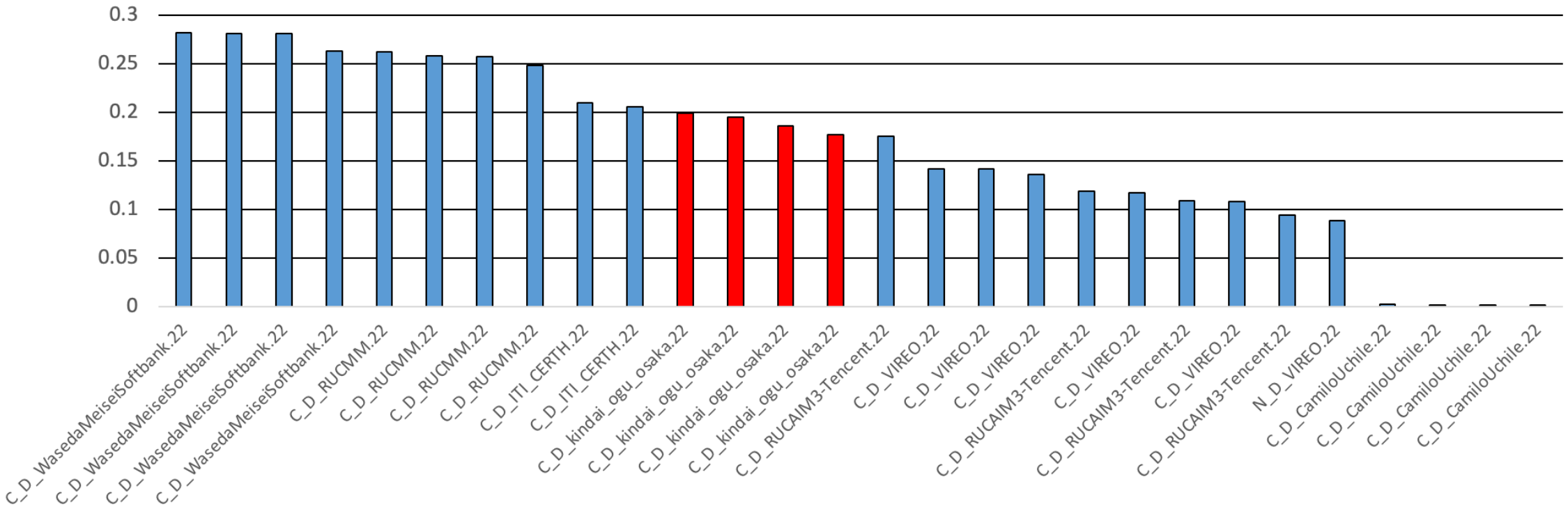
kindai_ogu_osaka at TRECVID 2022 AVS Task: A Hierarchical Approach to Dynamically Treat Meanings of Words and Phrases

Kimiaki Shirahama, Kazuma Fujioka, Taichi Shinno (Kindai University)

Takashi Matsubara (Osaka University)

Kuniaki Uehara (Osaka Gakuin University)

Our Submitted Runs



We are ranked at the fouth position among seven teams in the fully automatic cateogry of AVS task

Our main focus:

1. Retrieval by aligning regions in a frame with words and phrases in a topic
2. Expolation of CLIP models

Decomposition of a Frame and a Topic into Shared Concepts

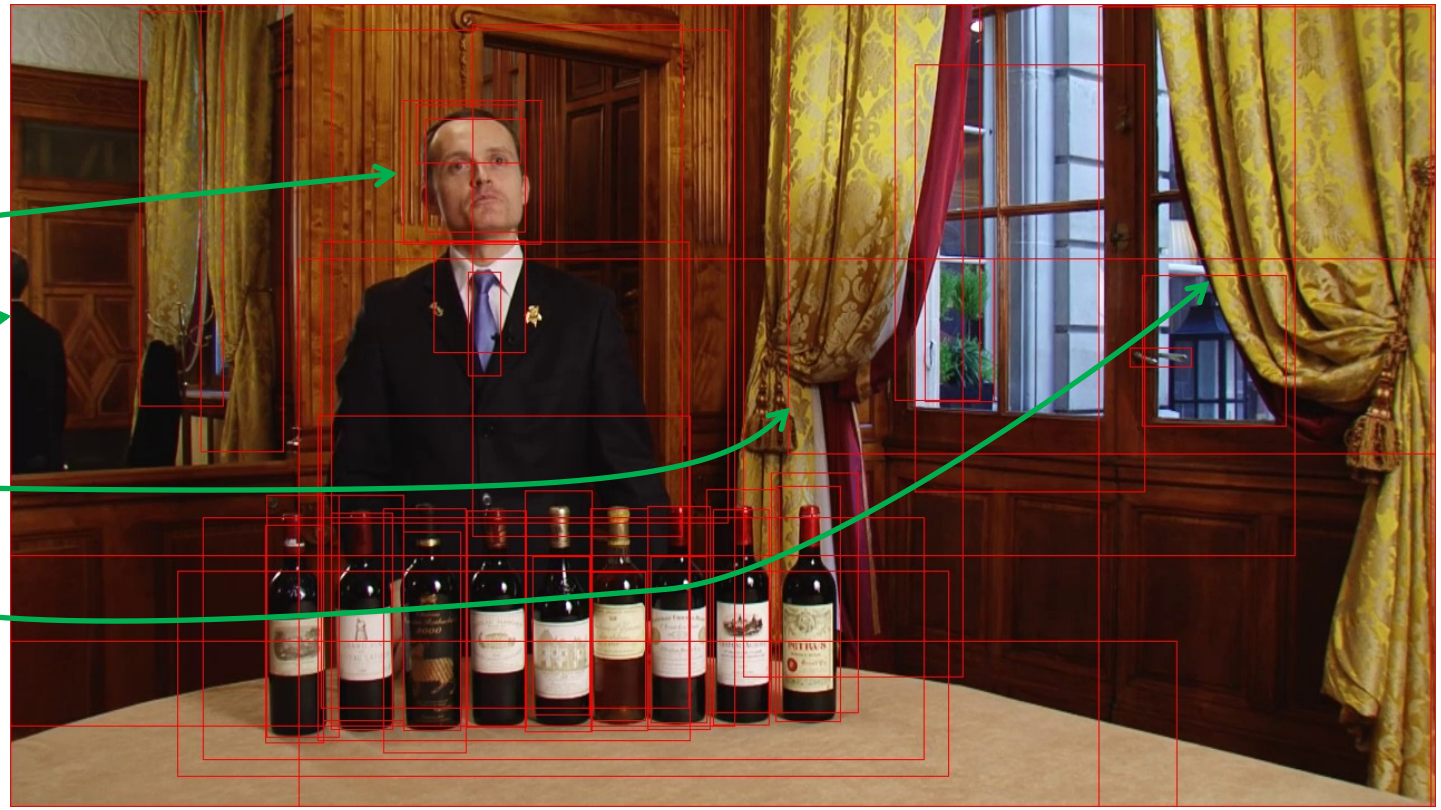
Human examines the relevance of a frame to a topic by unconciously matching regions with words and phrases

