Tutorial on Recent Advances in Visual Captioning

Luowei Zhou

06/15/2020

Outline

- Problem Overview
- Visual Captioning Taxonomy
- Image Captioning
- Datasets and Evaluation
- Video Description
- Grounded Caption Generation
- Dense Caption Generation
- Conclusion
- Q&A

Problem Overview

- Visual Captioning – Describe the content of an image or video with a natural language sentence.

A cat is sitting next to a pine tree, looking up.

A dog is playing piano with a girl.

Applications of Visual Captioning

- Alt-text generation (from PowerPoint)
- Content-based image retrieval (CBIR)
- Or just for fun!

A fun video running visual captioning model real-time made by Kyle McDonald. Source: https://vimeo.com/146492001

Visual Captioning Taxonomy

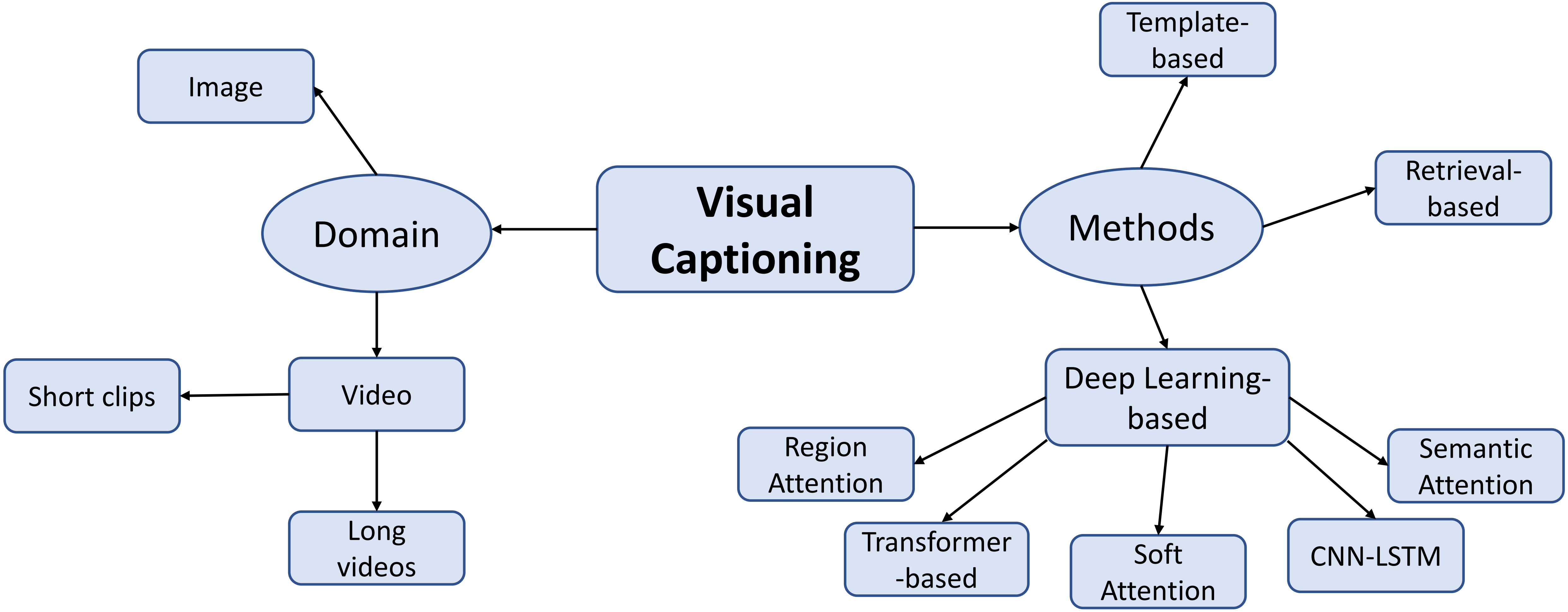


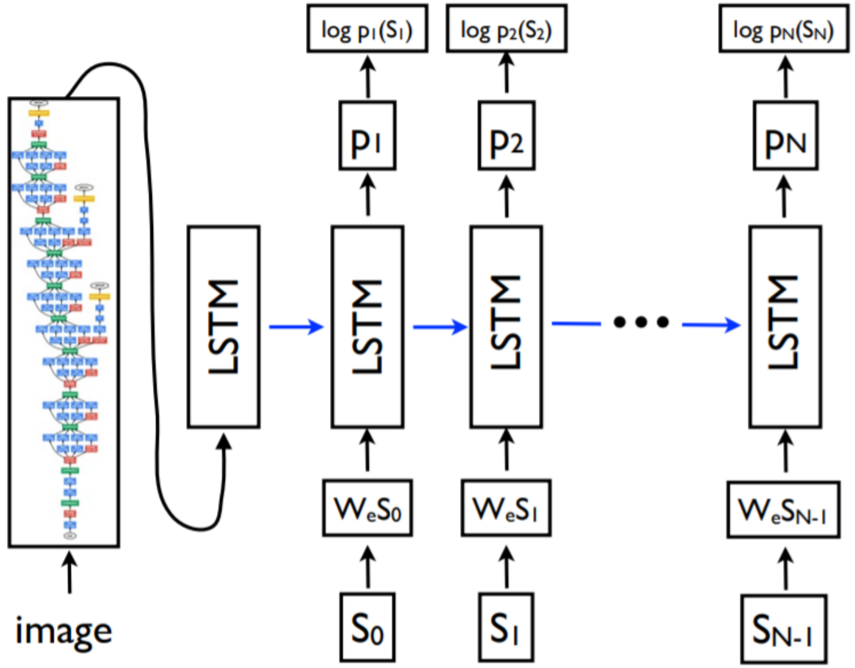
Image Captioning with CNN-LSTM

- Problem Formulation

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

"Show and Tell"



- The Encoder-Decoder framework

→ Visual Encoder → Language Decoder → "Cat sitting outside"

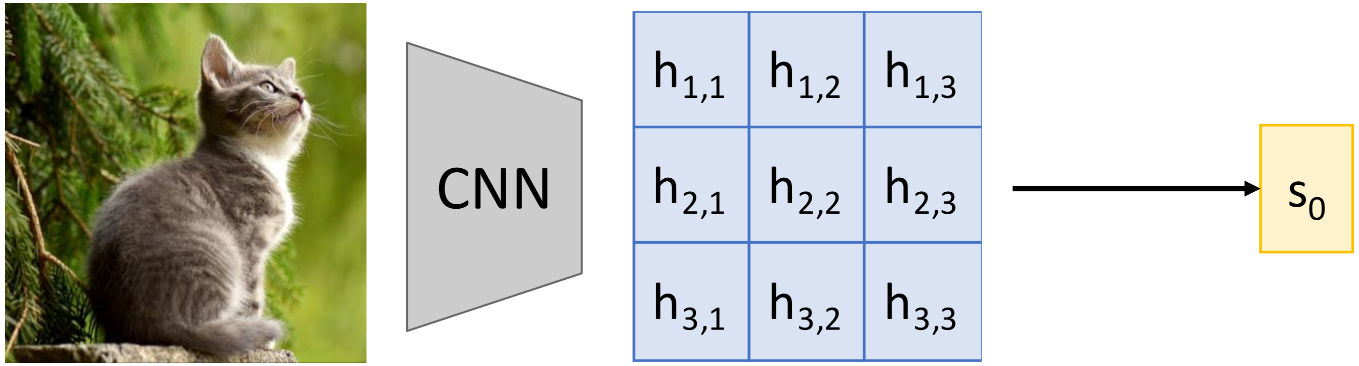
Image credit: Vinyals et al. "Show and Tell: A Neural Image Caption Generator", CVPR 2015.

Image Captioning with Soft Attention

- Soft Attention – Dynamically attend to input content based on query.
- Basic elements: query – q, keys - K, and values – V
- In our case, keys and values are usually identical. They come from the CNN activation map.
- Query q is determined by the global image feature or LSTM's hidden states.

Bahdanau et al. "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015.
Xu et al. "Show, Attend and Tell", ICML 2015.

Image Captioning with Soft Attention

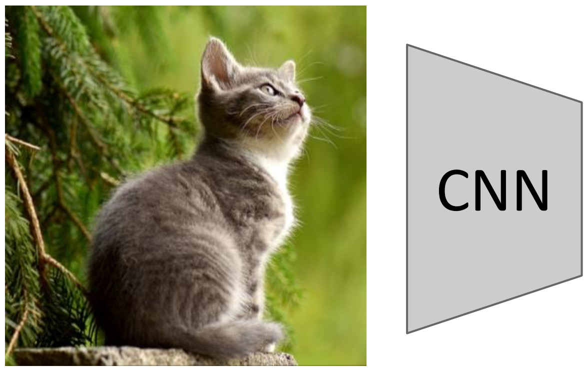


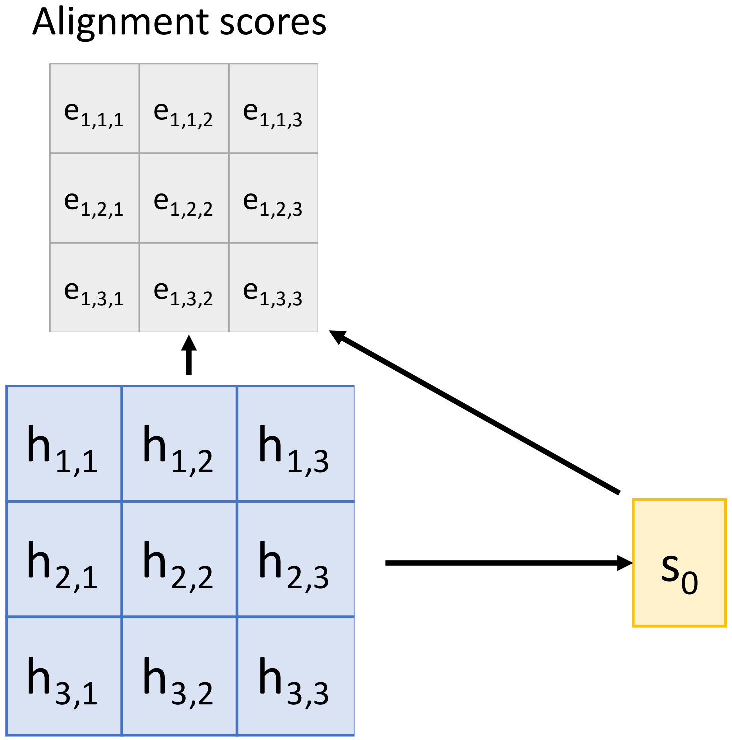
Use a CNN to compute a
grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$



