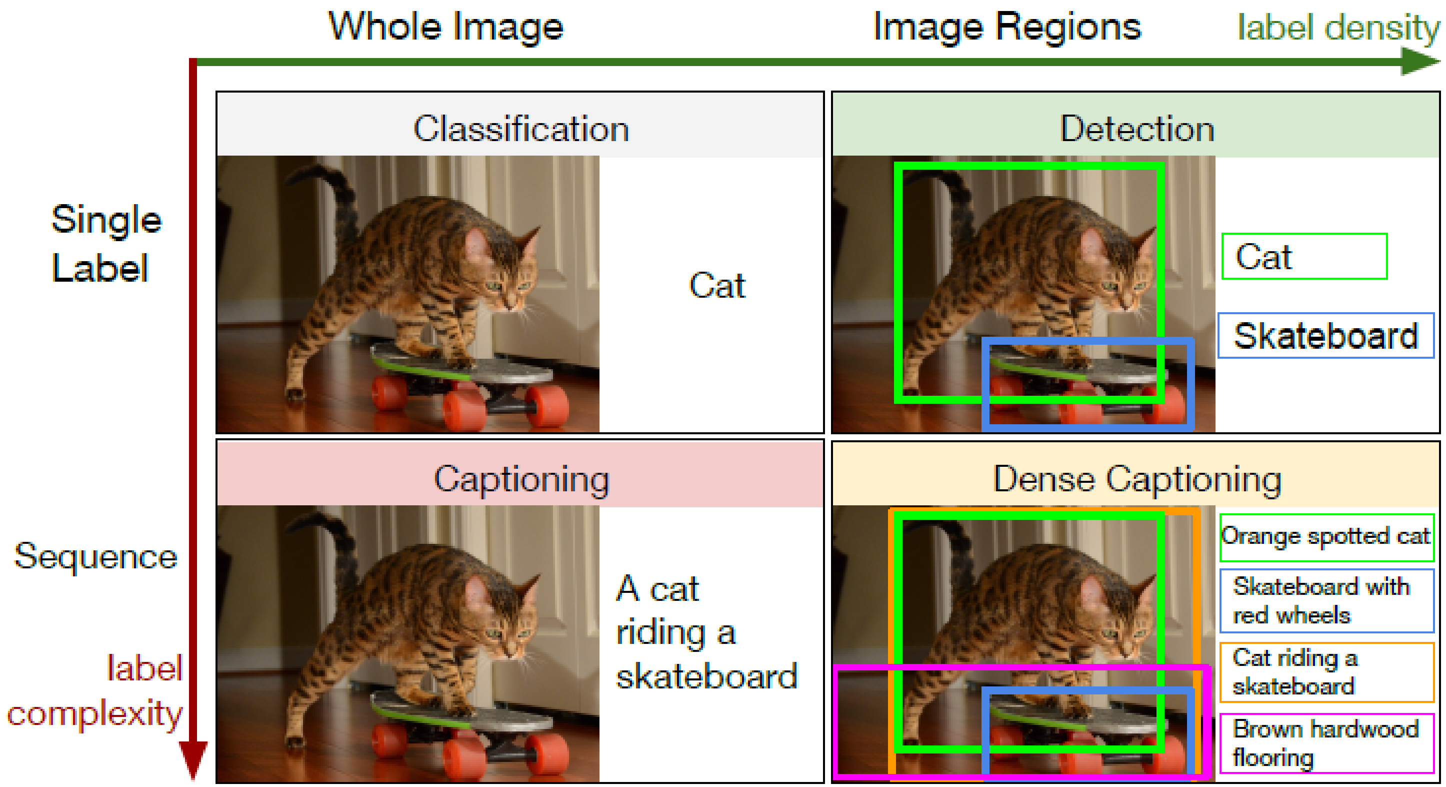
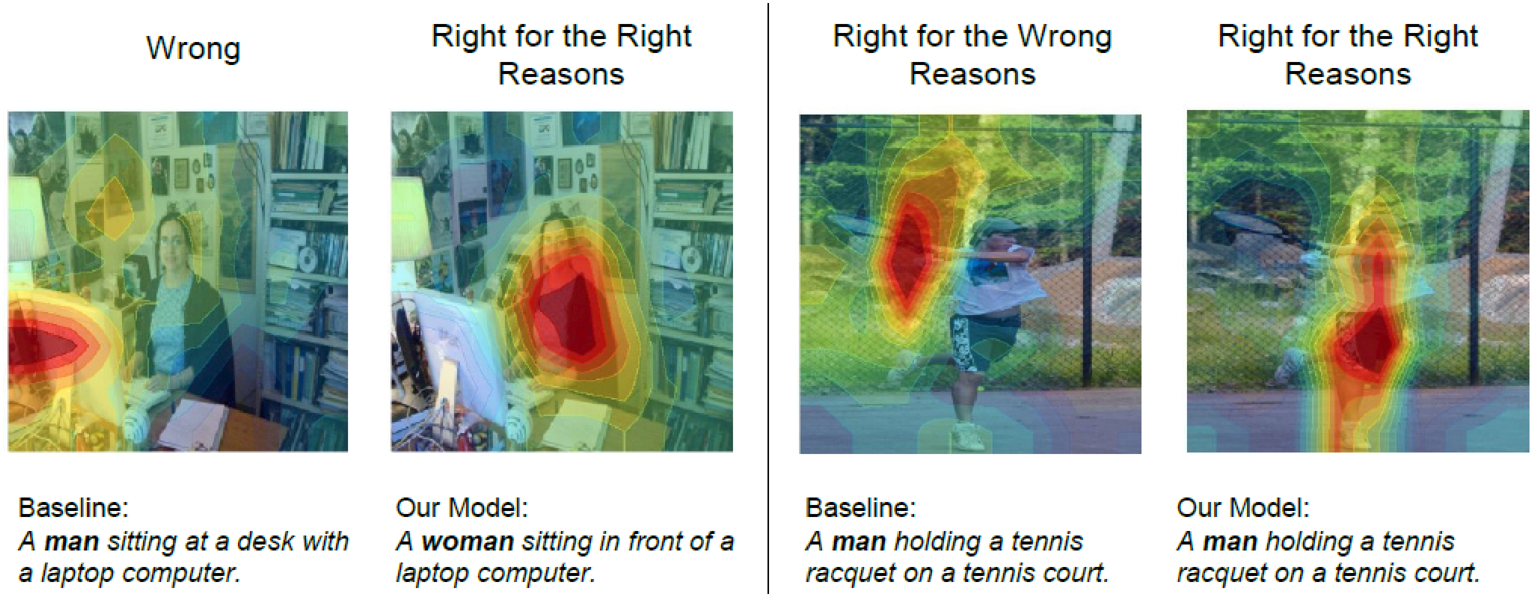