Alignment between regions and words/phrases

Shared concepts

indicating conceptually the same meanings

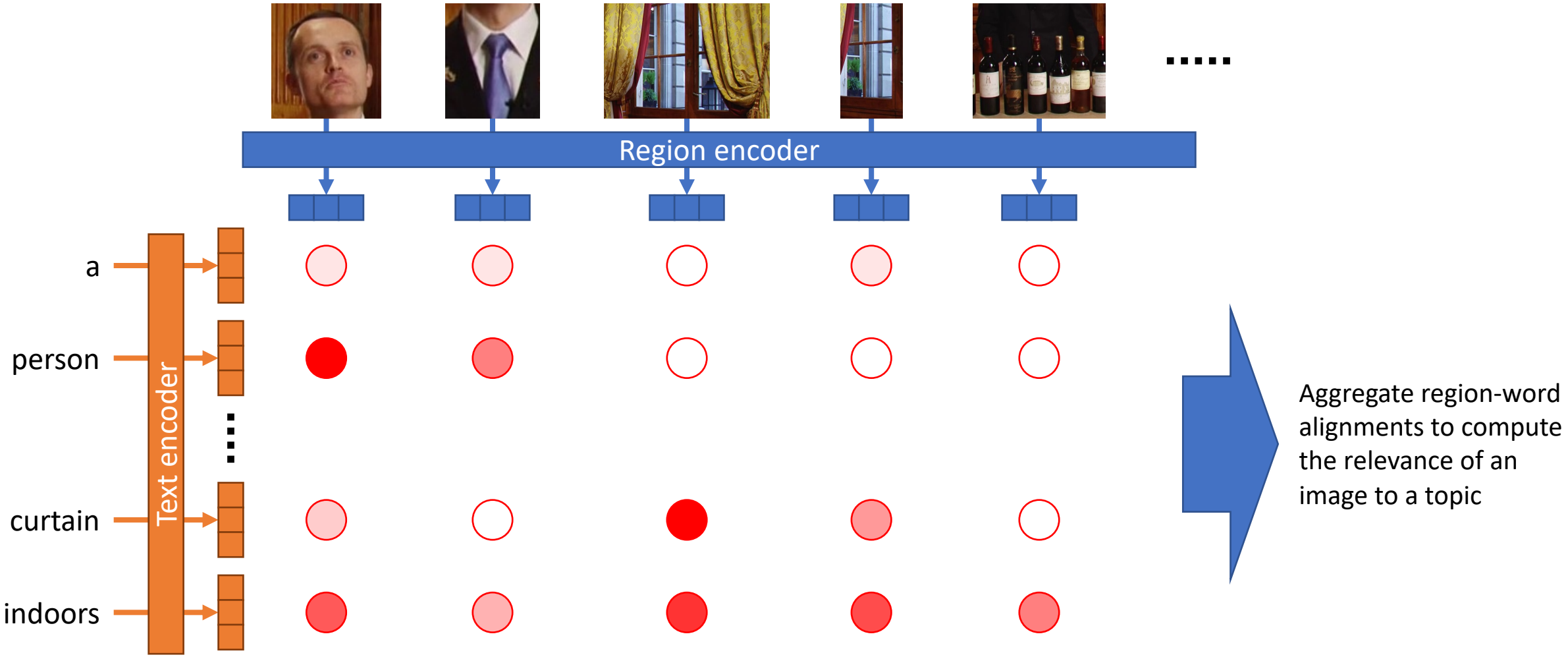
A person in front of a curtain indoors



SCAN: Stacked Cross Attention for Image-Text Maching

Learn to encode regions and words into vectors so as to match semantically relevant images (frames) and captions (topics)

A region and a word sharing the same concept are encoded into similar vectors (i.e., ***aligned***)



Our Extensions of SCAN

- Revise the setting
  - ➡ Use **Conceptual Captions** as large-scale training data and **VinVL** to extract the state-of-the-art visual features
- Align regions with words as well as phrases
  - ➡ Use a **neural parser** to obtain a constituency tree of a topic and **Tree-LSTM** to extract textual features for phrases
- Improve the consistency of aligning regions with words/phrases
  - ➡ Devise a **hierarchical alignment** approach to gradually address more specific phrases based on a constituency tree of a topic

Revising the Setting of SCAN

Until last year → This year

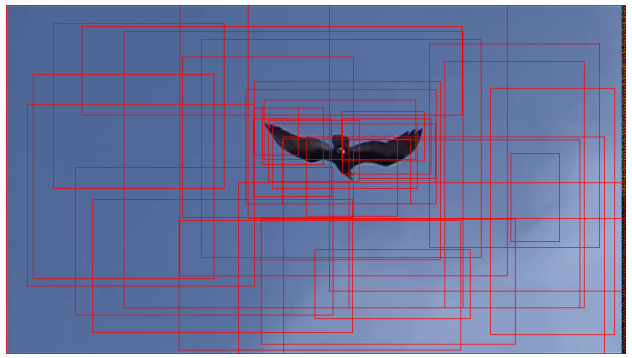
MS-COCO (0.56M image-caption pairs)

1 image is associated with 5 captions

Bottom-up attention

- Faster R-CNN based on ResNet101 backbone
  - Visual Genome
    98K images
    1600 object classes
    400 attribute classes

- Extract a **fixed number** of regions
  (36 regions)



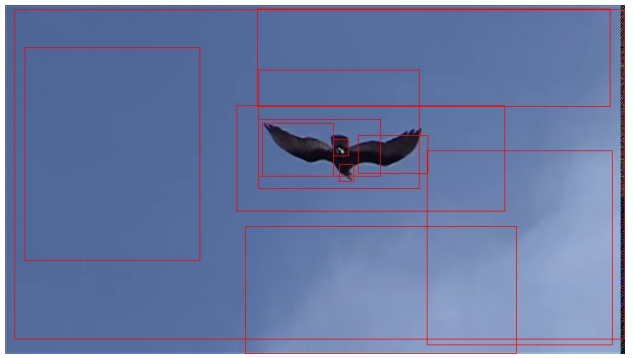
Conceptual Captions (2.94M image-caption pairs)

VinVL

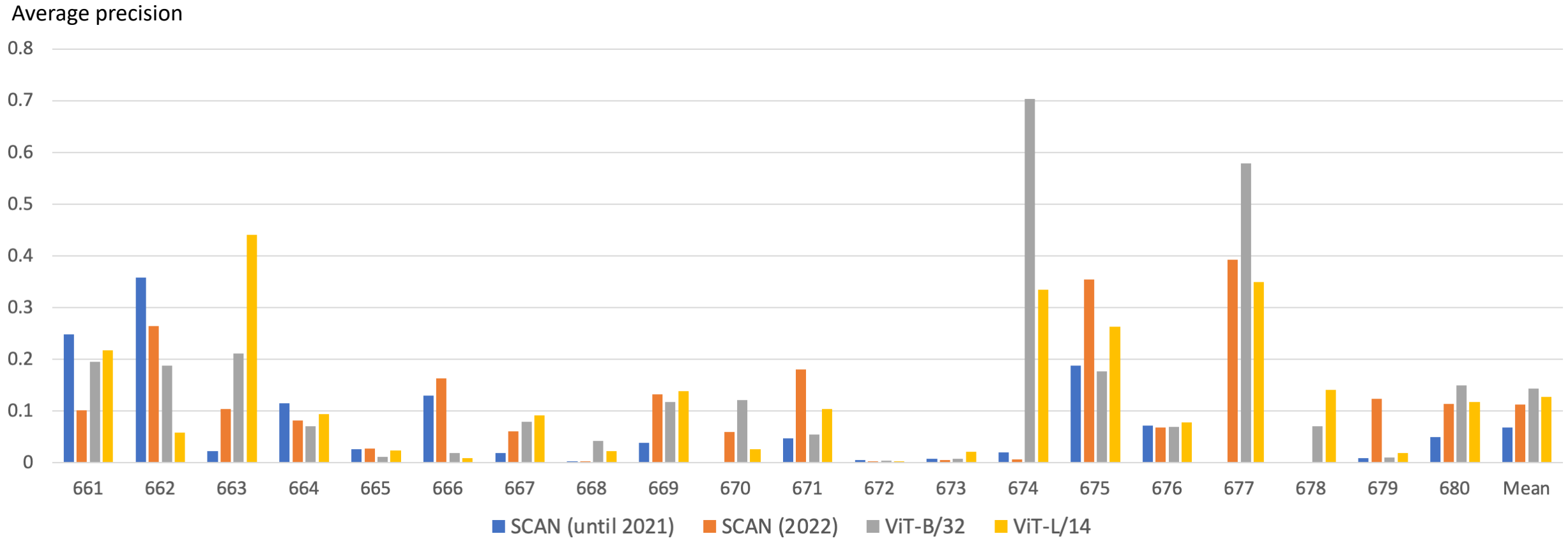
- Faster R-CNN based on ResNeXt-152 backbone
  - MS-COCO
  - OpenImages
  - Objects365
  - Visual Genome