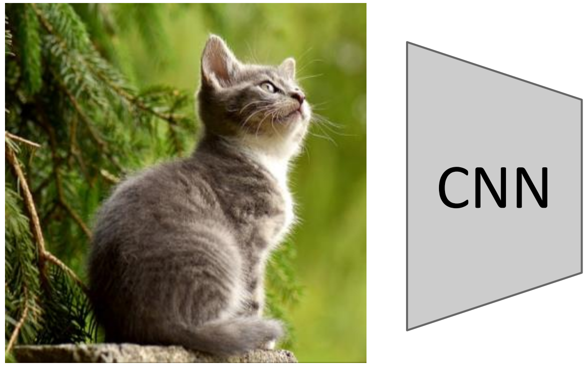


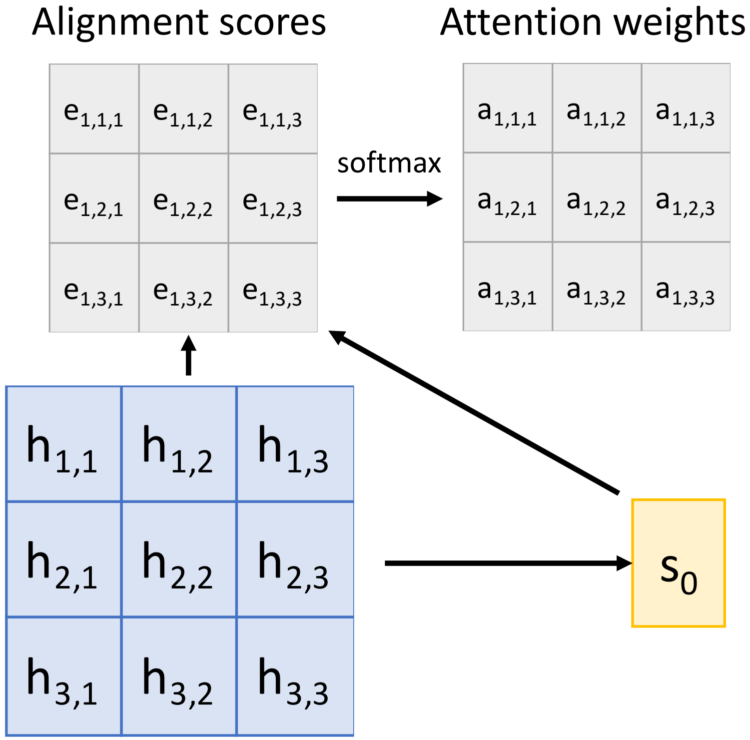
Use a CNN to compute a
grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$



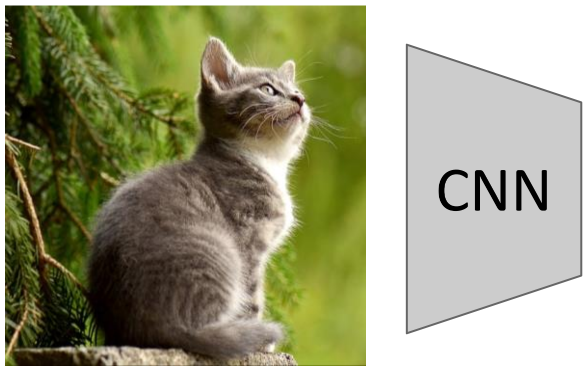


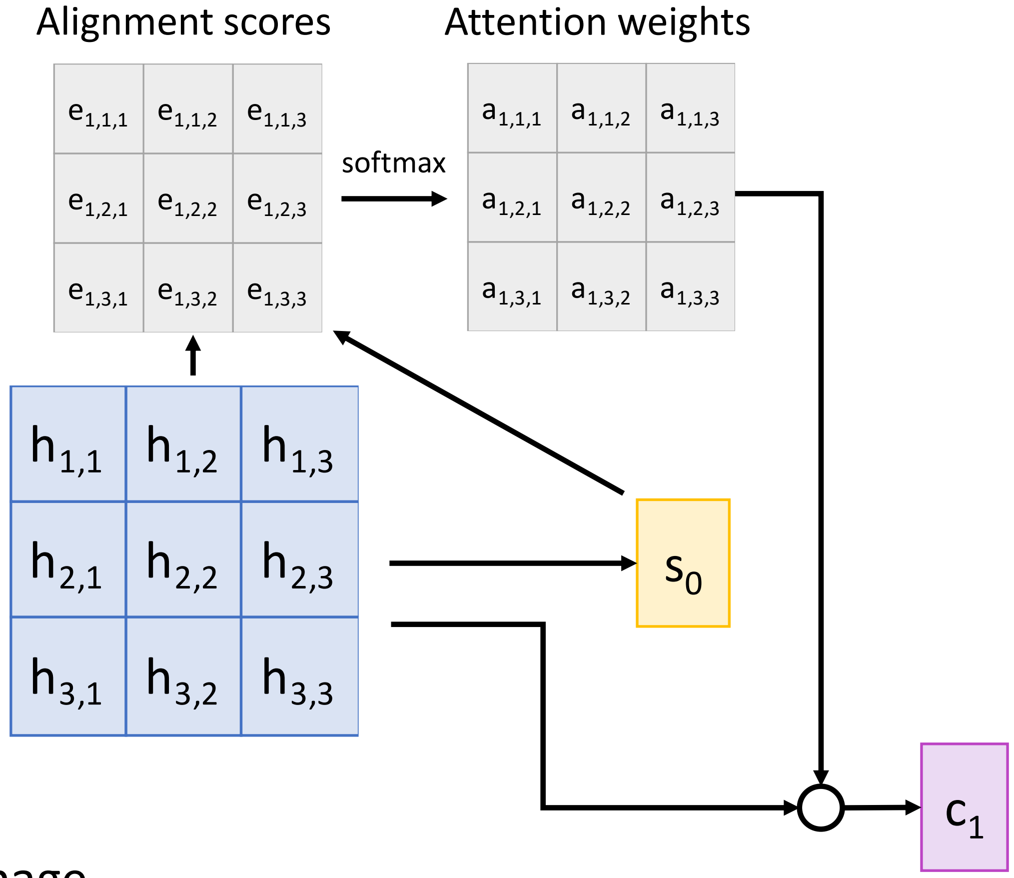
Use a CNN to compute a
grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$



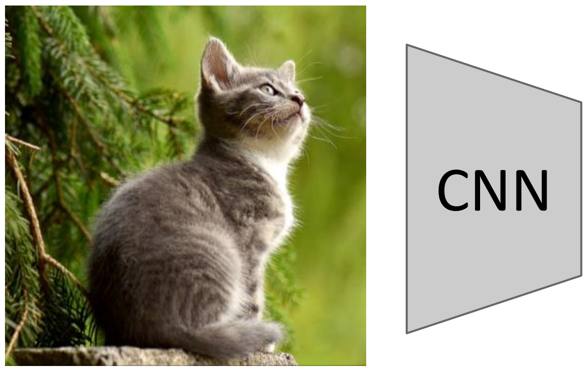


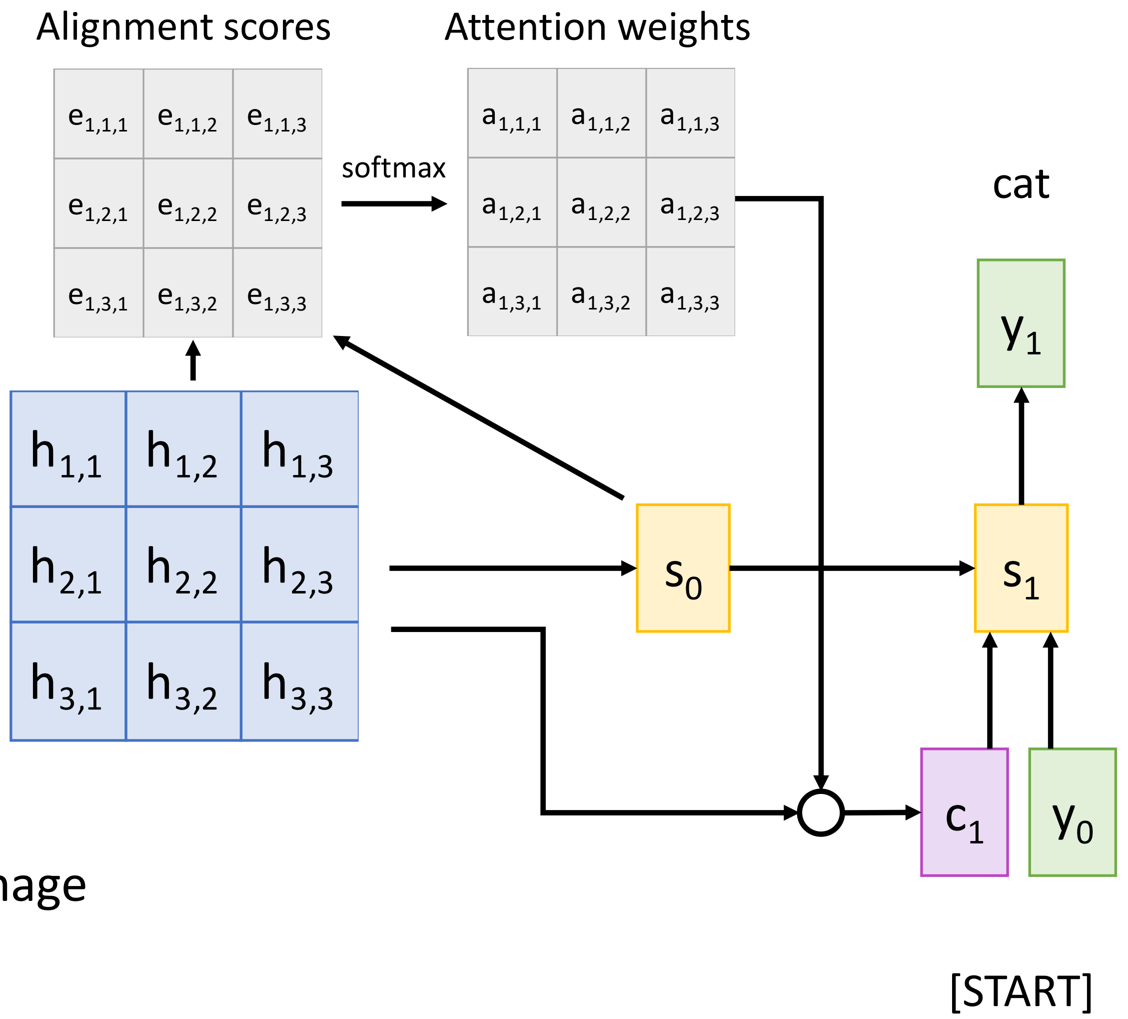
Use a CNN to compute a
grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$