Figure 1. We address the Dense Captioning task (bottom right) with a model that jointly generates both dense and rich annotations in a single forward pass.

J. Johnson, et al. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Image Captioning: Equalizer


Wrong
Right for the Right Reasons
Right for the Wrong Reasons
Right for the Right Reasons

Baseline:
*A **man** sitting at a desk with a laptop computer.*

Our Model:
*A **woman** sitting in front of a laptop computer.*

Baseline:
*A **man** holding a tennis racquet on a tennis court.*

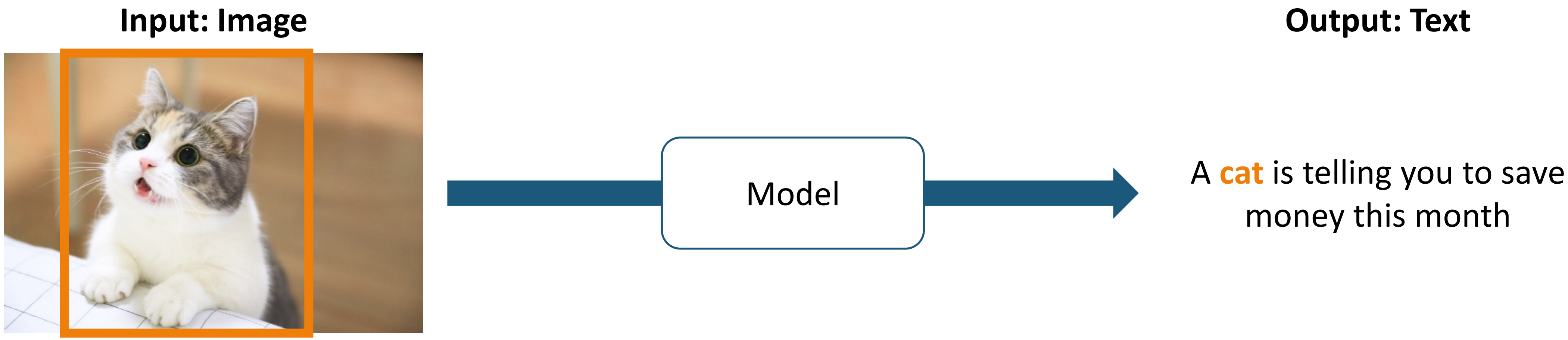
Our Model:
*A **man** holding a tennis racquet on a tennis court.*

Fig. 1: Examples where our proposed model (Equalizer) corrects bias in image captions. The overlaid heatmap indicates which image regions are most important for predicting the gender word. On the left, the baseline predicts gender incorrectly, presumably because it looks at the laptop (not the person). On the right, the baseline predicts the gender correctly but it does not look at the person when predicting gender and is thus not acceptable. In contrast, our model predicts the correct gender word and correctly considers the person when predicting gender.

K. Burns, et al. 2018. Women also Snowboard: Overcoming Bias in Captioning Models. In *Proceedings of the European Conference on Computer Vision*.

Image Captioning 101

- Automatically generating textual description of a digital image
- Multi-modal, Attention, Multiple Object Detection, Action Recognition, Text Generation, …