  5.43M images
  1848 object classes
  524 attribute classes

- Extract a **variable number** of regions



Performance Improvement by the Setting Revision



- A significant performance improvement is achieved compared to SCAN used until last year
- The performance of this year's SCAN is relatively close to those of CLIP models (*despite the insufficent number of epochs*)

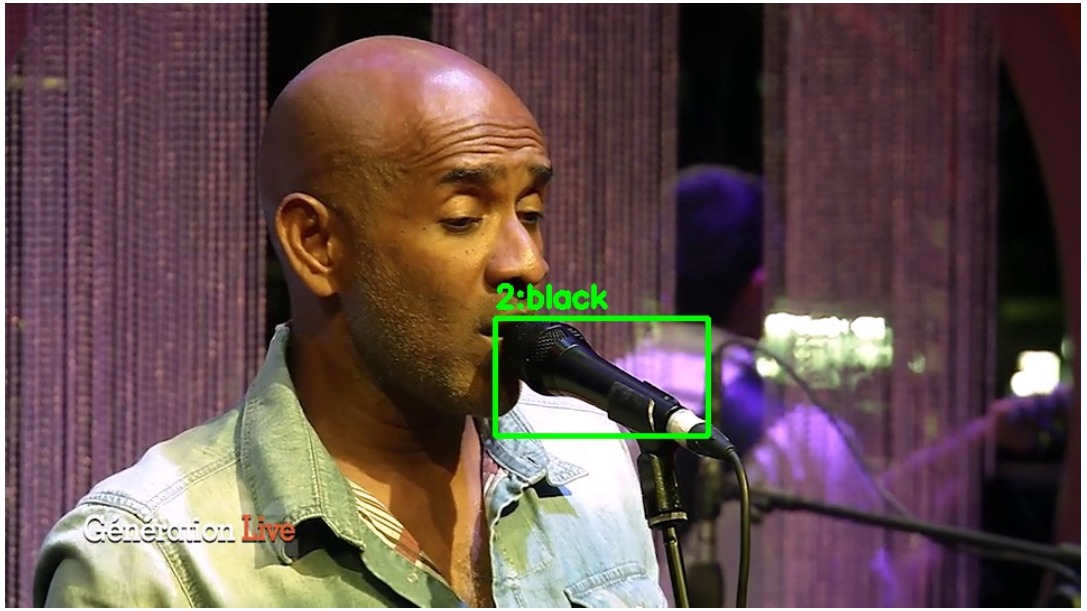
Why Do We Consider Phrases?

- Phrases representing specific meanings are more interesting for users
- Component words of a phrase are aligned with separate regions
- The meaning of a word dynamically changes depending on word combinations

Motivating examples

(A woman wearing a red dress outside in the daytime)

(A black man singing)



Extracting a Constituency Tree of a Caption

Attach-juxtapose parser

Sequential addition of each word to a tree by

- attaching it as a child node of an existing node

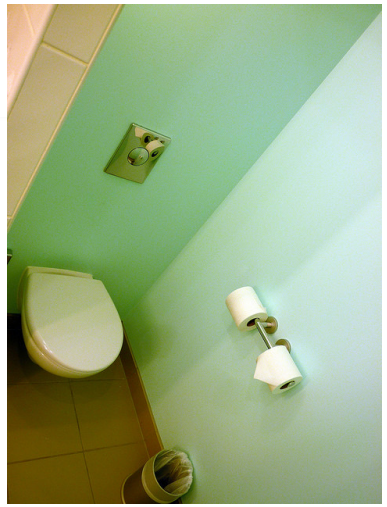
or

- juxtaposing it as a sibling to an existing node

Train a neural network to perform this action selection

Confirmed that the attach-juxtapose parser can extract semantically reasonable constituency trees for various captions

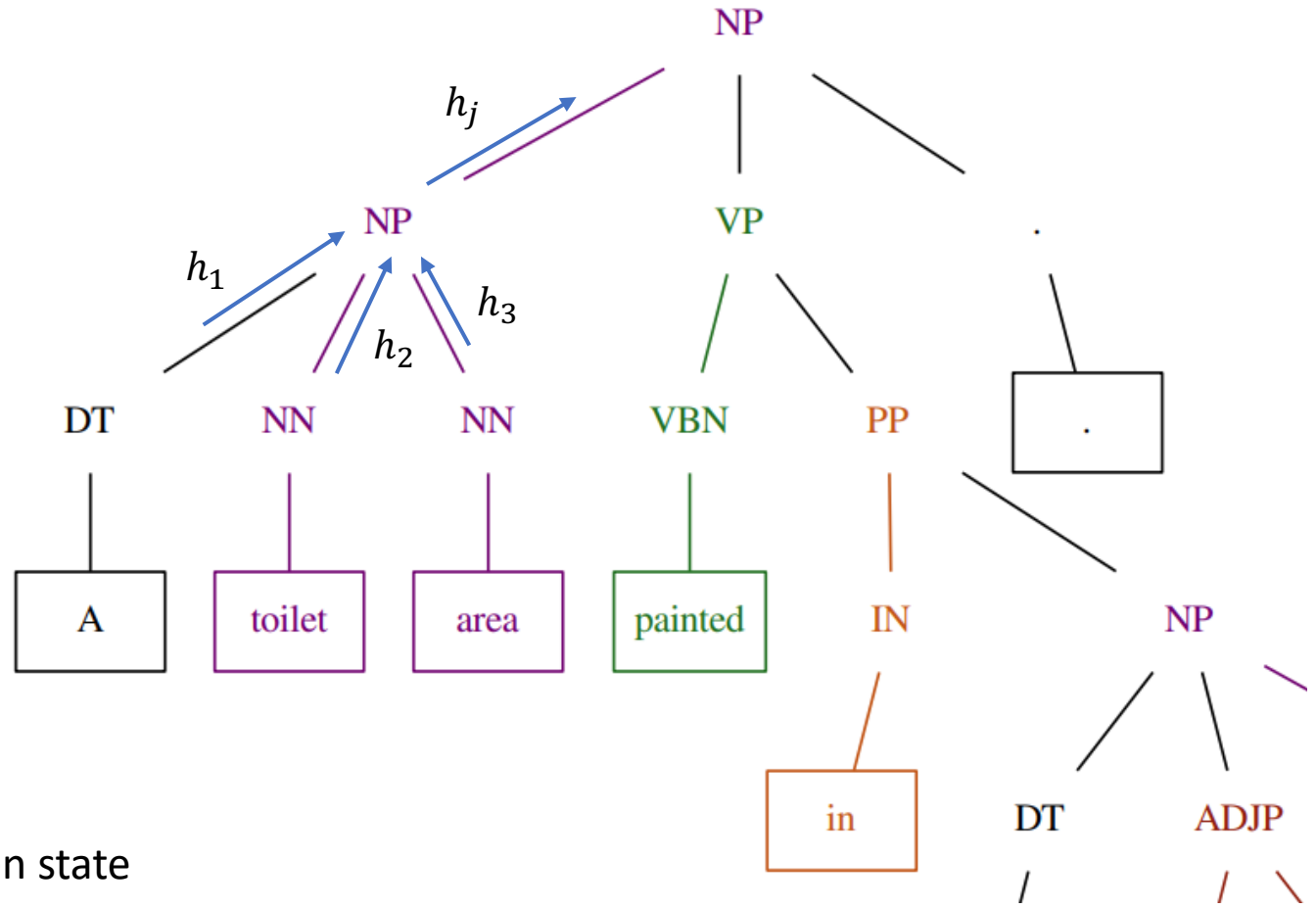
(A toilet area painted in a light green)