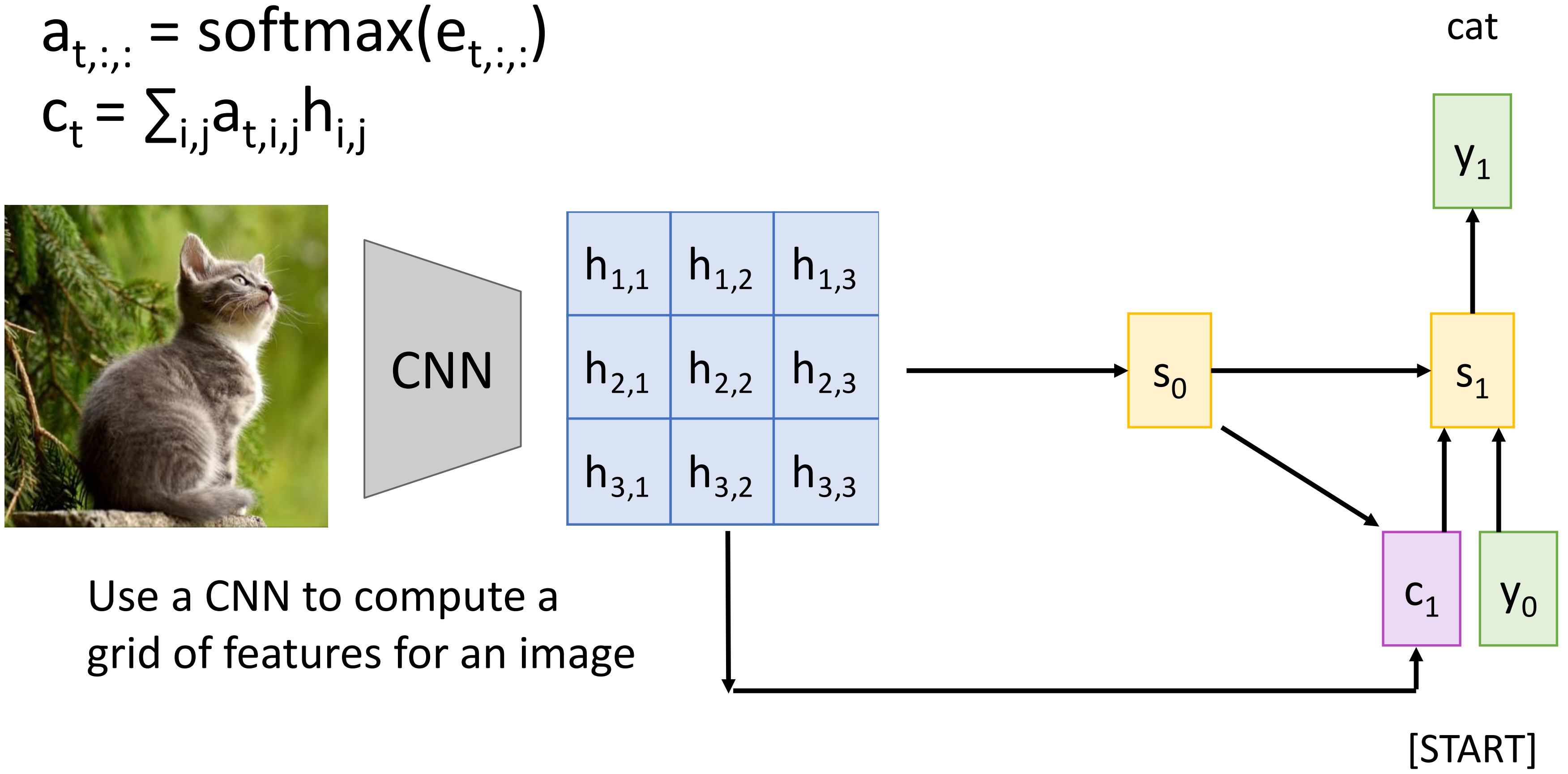


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

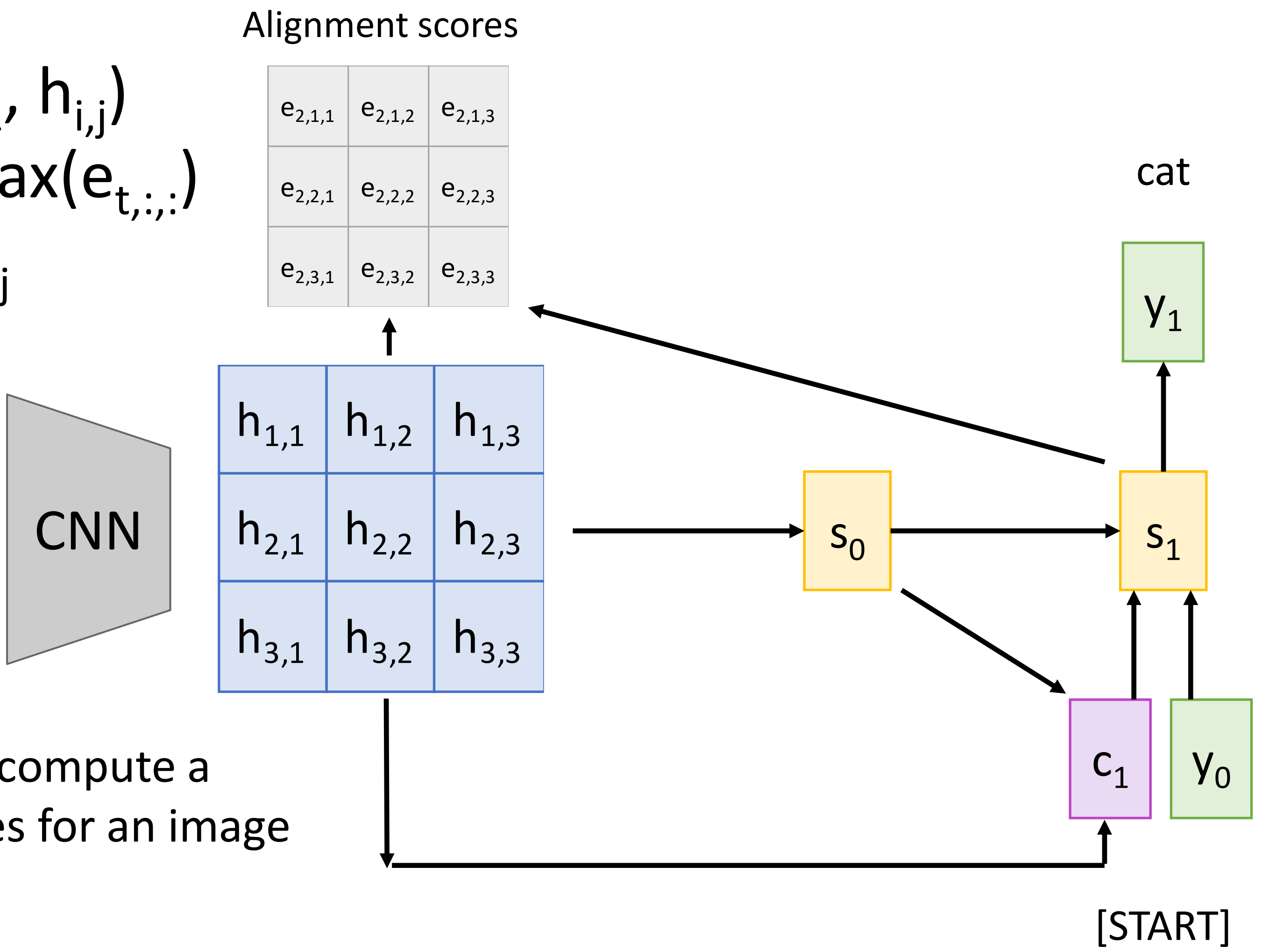


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: "Show, Attend and Tell" by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$

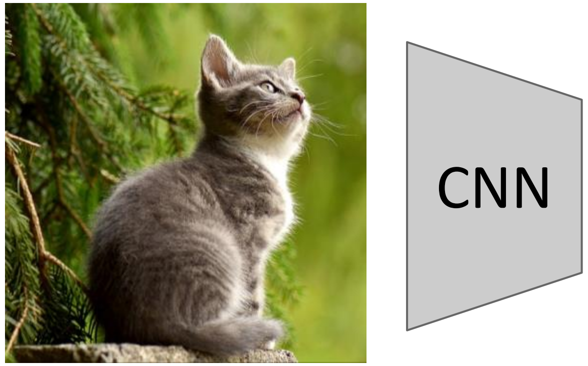


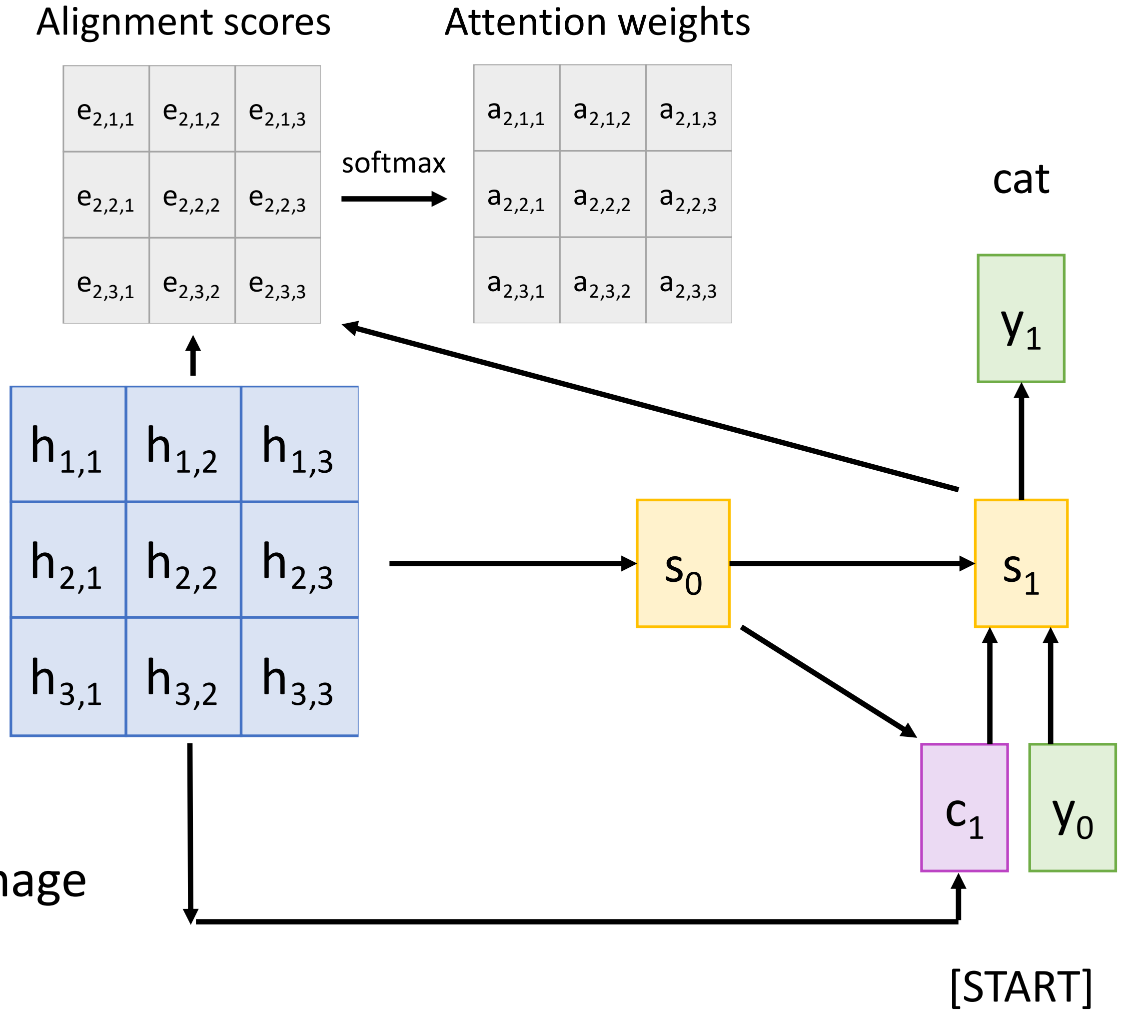
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \mathrm{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$