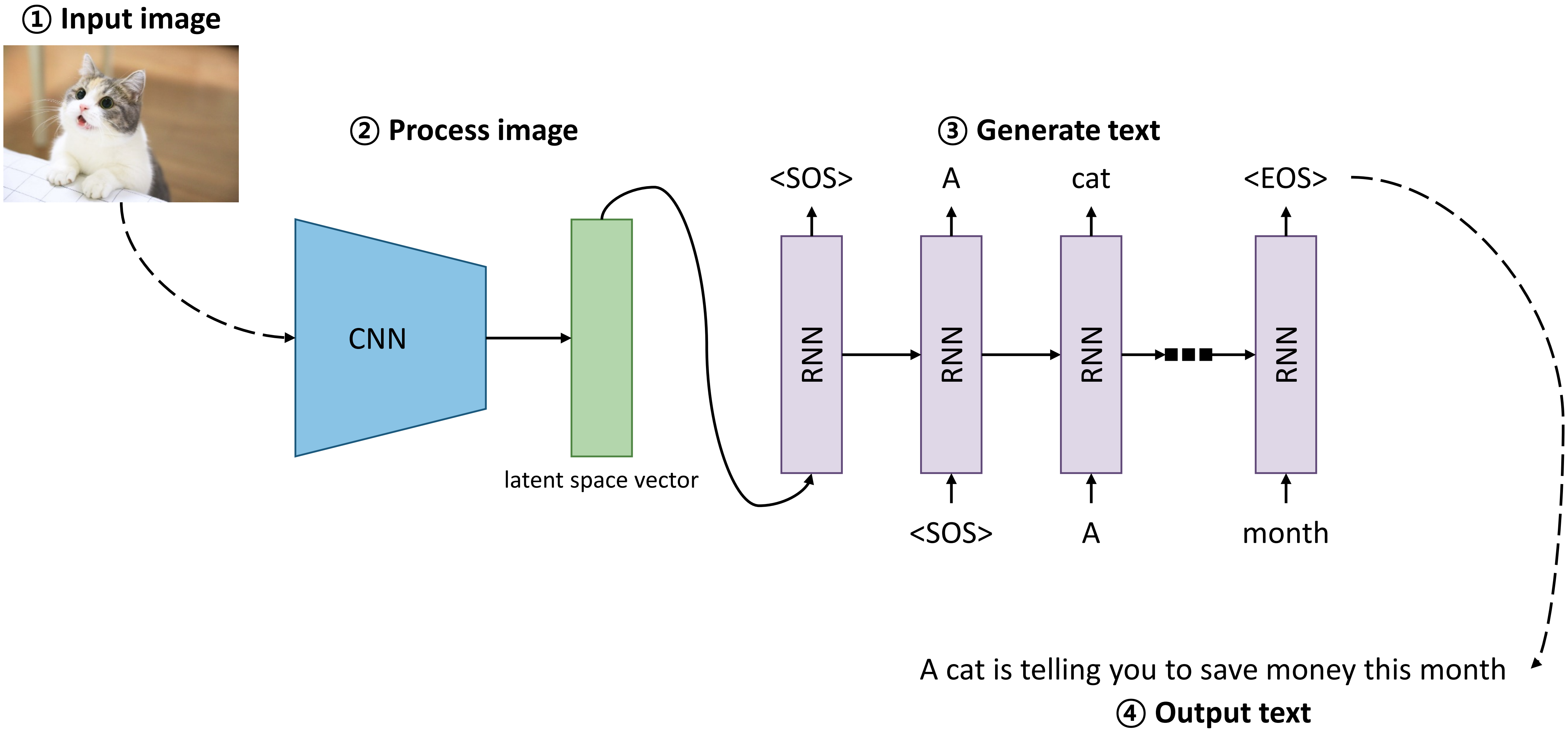
Input: Image

Model

Output: Text

A **cat** is telling you to save money this month

Image Captioning Steps



① Input image

② Process image

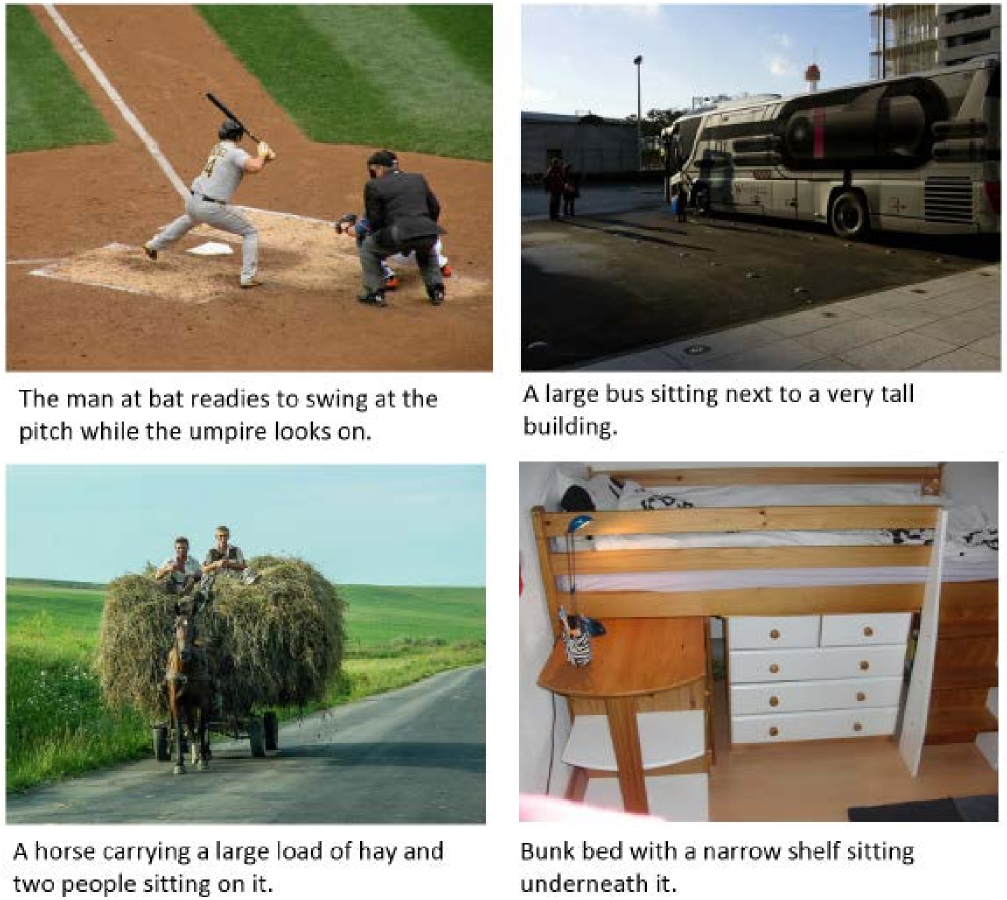
③ Generate text

A cat is telling you to save money this month

④ Output text

COCO Captions

- 413,915 captions for 82,783 images in training
- 202,520 captions for 40,504 images in validation
- 379,249 captions for 40,775 images in testing
- GitHub:
  - https://github.com/tylin/coco-caption.git