Feature Extraction for Phrases

Tree-LSTM: Extension of LSTM to propagate hidden states based on a tree structure

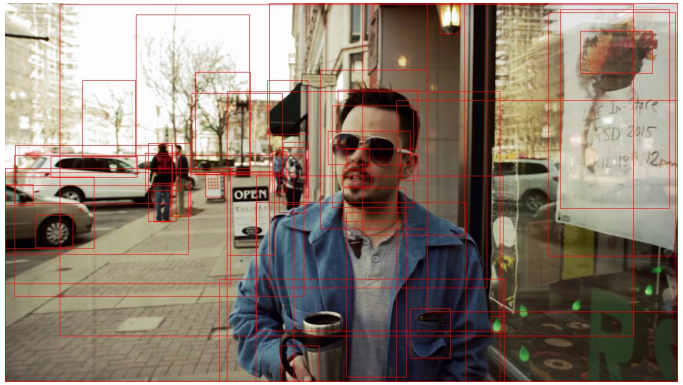
- Aggregated hidden state: $\tilde{h}_j = \sum_{k \in C(j)} h_k$
- Input: $u_j = \tanh\left(W^{(u)} x_j + U^{(u)} \tilde{h}_j + b^{(u)}\right)$
- Input gate: $i_j = \sigma\left(W^{(i)} x_j + U^{(i)} \tilde{h}_j + b^{(i)}\right)$
- Forget gate: $f_{jk} = \sigma\left(W^{(f)} x_j + U^{(f)} h_k + b^{(f)}\right)$
- Output gate: $o_j = \sigma\left(W^{(o)} x_j + U^{(o)} \tilde{h}_j + b^{(o)}\right)$
- Memory cell: $c_j = i_j \odot u_j + \sum_{k \in C(j)} f_{jk} \odot c_k$
- Hidden state: $h_j = o_j \odot \tanh\left(c_j\right)$

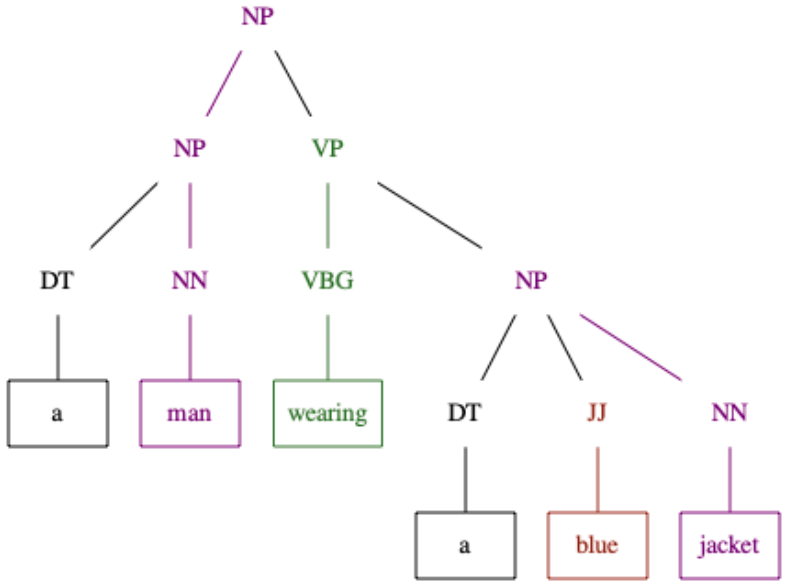


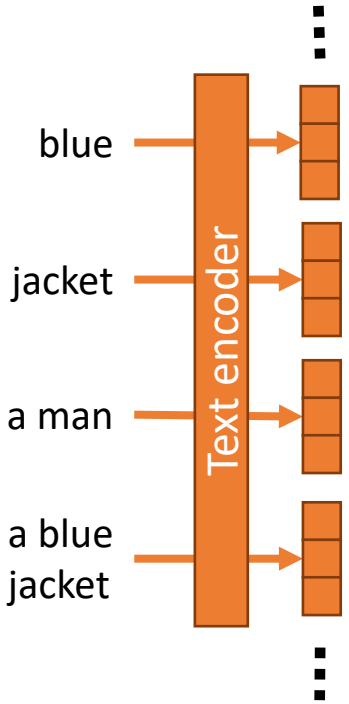
Starting with the 300-D feature of each word, the hidden state of each node is computed in a bottom-up fashion

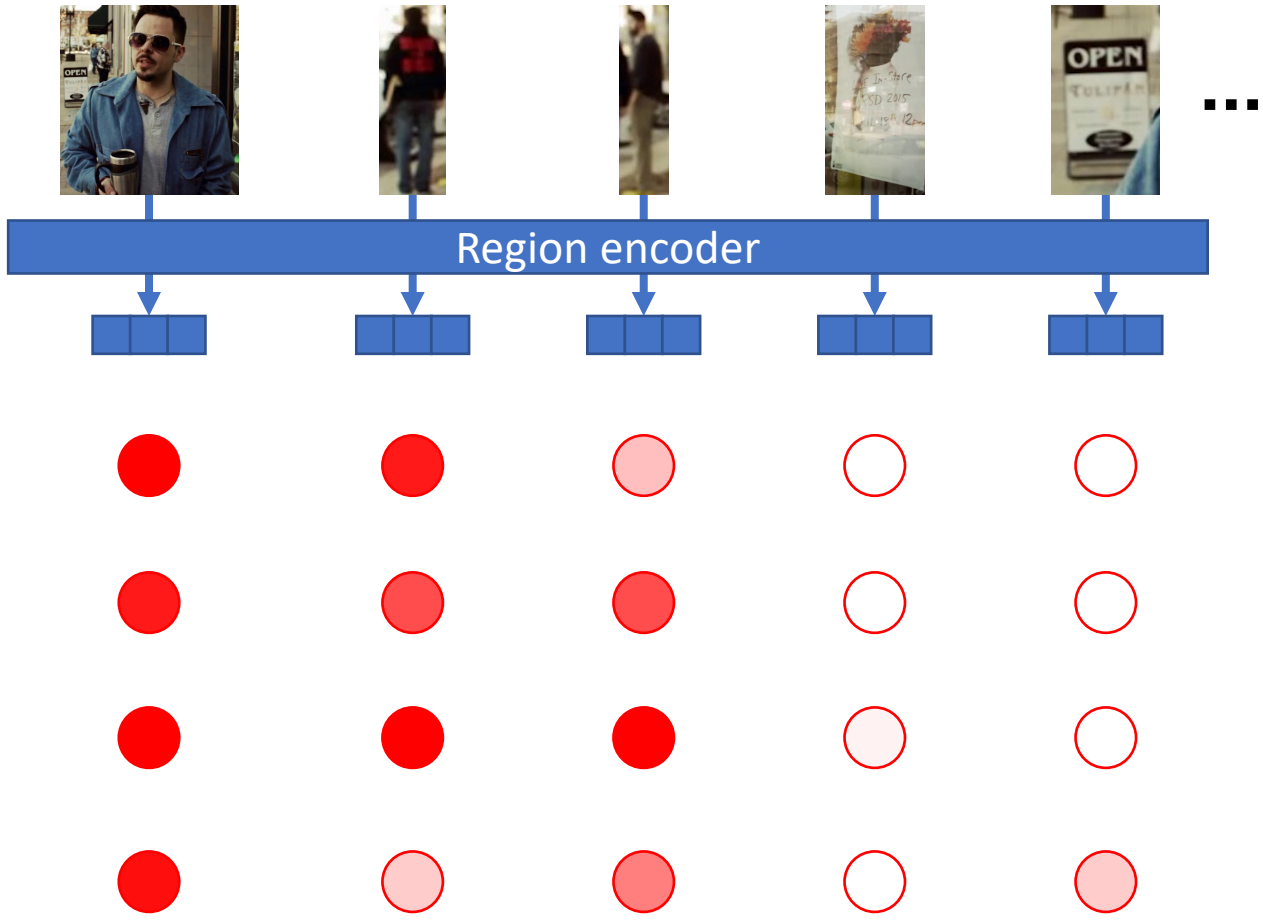
Alignment between Regions and Words/phrases