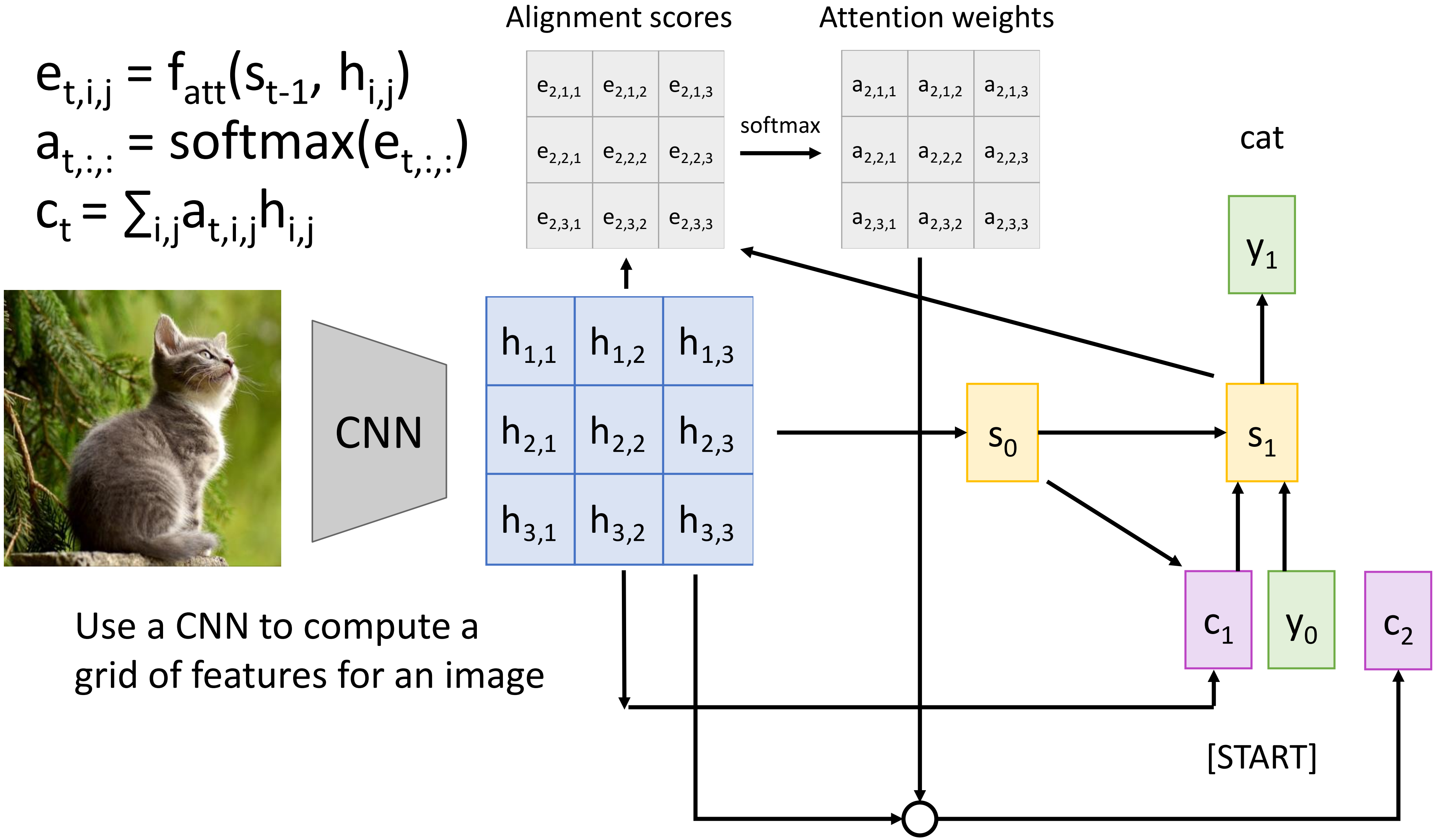


Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \text{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$

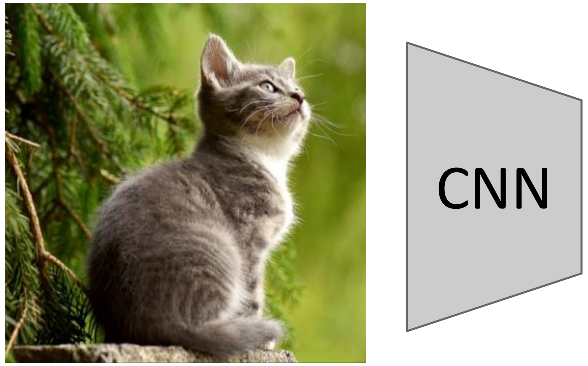


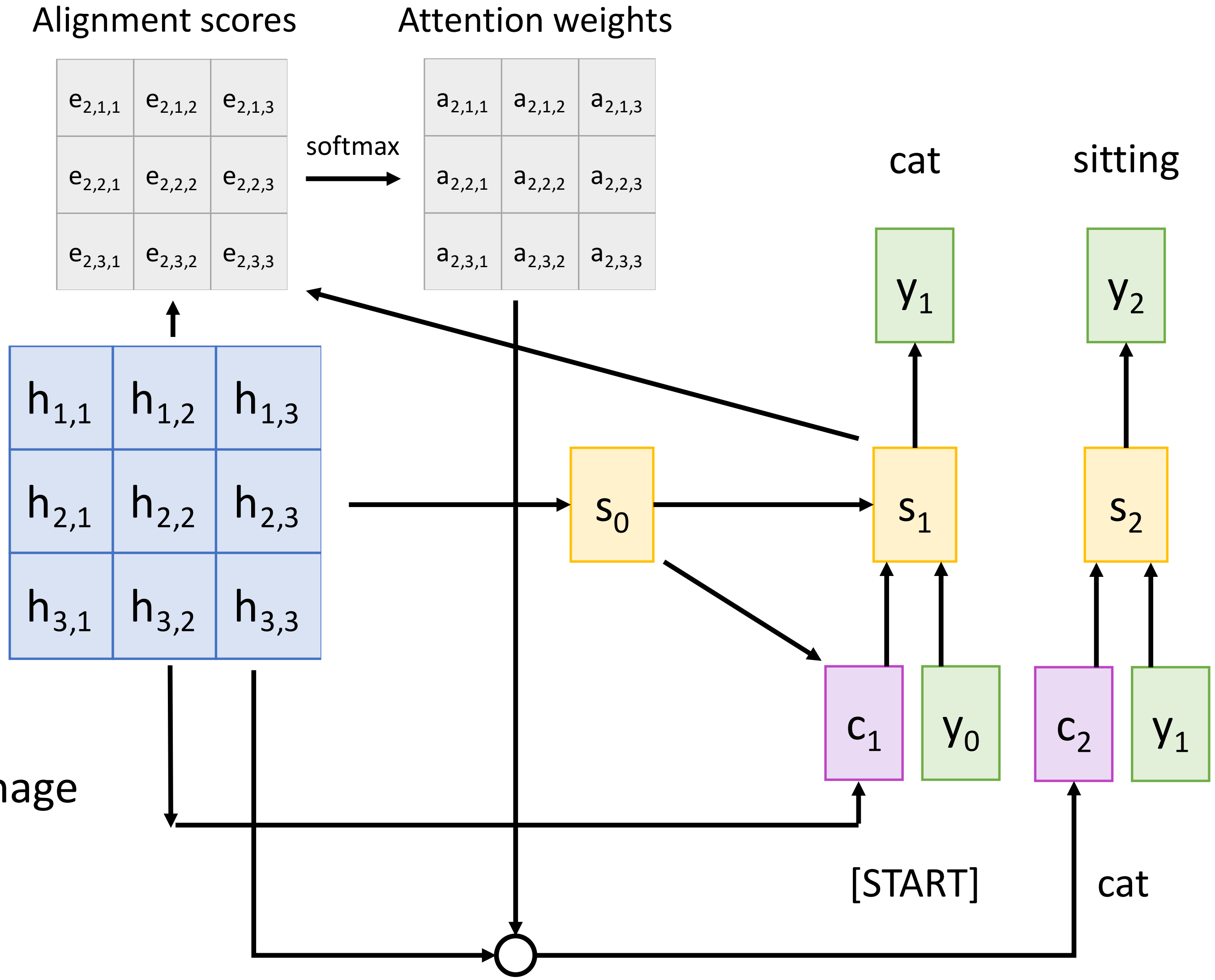
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$
$a_{t,:,:} = \text{softmax}(e_{t,:,:})$
$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$





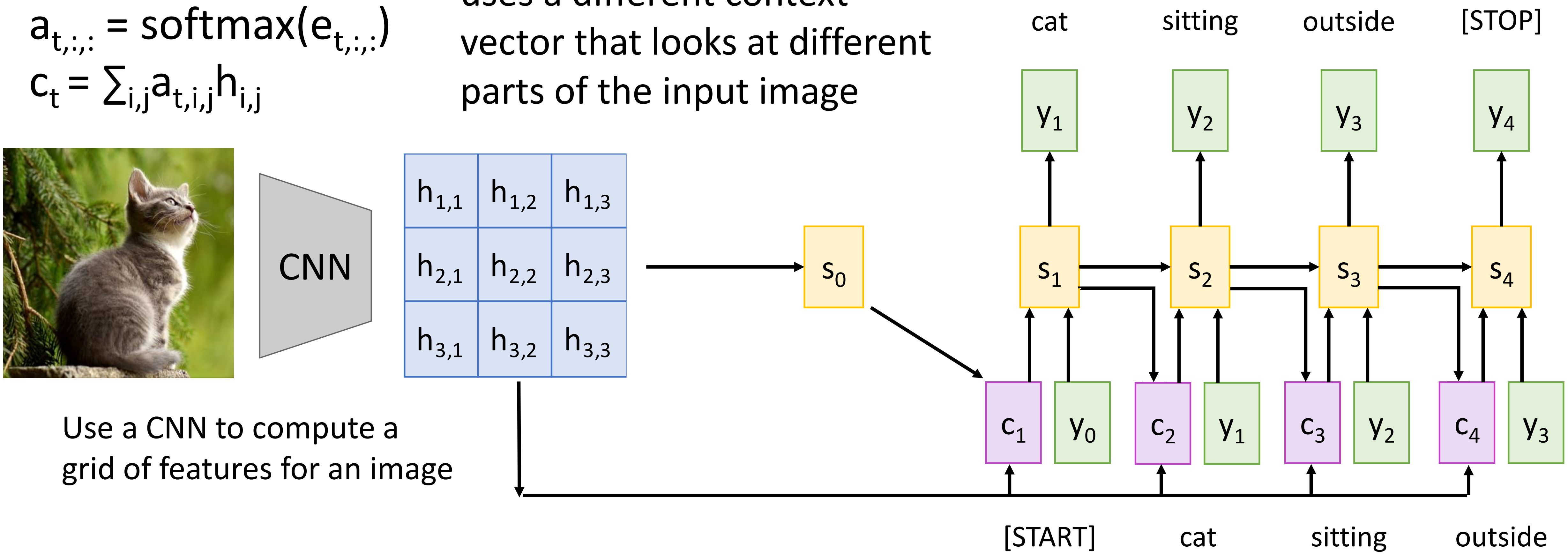
Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

$$e_{t,i,j} = f_{att}(s_{t-1}, h_{i,j})$$
$$a_{t,:,:} = \text{softmax}(e_{t,:,:})$$
$$c_t = \sum_{i,j} a_{t,i,j} h_{i,j}$$

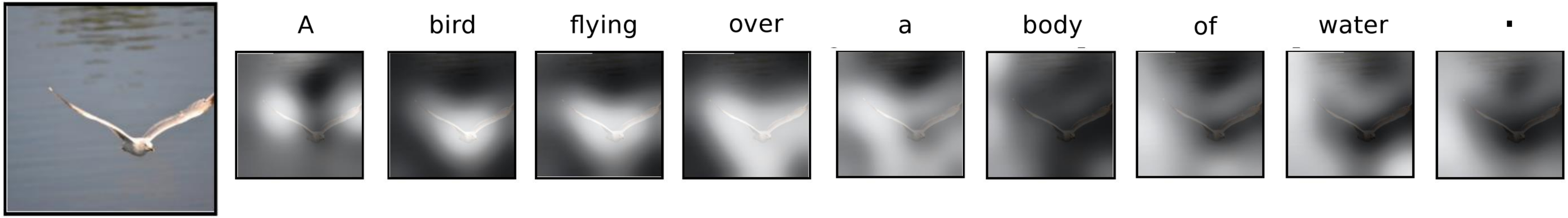
Each timestep of decoder uses a different context vector that looks at different parts of the input image



Use a CNN to compute a grid of features for an image

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention



Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Soft Attention

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Slide credit: UMich EECS 498/598 DeepVision course by Justin Johnson. Method: “Show, Attend and Tell” by Xu et al. ICML 2015.