The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Fig. 1: Example images and captions from the Microsoft COCO Caption dataset.

X. Chen, et al. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv preprint arXiv:1504.00325.

Evaluation Metrics

- BLEU
  - K. Papineni, et al. “BLEU: A Method for Automatic Evaluation of Machine Translation,” in *ACL*, 2002.
- METEOR
  - M. Denkowski and A. Lavie. “Meteor Universal: Language Specific Translation Evaluation for Any Target Language,” In *EACL Workshop on Statistical Machine Translation*, 2014.
- ROUGE-L
  - C. Lin. “ROUGE: A Package for Automatic Evaluation of Summaries,” In *ACL Workshop*, 2004.
- CIDEr
  - R. Vedantam, et al. “CIDEr: Consensus-based Image Description Evaluation,” In *CVPR*, 2015.
- SPICE
  - P. Anderson, et al. “SPICE: Semantic Propositional Image Caption Evaluation,” In *ECCV*, 2016.

One Step Further

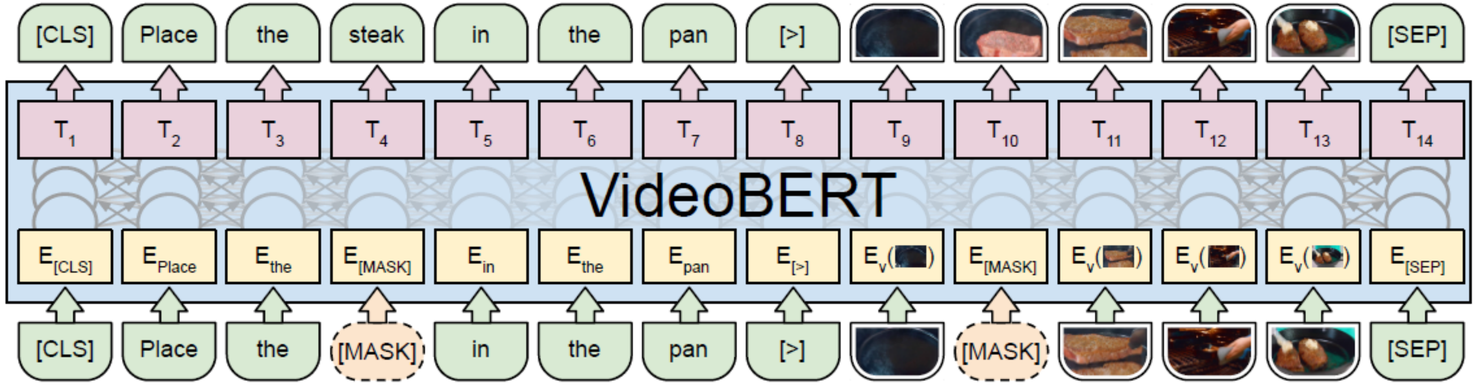
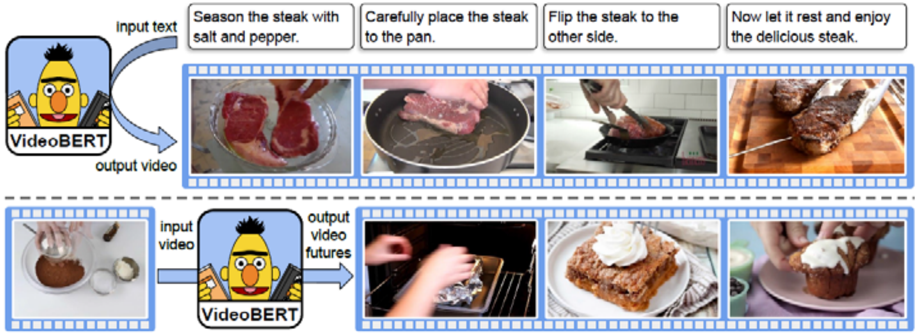
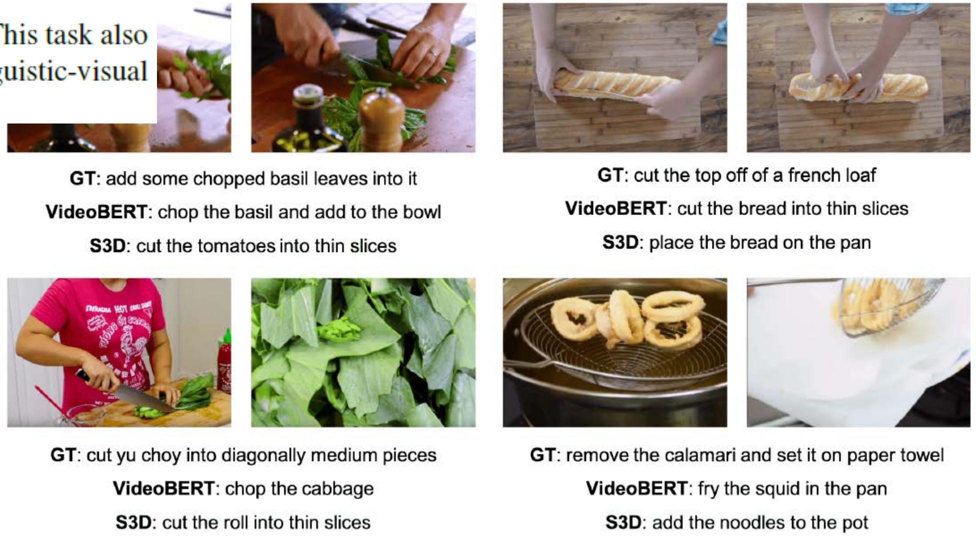