Regard phrases as additional words to be aligned





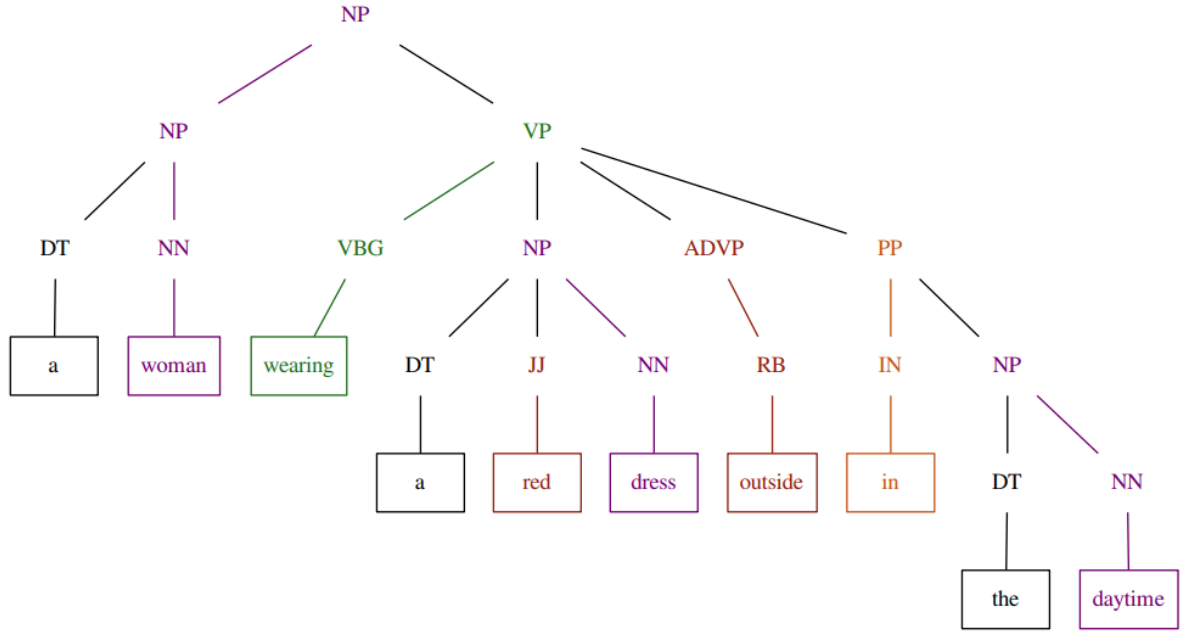




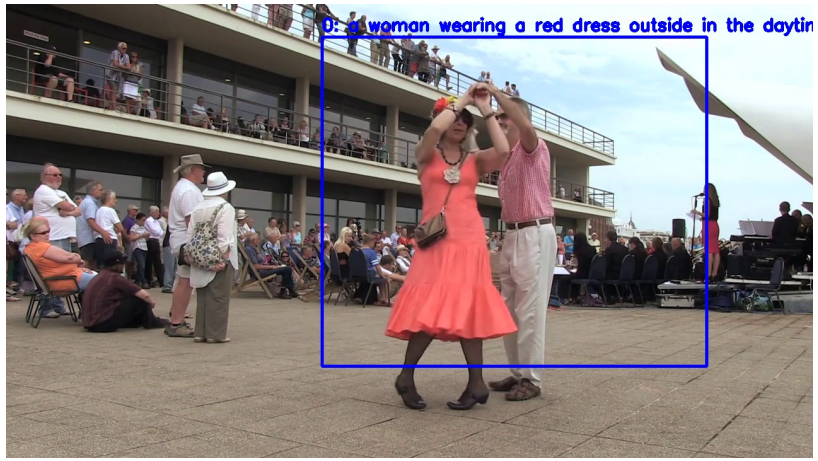
Inconsistent Region Alignment for Words and Phrases

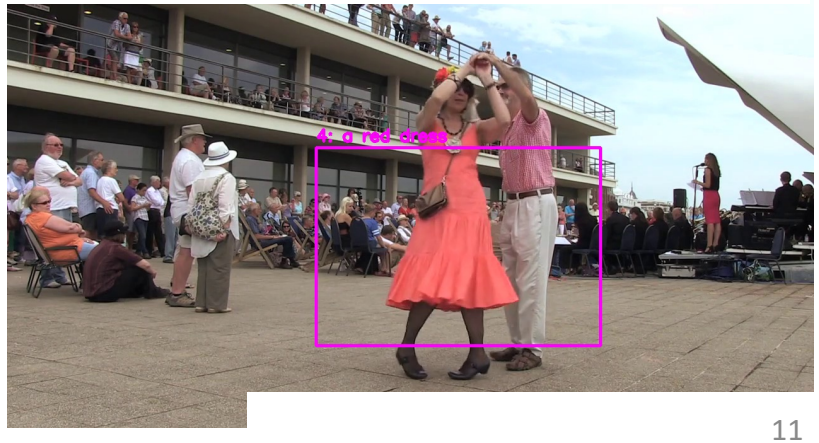
A phrase represents a more specific meaning than words (or sub-phrases)

The region aligned with the former is usually the same or smaller than those of the latter