Image Captioning with Region Attention

- Variants of Soft Attention based on the feature input
  - Grid activation features (covered)
  - Region proposal features

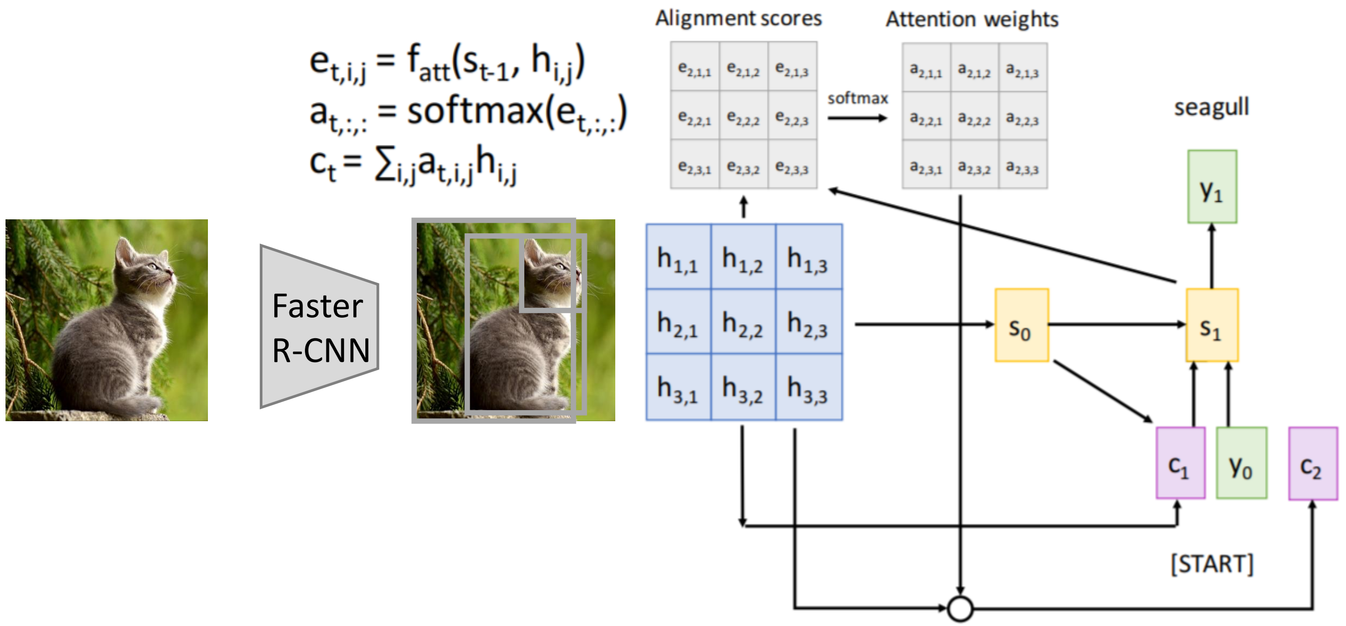
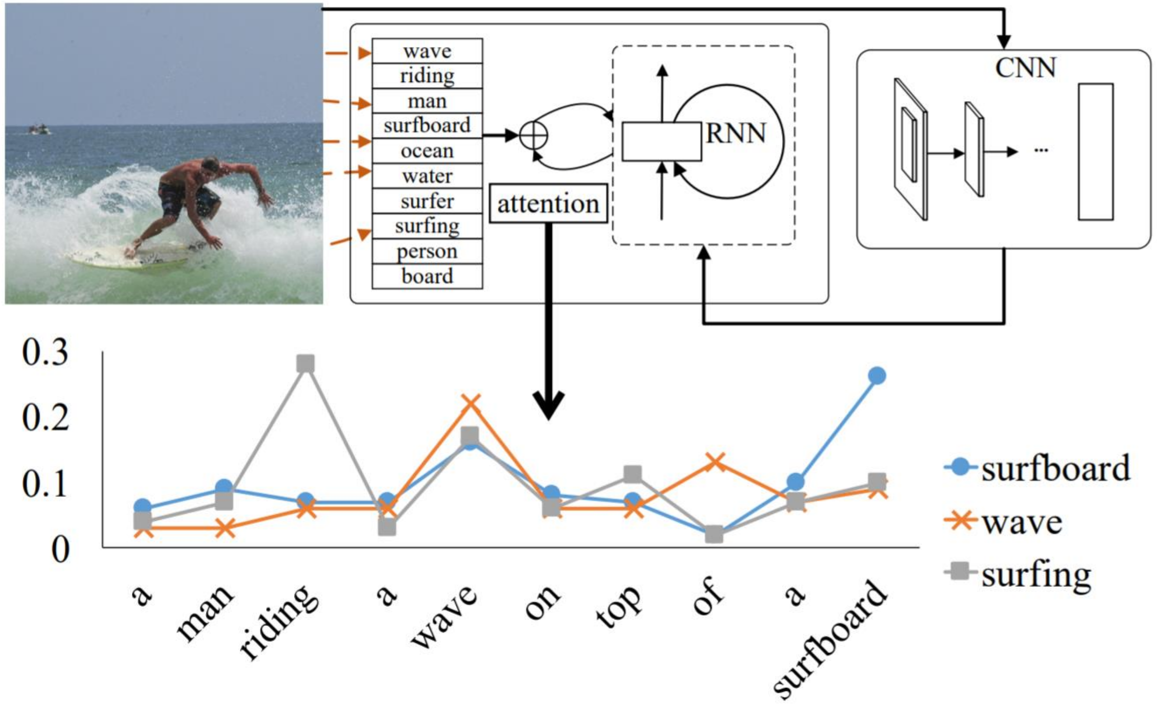


Image Captioning with "Fancier" Attention

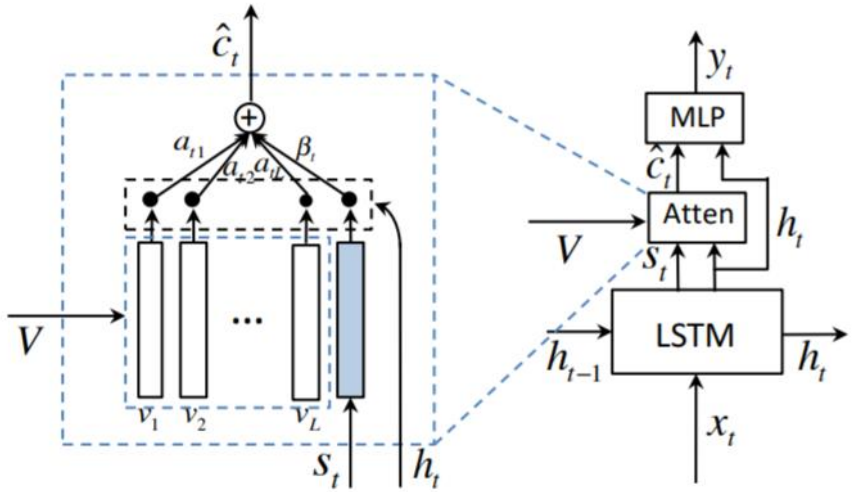
Semantic attention

- Visual attributes



Adaptive Attention

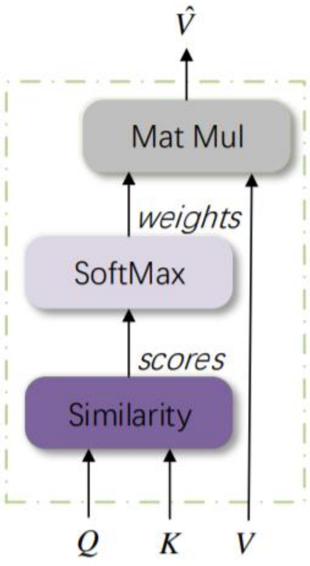
- Knowing when to & not to attend to the image



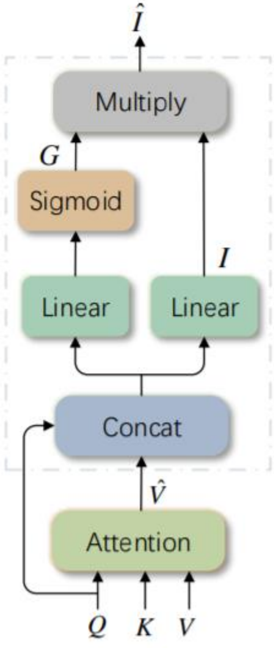
You et al. "Image captioning with semantic attention", CVPR 2016. Yao et al. "Boosting Image Captioning with Attributes", ICCV 2017.
Lu et al. "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning", CVPR 2017.

Image Captioning with "Fancier" Attention

Attention on Attention



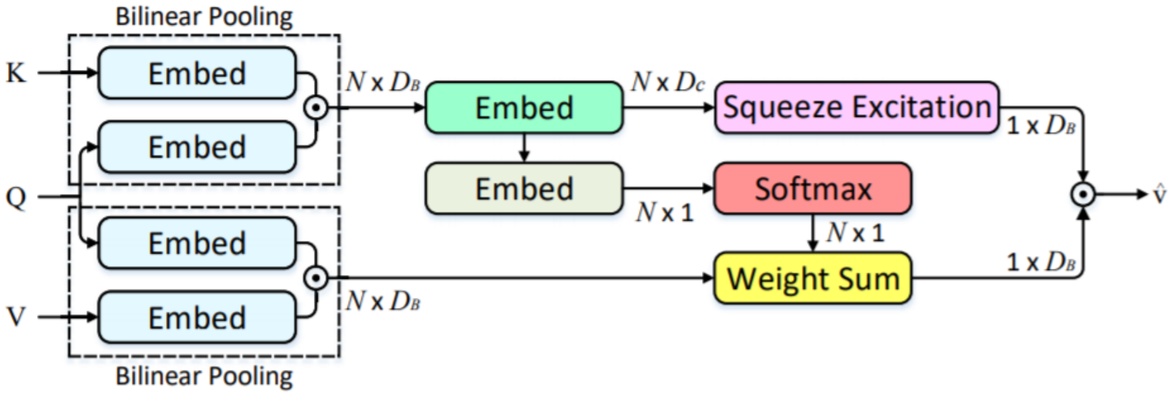
(a) Attention



(b) Attention on Attention

X-Linear Attention

- Spatial and channel-wise bilinear attention

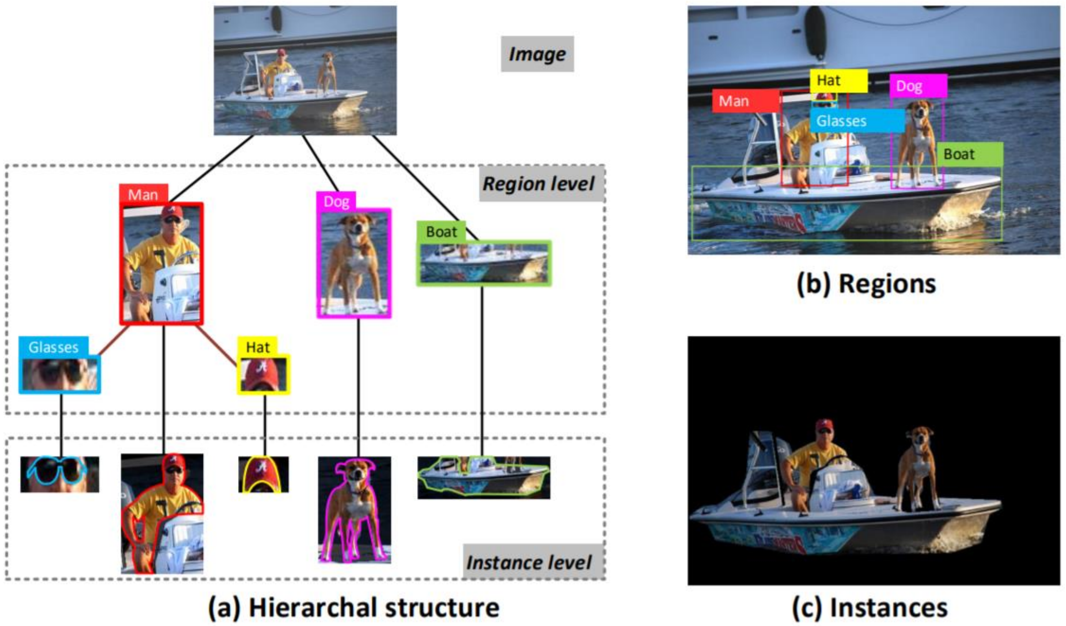


Huang et al. "Attention on Attention for Image Captioning", ICCV 2019.
Pan et al. "X-Linear Attention Networks for Image Captioning", CVPR 2020.