Figure 3: Illustration of VideoBERT in the context of a video and text masked token prediction, or *cloze*, task. This task also allows for training with text-only and video-only data, and VideoBERT can furthermore be trained using a linguistic-visual alignment classification objective (not shown here, see text for details).



input text
Season the steak with salt and pepper.
Carefully place the steak to the pan.
Flip the steak to the other side.
Now let it rest and enjoy the delicious steak.
VideoBERT
output video
input video
VideoBERT
output video futures

Figure 1: **VideoBERT text-to-video generation and future forecasting.** (Above) Given some recipe text divided into sentences, $y = y_{1:T}$, we generate a sequence of video tokens $x = x_{1:T}$ by computing $x_t^* = \arg\max_k p(x_t = k|y)$ using VideoBERT. (Below) Given a video token, we show the top three future tokens forecasted by VideoBERT at different time scales. In this case, VideoBERT predicts that a bowl of flour and cocoa powder may be baked in an oven, and may become a brownie or cupcake. We visualize video tokens using the images from the training set closest to centroids in feature space.



GT: add some chopped basil leaves into it
VideoBERT: chop the basil and add to the bowl
S3D: cut the tomatoes into thin slices

GT: cut the top off of a french loaf
VideoBERT: cut the bread into thin slices
S3D: place the bread on the pan

GT: cut yu choy into diagonally medium pieces
VideoBERT: chop the cabbage
S3D: cut the roll into thin slices

GT: remove the calamari and set it on paper towel
VideoBERT: fry the squid in the pan
S3D: add the noodles to the pot

Figure 6: Examples of generated captions by VideoBERT and the S3D baseline. In the last example, VideoBERT fails to exploit the full temporal context, since it misses the paper towel frame.

C. Sun, et al. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *Proceedings of the International Conference on Computer Vision*.