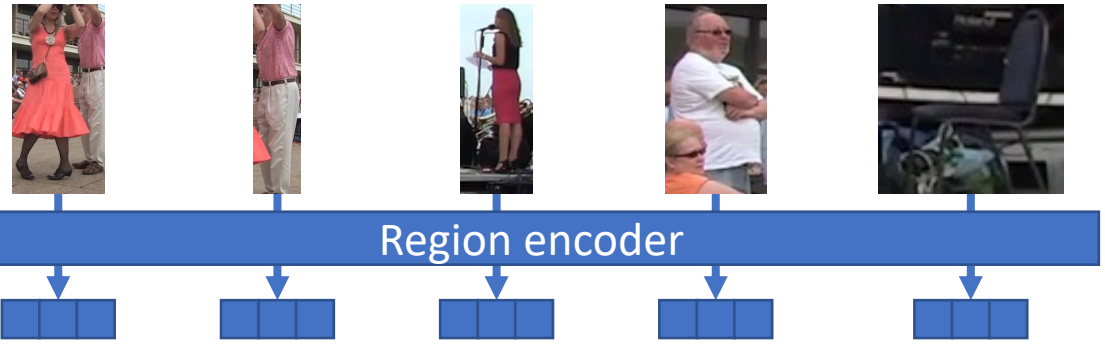


A woman wearing a red dress outside in the daytime





Hierarchical Alignment between Words/phrases and Regions



Region encoder

a red dress

Text encoder

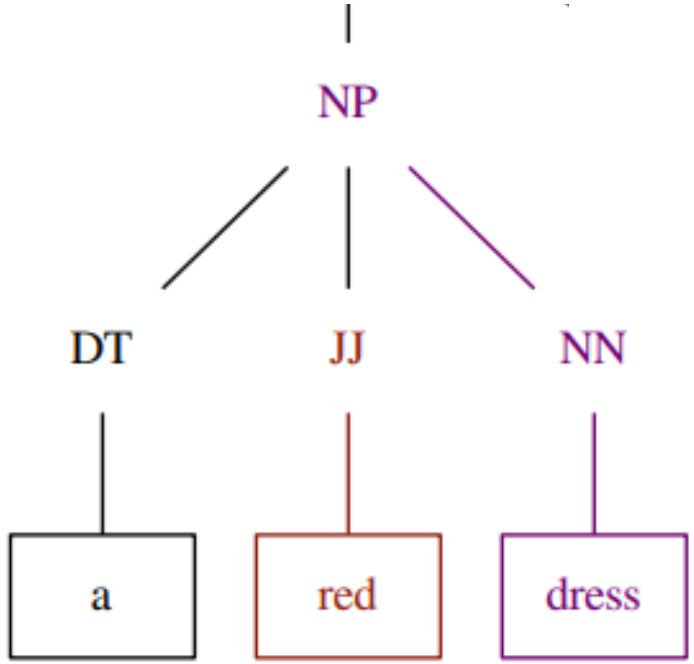
a

red

dress

Propagate aligment results of words (phrases) to the alignment of their parent phrase

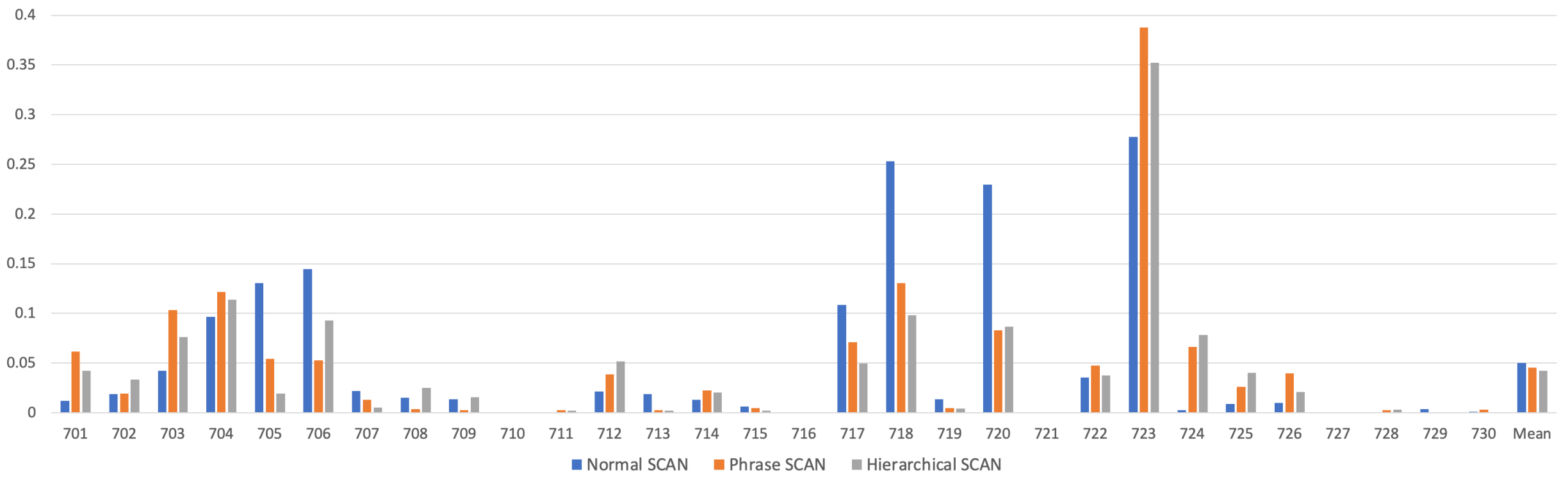
$$\alpha'_j = \alpha_j + \beta \sum_{k \in C(j)} \alpha'_k$$



Performance Comparison among Different Alignment Approaches

- **Normal SCAN:** Alignment of regions only with words
- **Phrase SCAN:** Non-hierarchical alignment of regions with words and phrases
- **Hierarchical SCAN:** Hierarchical alignment of regions with words and phrases

All models are trained on MSCOCO dataset



Currently no performance improvement is obtained by adopting hierarchical alignment, **BUT**

Case Study on Different Alignment Approaches

Phrase SCAN > Normal SCAN | Hierarchical SCAN > Normal SCAN | Too low performance

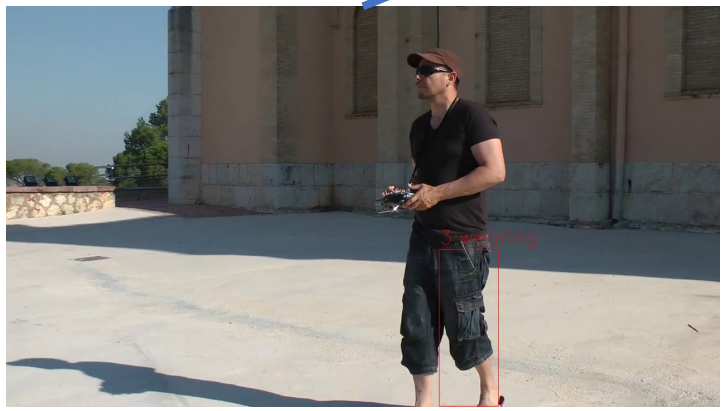
	Topic	Normal SCAN	Phrase SCAN	Hierarchical SCAN
701	A man with a white beard	0.0122	0.0618	0.0423
702	A room with blue wall	0.0187	0.0195	0.0337
703	A construction site	0.0422	0.1032	0.076
704	A parked white car	0.0964	0.1214	0.114
705	A type of cloth hanging on a rack, hanger, or line	0.1304	0.0543	0.0194
706	Building with columns during daytime	0.1444	0.053	0.0928
707	A person is mixing ingredients in a bowl, cup, or similar type of containers	0.0219	0.0131	0.0053
708	A female person bending downwards	0.0152	0.0035	0.0253
709	A person is in the act of swinging	0.0136	0.0029	0.0158
710	A person wearing a light t-shirt with dark or black writing on it	0.0006	0.0005	0.0008
711	A woman wearing a head kerchief	0.0003	0.0029	0.002
712	A man wearing black shorts	0.0212	0.0389	0.0516
713	A kneeling man outdoors	0.0188	0.0029	0.0019
714	Two or more persons in a room with a fireplace	0.0132	0.0227	0.0203
715	An Asian bride and groom celebrating outdoors	0.0065	0.0049	0.0019
716	A drone landing or rising from the ground	0	0	0
717	A black bird seen on a dry area sitting, walking, or eating	0.1084	0.0711	0.0494
718	A large stone building from the outside	0.2529	0.1304	0.0984
719	A piece of heavy farm equipment or machine seen outdoors	0.0136	0.0049	0.0042
720	A clock on a wall in a room	0.2294	0.0829	0.0865
721	Two persons are seen while at least one of them is speaking in a non-English language outdoors	0.0001	0	0
722	A woman is eating something outdoors	0.0354	0.0476	0.0377
723	A person is biking through a path in a forest	0.2777	0.388	0.3522
724	A man and a bike in the air after jumping from a ramp	0.0026	0.0661	0.0783
725	A woman holding or smoking a cigarette	0.0087	0.0261	0.0403
726	Two teams playing a game where one team have their players wearing white t-shirts.	0.0098	0.0396	0.021
727	Two persons wearing white outfits and black belts demonstrate martial arts in a room with floor mats	0	0.0001	0
728	Two adults are seated in a flying paraglider in the air	0	0.0027	0.0031
729	A ring shown on the left hand of a person	0.0036	0.0001	0.0001
730	A man is holding a knife in a non-kitchen location	0.001	0.003	0.0006