Image Captioning with “Fancier” Attention

Hierarchy Parsing and GCNs

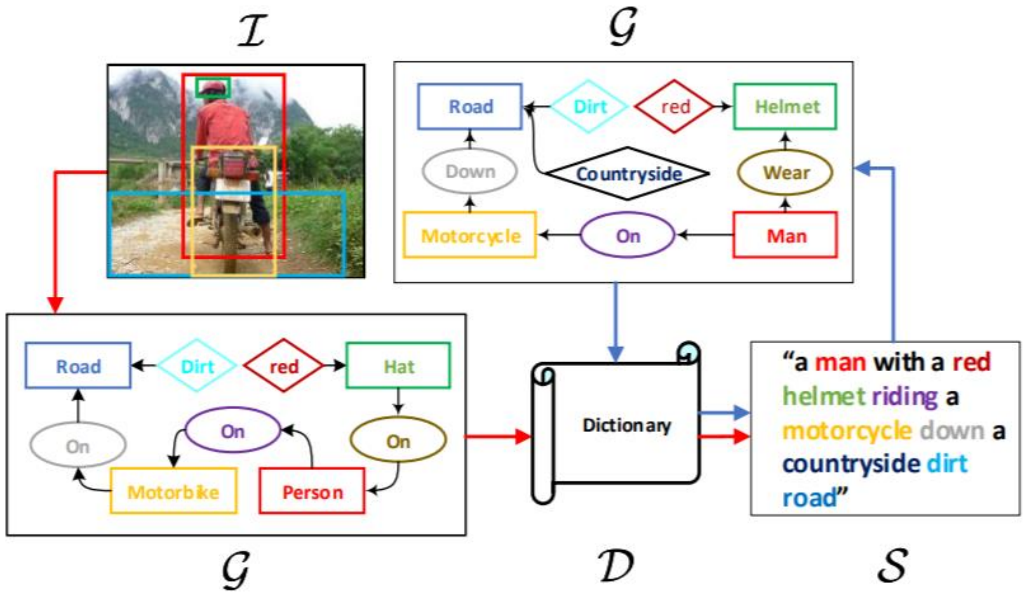
- Hierarchal tree structure in image



(a) Hierarchal structure

Auto-Encoding Scene Graphs

- Scene Graphs in image and text



Yao et al. “Hierarchy Parsing for Image Captioning”, ICCV 2019.
Yang et al. “Auto-encoding scene graphs for image captioning”, CVPR 2019.

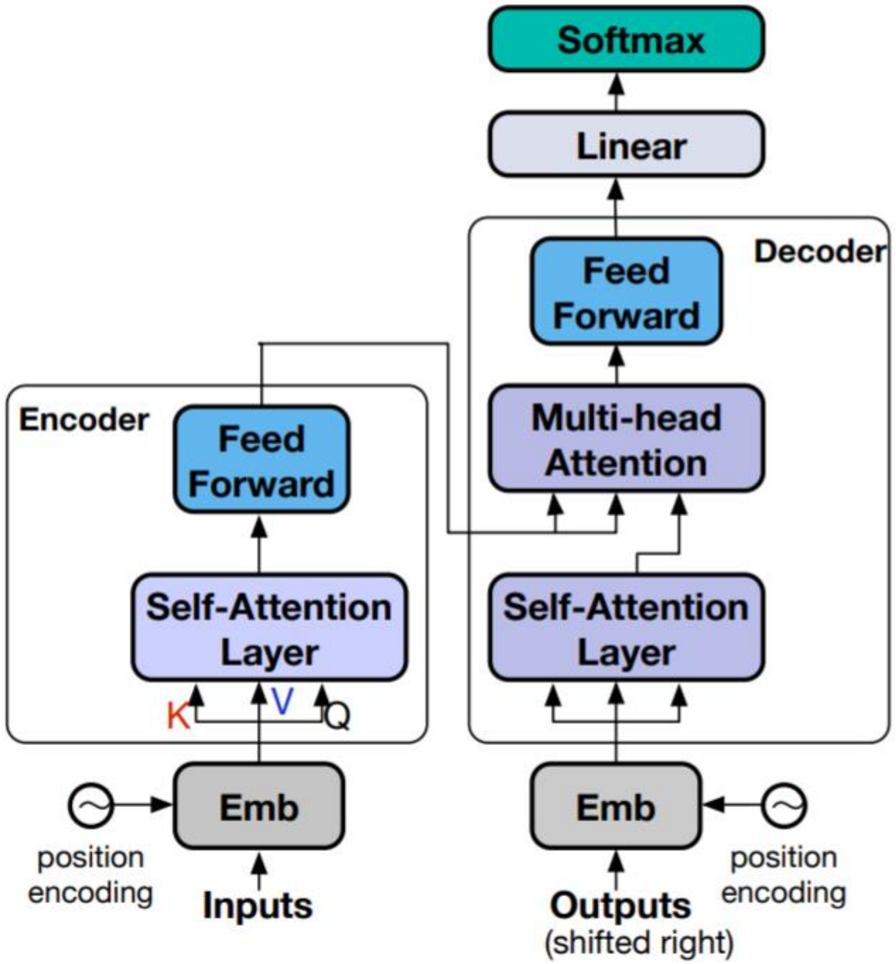
Image Captioning with Transformer

- Transformer performs sequence-to-sequence generation.
- Self-Attention – A type of soft attention that "attends to itself".
- Self-Attention is a special case of Graph Neural Networks (GNNs) that has a fully-connected graph.
- Self-attention is sometimes used to model relationship between object regions, similar to GCNs.

Vaswani et al. "Attention is all you need", NIPS 2017.
Yao et al. "Exploring visual relationship for image captioning", ECCV 2018.
Further readings: https://graphdeeplearning.github.io/post/transformers-are-gnns/

Image Captioning with Transformer

- Transformer is first adapted for captioning in Zhou et al.
- Others: Object Relation Transformer, Meshed-Memory Transformer



Zhou et al. “End-to-end dense video captioning with masked transformer”, CVPR 2018.
Herdade et al. “Image Captioning: Transforming Objects into Words”, NeurIPS 2019.
Cornia et al. “Meshed-Memory Transformer for Image Captioning”, CVPR 2020.

Vision-Language Pre-training (VLP)

- Two-stage training strategy: **pre-training** and **fine-tuning**.
- **Pre-training** is performed on a large dataset. Usually with auto-generated captions. The training objective is *unsupervised*.
- **Fine-tuning** is task-specific *supervised* training on downstream tasks.
- All methods are based on BERT (a variant of Transformer).

Zhou et al. "Unified vision-language pre-training for image captioning and vqa", AAAI 2020.
Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL 2019.

Vision-Language Pre-training (VLP)

Separate Encoder-Decoder

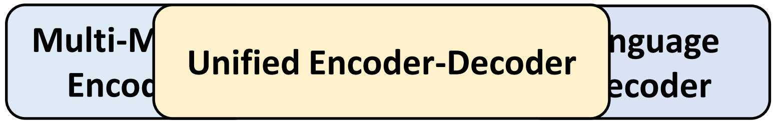
- Methods: VideoBERT and Oscar

- Only the encoder is pre-trained

Unified Encoder-Decoder

- Methods: Unified VLP

- Both encoder and decoder are pre-trained



Sun et al. “Videobert: A joint model for video and language representation learning,” ICCV 2019.
Li et al. “Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks,” arXiv 2020.
Zhou et al. “Unified vision-language pre-training for image captioning and vqa”, AAAI 2020.

Evaluation – Benchmark Dataset

COCO Captions

- Train / val / test: 113k / 5k / 5k
- Hidden test (leaderboard): 40k
- Vocabulary (≥ 5 occurrences): 9,587
- Most-adopted!

Flirckr30K

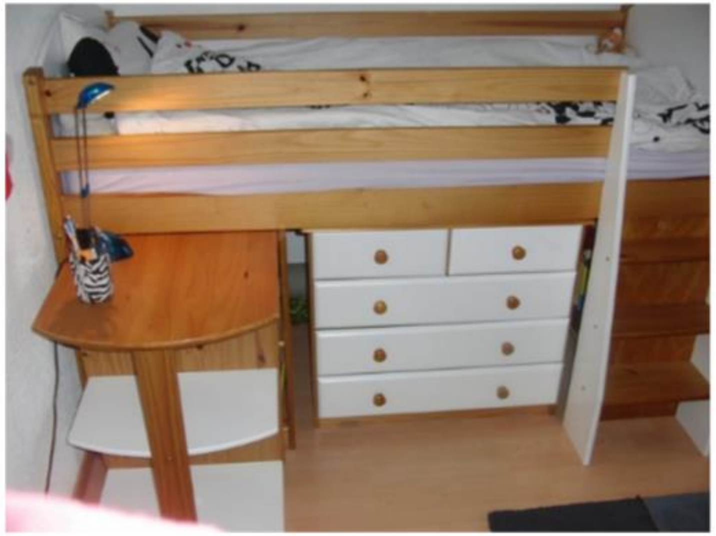
- Train / val / test: 29k / 1k / 1k
- Vocabulary (≥ 5 occurrences): 6,864

Chen et al. "Microsoft coco captions: Data collection and evaluation server", arXiv 2015.

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.

Image credit: Chen et al. “Microsoft coco captions: Data collection and evaluation server”, arXiv 2015.

Evaluation – Metrics

- Most commonly-used: BLEU / METEOR / CIDEr / SPICE
  - BLEU: based on n-gram based precision
  - METEOR: ordering sensitive through unigram matching
  - CIDEr: gives more weight-age to important n-grams through TF-IDF
  - SPICE: F1-score over caption scene-graph tuples

- Further readings: Sanja Fidler’s lecture slides http://www.cs.toronto.edu/~fidler/slides/2017/CSC2539/Kaustav_slides.pdf

Evaluation – Results on COCO

Method	BLEU@4	METEOR	CIDEr	SPICE
CNN-LSTM	20.3	-	-	-
Soft Attention	24.3	23.9	-	-
Semantic Attention	30.4	24.3	-	-
Adaptive Attention	32.5	26.6	108.5	19.5
Region Attention*	36.3	27.7	120.1	21.4
Attention on Attention*	38.9	29.2	129.8	22.4
Transformer (vanilla)*	38.2	28.9	128.4	22.2
M^2 Transformer*	39.1	29.2	131.2	22.6
X-Transformer*	39.7	29.5	132.8	23.4
VLP (with pre-training)*	39.5	29.3	129.3	23.2
Oscar (with pre-training)*	**41.7**	**30.6**	**140.0**	**24.5**

Note that all methods use a single model.
* Indicates with CIDEr optimization

Image Captioning – Other Topics

- Dense Captioning
- Novel Object Captioning
- Stylized Captioning (GAN)
- RL-based (e.g., SCST)

Video ~~Captioning~~ Description

- Now, we extend our scope to the video domain.