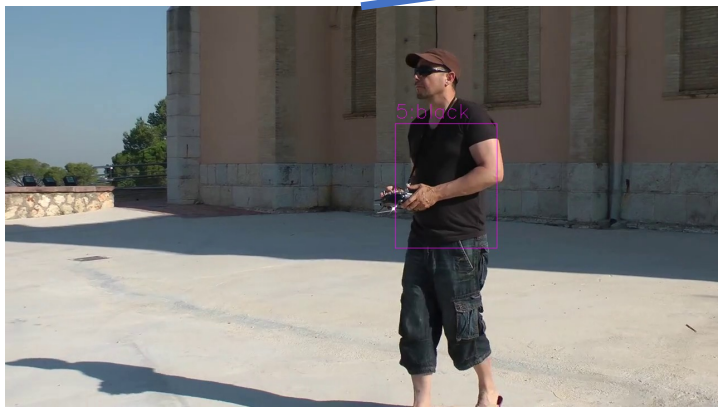
It seems that performance improvements are obtained for topics where phrases are important!

An Example of Alignment Result

A man wearing black shorts

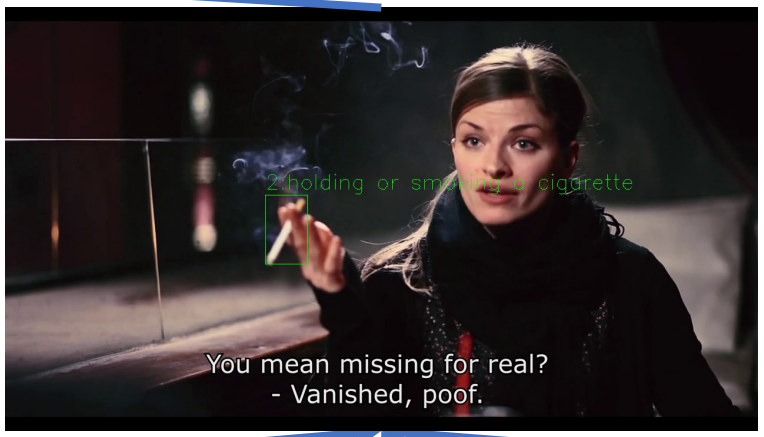
Semantically cosistent alignment between regions and words/phrases





Another Example of Alignment Result

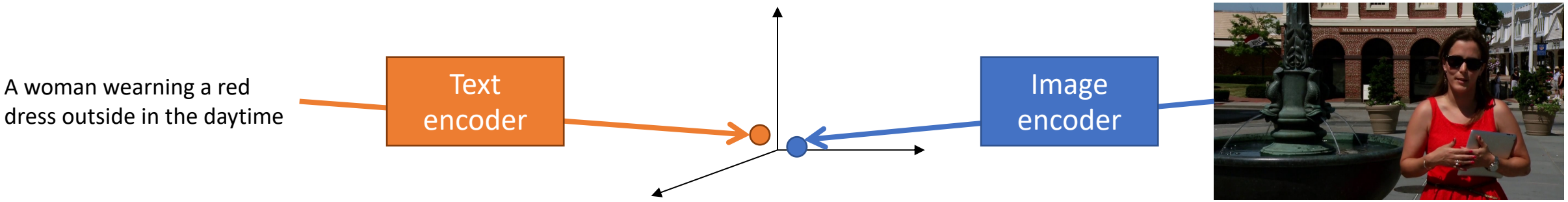
A woman holding or smoking a cigarette



Words and phrases in a topic are reasonably aligned with regions in a frame!

Exploration of CLIP Models

Transformer-based embedding model trained on a very large dataset containing 400M image-caption pairs
Most teams highly ranked at last year's AVS task used CLIP models

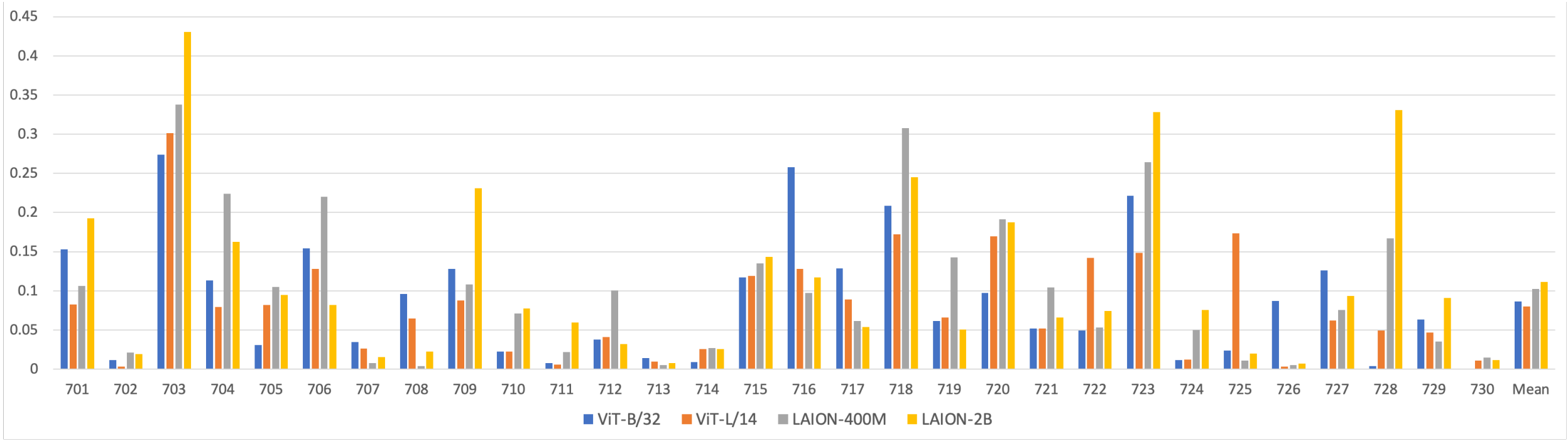


Use the image and text encoders to encode a frame and a topic into feature vectors, respectively

Four CLIP models we tested

1. **ViT-B/32:** Vision transformer based on a BERT base model accepting 32x32 pixel patches in a 224x224 pixel image
2. **ViT-L/14:** Vision transformer based on a BERT large model accepting 14x14 pixel patches in a 336x336 pixel image
   (ViT-B/32 and ViT-L/14 are trained on WebImageText (WIT) dataset containing 400M image-caption pairs)
3. **LAION-400M:** ViT-B/32 trained on an open dataset containing 400M image-caption pairs
4. **LAION-2B:** ViT-B/32 trained on an open dataset containing 2B image-caption pairs