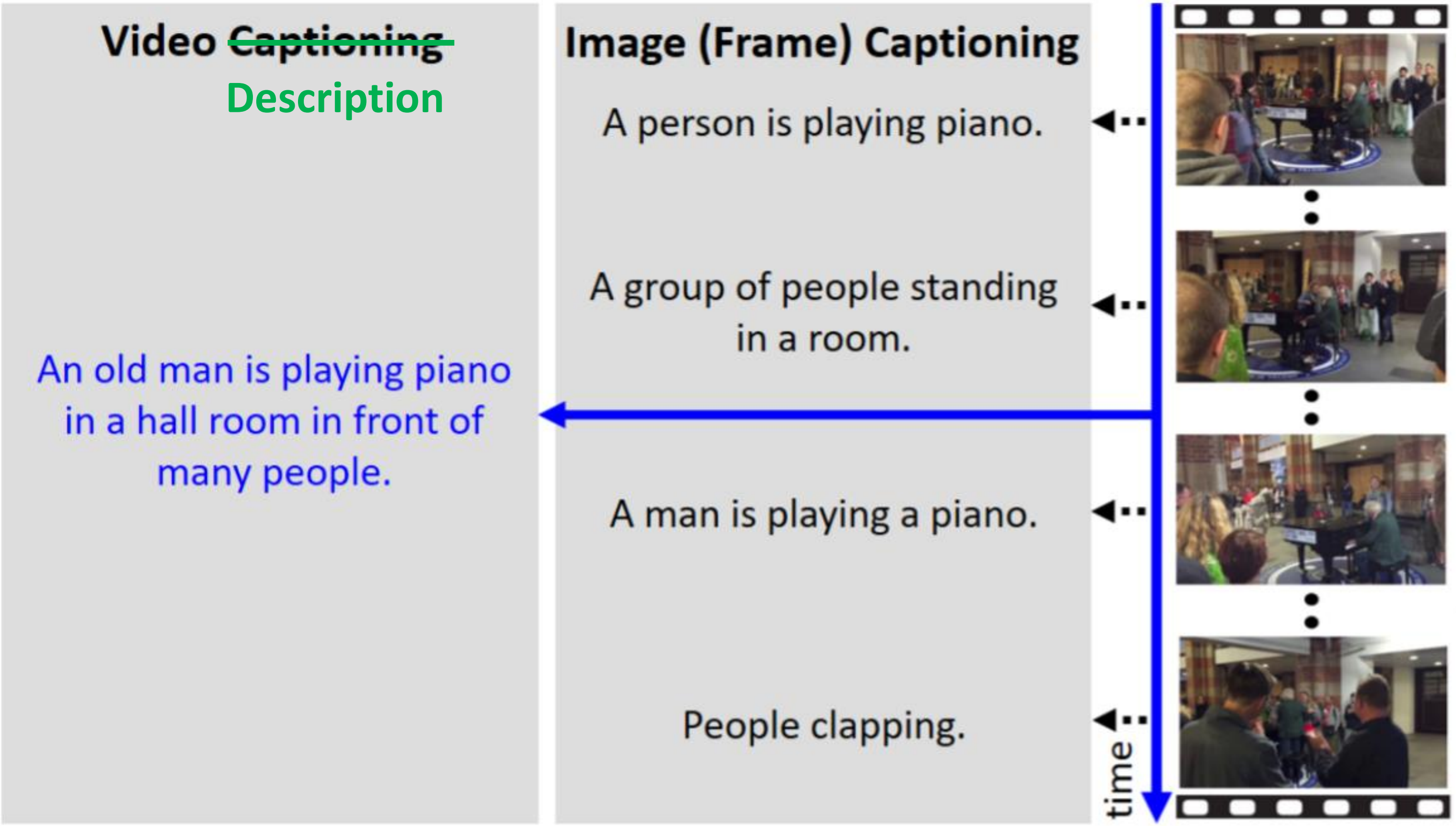


Image credit: Aafaq et al. "Video Description: A Survey of Methods, Datasets and Evaluation Metrics" ACM Computing Surveys 2019.

Video Description

- Method-wise, almost no difference! (enc-dec, attention, Trans. etc.)
- The temporal info is aggregated through the following methods:

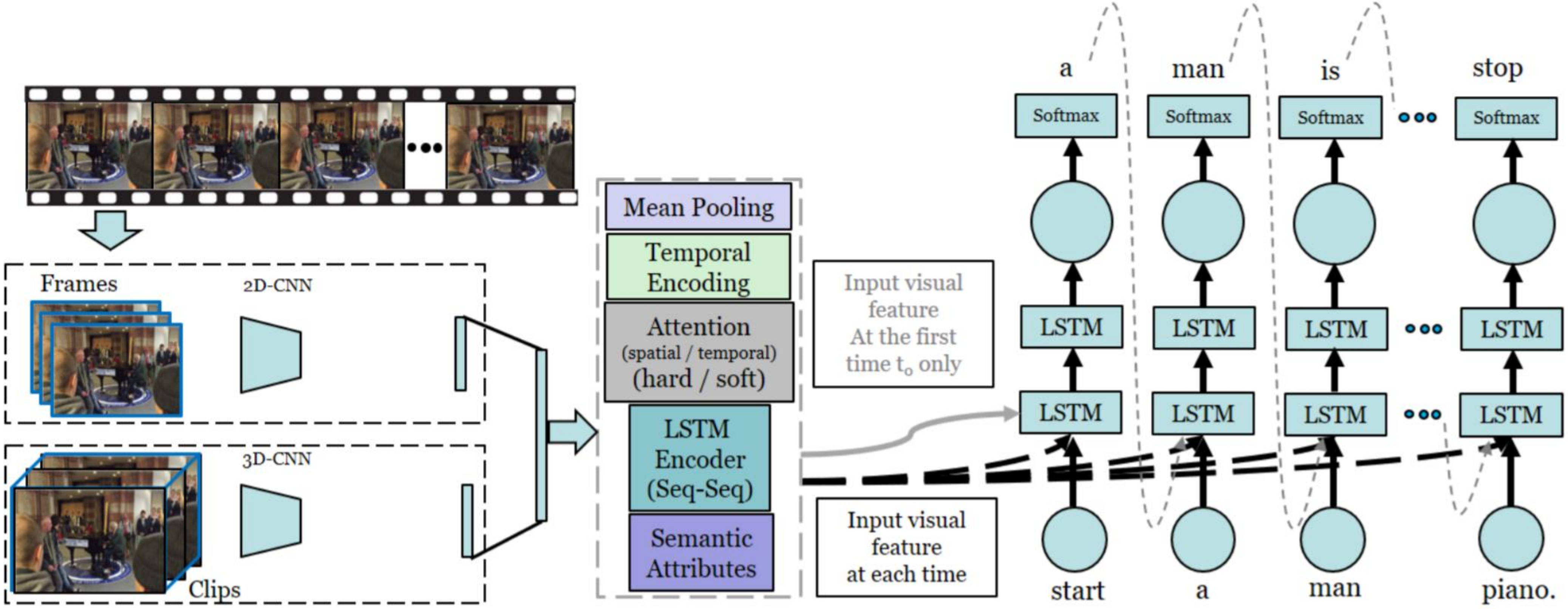
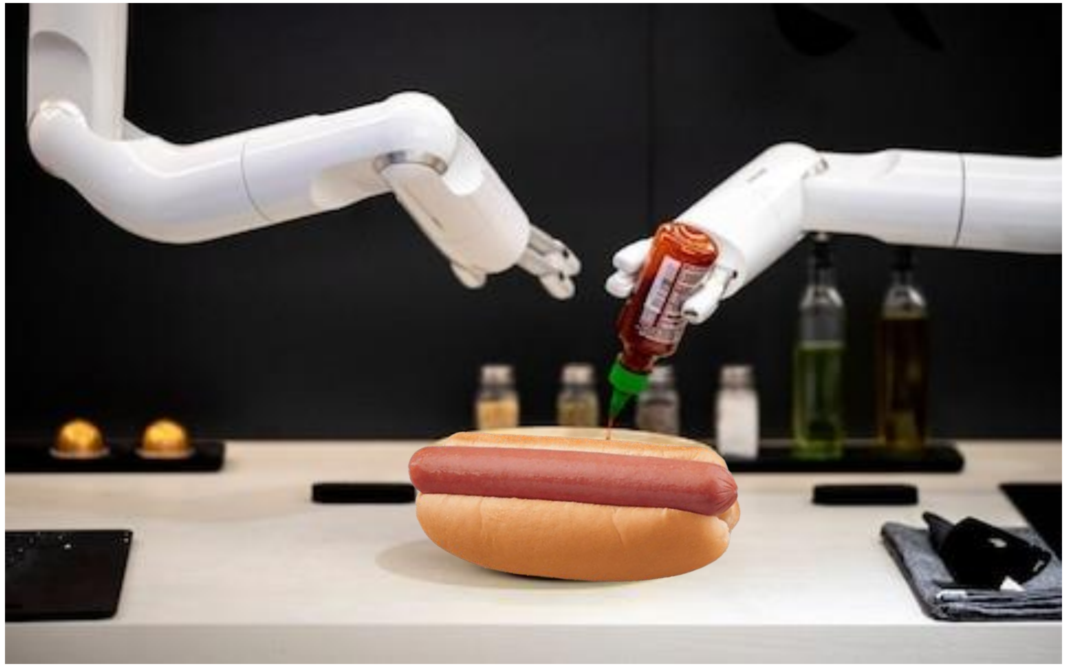


Image credit: Aafaq et al. “Video Description: A Survey of Methods, Datasets and Evaluation Metrics” ACM Computing Surveys 2019.

Description alone might fail...

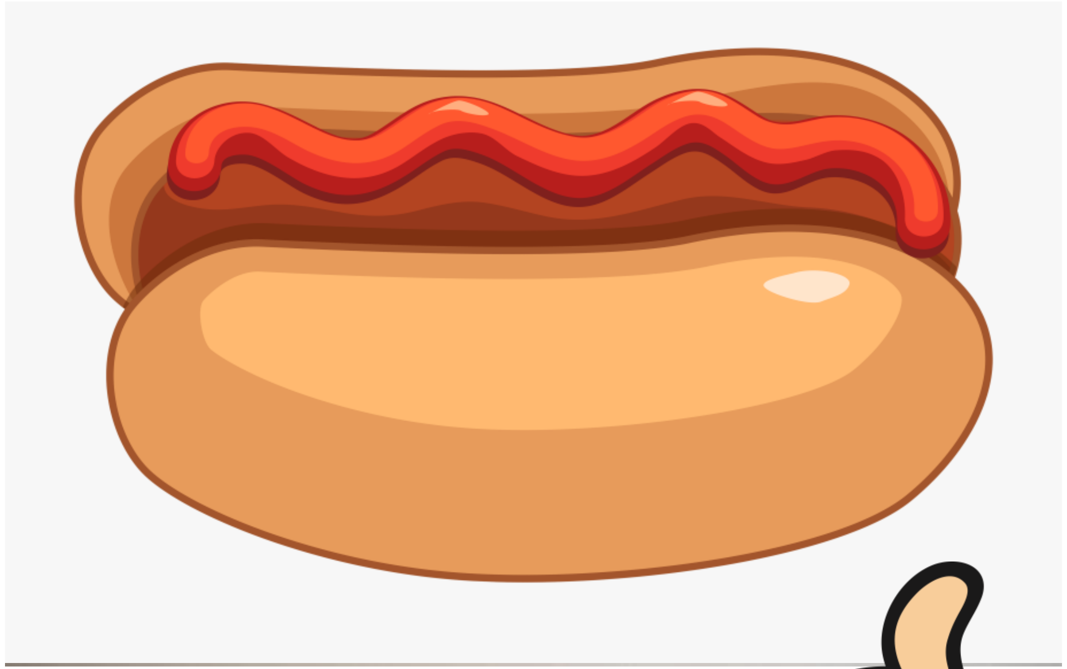


Description: A bottle of ketchup and a bottle of sriracha are on a table.

See notes for the image credits.

Description alone might fail...

Description: A bottle of ketchup and a bottle of sriracha are on a table.



See notes for the image credits.

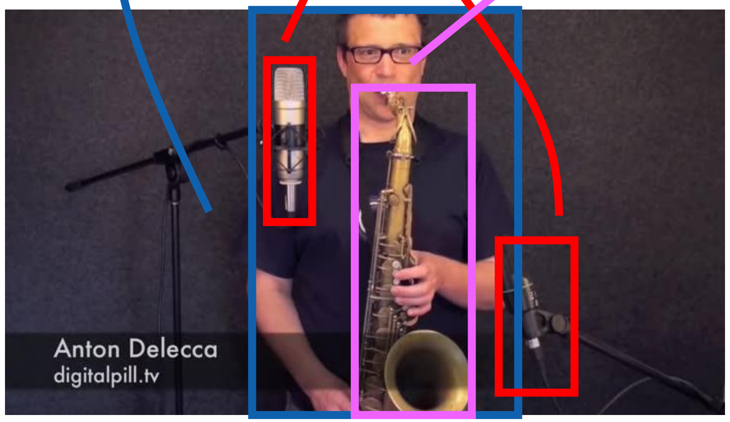
Grounded Visual Description

- Essentially, visual description + object grounding or detection
- In the image domain, Neural Baby Talk
- In the video domain, Grounded Video Description
- Requires special dataset that has both description and bounding box

Lu et al. “Neural Baby Talk”, CVPR 2018.
Zhou et al. “Grounded video description”, CVPR 2019.

Single-Frame Annotation

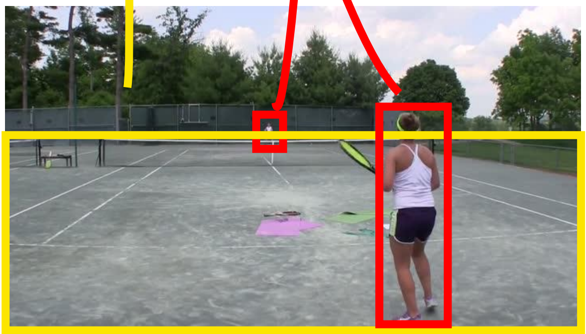
We see a man playing a saxophone in front of microphones.



From ActivityNet-Entities dataset. Zhou et al. “Grounded video description”, CVPR 2019.

Multi-Frame Annotation

Two women are on a tennis court, showing the technique to posing and hitting the ball.



From ActivityNet-Entities dataset. Zhou et al. “Grounded video description”, CVPR 2019.

Grounded Video Description (GVD) model

- Architecture: grounding module + caption decoder

- Grounding happens simultaneously with caption generation.

- GVD adopts **three proxy tasks** to leverage the BBox annotations:
  - Supervised attention
  - Supervised grounding
  - Region classification

- Details: https://www.youtube.com/watch?v=7AVCgn21noM

Zhou et al. “Grounded video description”, CVPR 2019.

Video Description

- The Encoder-Decoder framework works fairly well for images and short video clips.

- How about long videos?

- The average video length on YouTube is 4.4 minutes!

Video Paragraph Description

Add chopped bacon to a hot pan and stir. Remove the bacon from the pan. Place the beef into a hot pan to brown. Add onion and carrots to the pan. Pour the meat back into the pan and add flour. Place the pan into the oven. Add bay leaves thyme red wine beef stock garlic and tomato paste to the pan and boil. Add pearl onions to a hot pan and add beef stock bay leaf and thyme. Add mushrooms to a hot pan. Add the mushrooms and pearl onions to the meat...

Yu et al. “Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks” *CVPR 2016*.

Dense Video Description

- Objective – Localize and describe events from a video.
- Input: Video. Output: Triplets of event start, end time and description

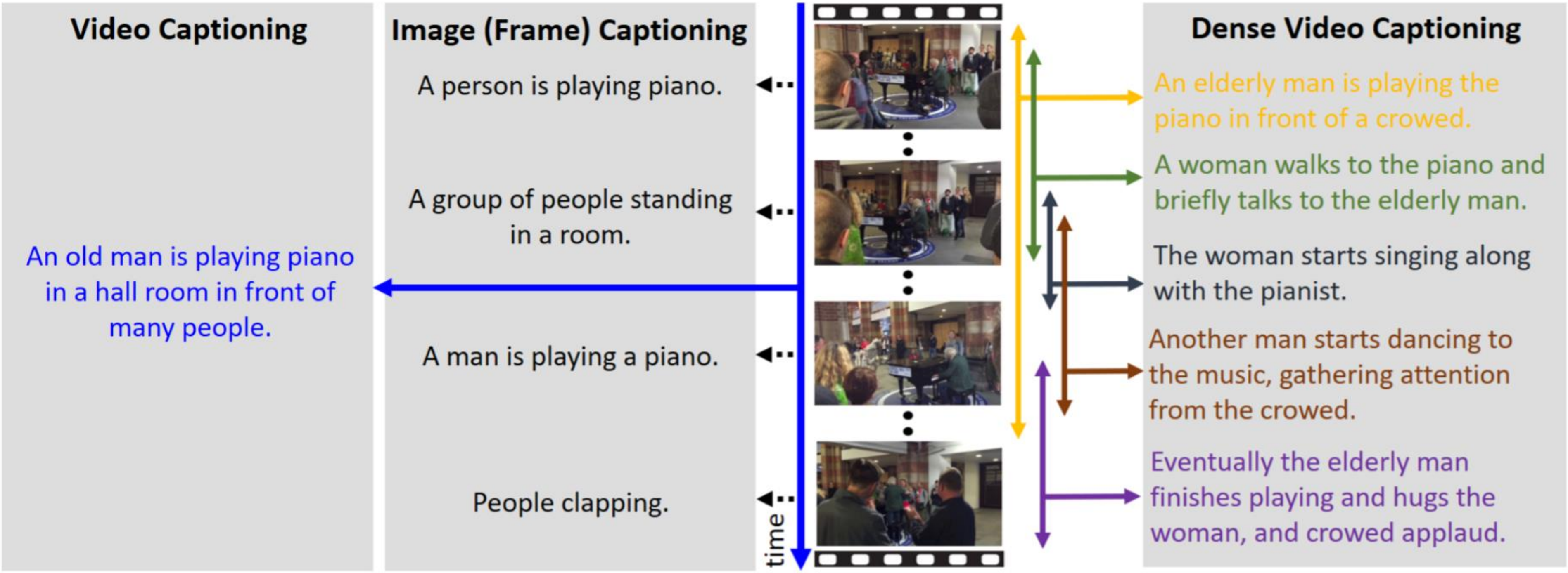
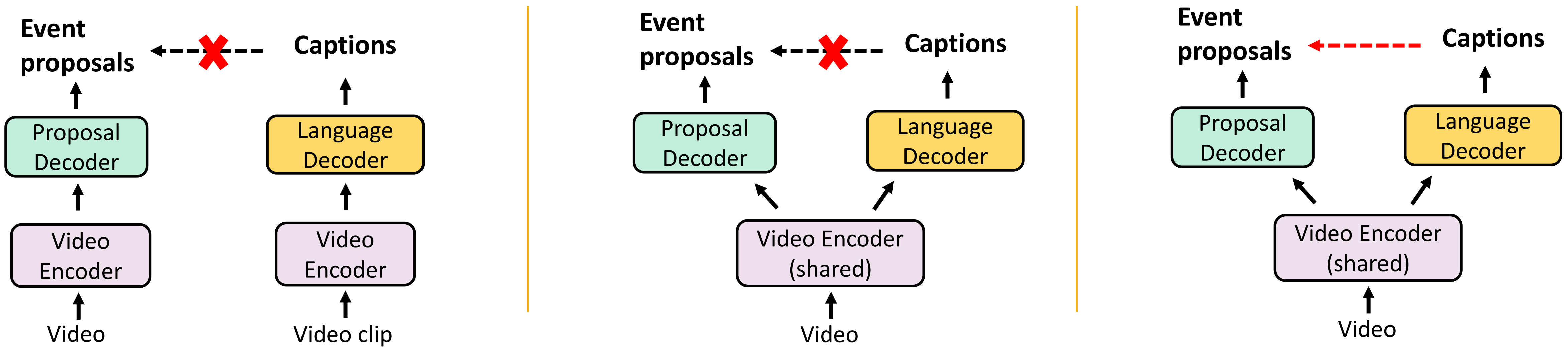


Image credit: Aafaq et al. “Video Description: A Survey of Methods, Datasets and Evaluation Metrics”, ACM Computing Surveys 2019.

Dense Video Description

- Existing methods usually contain two modules: event proposal and video description.

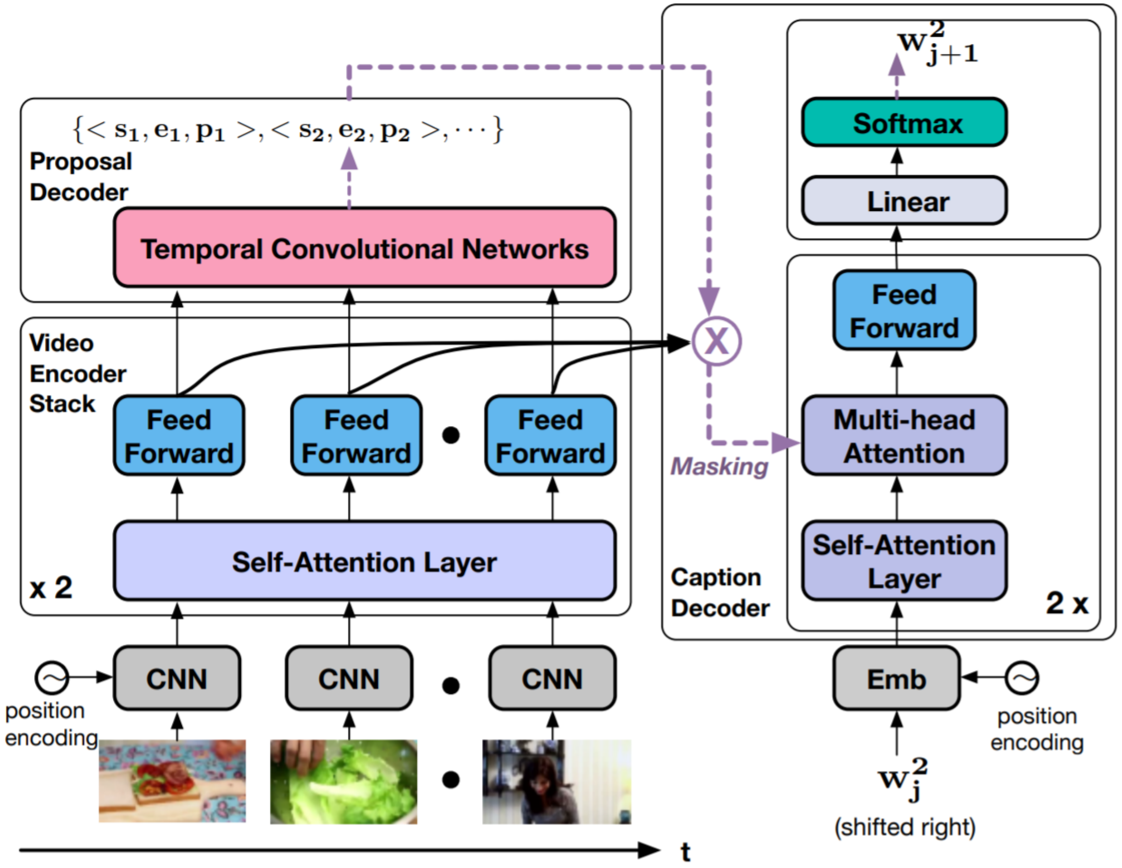


Separate Training | Alternating Training | End-to-End Training

Krishna et al. “Dense-captioning events in videos”, ICCV 2017.
Zhou et al. “End-to-end dense video captioning with masked transformer”, CVPR 2018.

Dense Video Description

- End-to-End Masked Transformer



Zhou et al. "End-to-end dense video captioning with masked transformer", CVPR 2018.

Conclusions

- We have seen aggressive progresses in the field…
  - On COCO Captions, CIDEr goes from <100 to 140
- Motivation is important. Avoid piling up “Legos”.
- To achieve better result interpretability, we need grounding.
- Towards generalizable and robust models, pre-training is one option

Limitations

- Still a long way to go before production-ready due to…
- Recognition failure -> better feature
- Object hallucination -> better grounding/detection
- Model bias -> alleviating biases

Rohrbach et al. "Object hallucination in image captioning," EMNLP 2018.
Burns et al. "Women also Snowboard: Overcoming Bias in Captioning Models," ECCV 2018.

Future Directions

- Evaluation metrics that correlate better with human judgement.
- Revisit grid features and simplify the model pipeline.
- In vision-language pre-training, how to close the gap between pre-training domain and downstream domain.

Thank you!
Any questions?