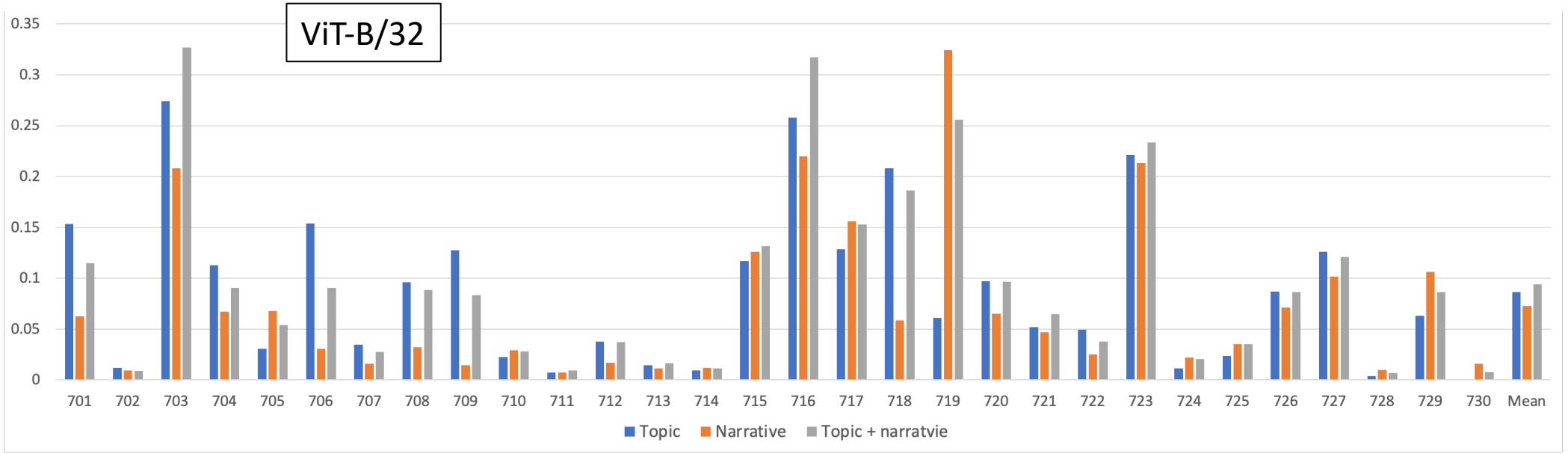
Performances of Individual CLIP Models

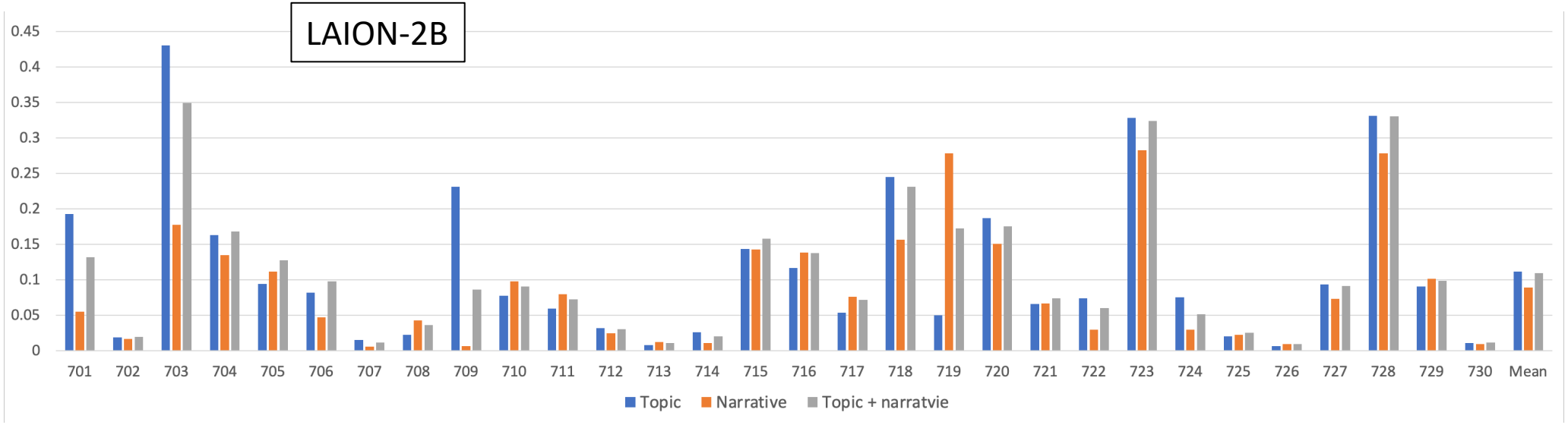


- LAION-2B achieves the best performance
- The performance of ViT-L/14 is unexpectedly the worst despite the tradition that processing larger resolution images usually leads to a better performance.

Effect of Narratives

Fusion is done by simply adding scores obtained using a topic and a narrative





- Although the narrative of a topic seems a more detailed description than the topic itself, additional use of narratives does not lead to a significant performace improvement.
- Using narratives degrades the performance for LAION-400M and LAION-2B.
- The performance trends are relatively similar in the case of using topics and the case of using narratives.

How to Utilise CLIP-based Features for Shots

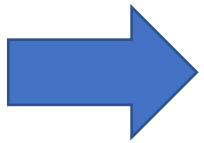
- **Keyframe only:** Represent a shot only with the feature of the keyframe
- **Average pooling:** Represent a shot with the average of features extracted for equidistantly sampled 10 frames and the keyframe
- **Max selection:** Represent a shot as a set of features for equidistantly sampled 10 frames and the keyframe, and measure its similarity to a topic as the highest frame-topic similarity

ViT-B/32	Keyframe only	0.0943 (topic: 0.0865, narrative 0.0728)
	Average pooling	**0.1207** (topic: 0.1095, narrative: 0.0892)
	Max selection	0.1199 (topic: 0.1070, narrative: 0.0893)
ViT-L/14	Keyframe only	0.0899 (topic: 0.0801, narrative: 0.0689)
	Average pooling	**0.1202** (topic: 0.1099, narrative: 0.0910)
	Max selection	0.1187 (topic: 0.1114, narrative: 0.0846)
LAION-400M	Keyframe only	0.0913 (topic: 0.1025, narrative: 0.0644)
	Average pooling	0.1047 (topic: **0.1133**, narrative: 0.0746)
	Max selection	0.0963 (topic: 0.1117, narrative: 0.0673)
LAION-2B	Keyframe only	0.1092 (topic: 0.1116, narrative: 0.089)
	Average pooling	0.1319 (topic: **0.1328**, narrative: 0.1046)
	Max selection	0.1214 (topic: 0.1247, narrative: 0.0946)

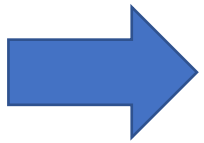
- Processing 10 frames in addition to a keyframe leads to a performance improvement
- Average pooling is better than max selection
- For LAION models, additional use of narratives causes performance degradations

Fusion of SCAN and CLIP models

Keyframe only	ViT-B/32	0.0943
	Vit-L/14	0.0899
	LAION-400M	0.0913
	LAION-2B	0.1092
Average pooling	ViT-B/32	0.1207
	Vit-L/14	0.1202
	LAION-400M	0.1047
	LAION-2B	0.1319
Max selection	ViT-B/32	0.1199
	Vit-L/14	0.1187
	LAION-400M	0.0963
	LAION-2B	0.1214



Keyframe only (Fusion of CLIP models)	0.1663
Average pooling (Fusion of CLIP models)	0.1890
Max selection (Fusion of CLIP models)	0.1757



Keyframe only with SCAN models (Submitted run)	0.199
Average pooling with SCAN models	0.1935
Max selection with SCAN models	0.1921

All fusion results converge to similar performances

Is it enough to analyse keyframes when fusing results by various models?

Conclusion and Future Work

Conclusion

Extensions of SCAN to perform semantically meaningful alignment between regions in a frame and words/phrases in a topic

1. **Revising the setting:** Conceptual Captions dataset and VinVL-based features
2. **Aligning regions with both words and phrases:** Attach-juxtapose parser and Tree-LSTM
3. **Improving the alignment consistency:** Hierarchical alignment based on a constituency tree

Expolation of CLIP models

1. Test four CLIP models, ViT-B/32, ViT-L/14, LAION-400M and LAION-2B
2. Investigate the effect of narratives and the combination of topics and narratives
3. Examine how to utilise CLIP-based features for shots

Future work

- Hyperparameter tuning for the hierarchical alignment
- More thorough performance evaluation
- Examine the necessity of processing multiple frames in a shot
- Adoption of motion and audio features

Thank you!