MAGNet: Multi-Region Attention-Assisted Grounding of Natural Language Queries at Phrase Level

Amar Shrestha, Krittaphat Pugdeethosapol, Haowen Fang, and Qinru Qiu

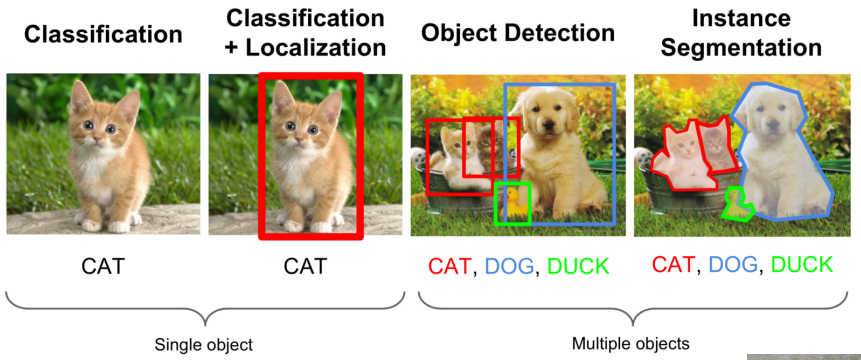
Syracuse University

Department of Electrical Engineering & Computer Science, Syracuse University, Syracuse, NY 13244, USA

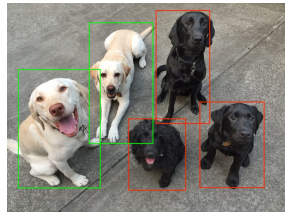
January 2021

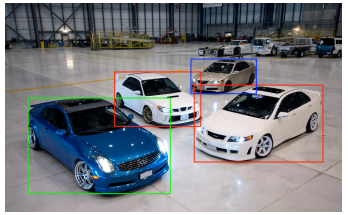
Motivation

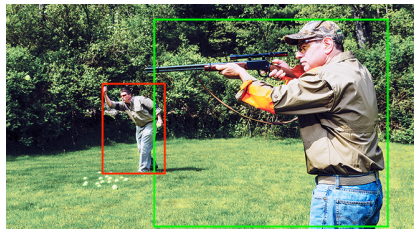
- Computer Vision Tasks
  - Classification, Classification + Localization, Object Detection, Instant Segmentation



- Limitations
  - Fix number of classes
  - The model has pre-defined number of classes.
  - Training and inferences have the same set of classes.
  - No context and description awareness

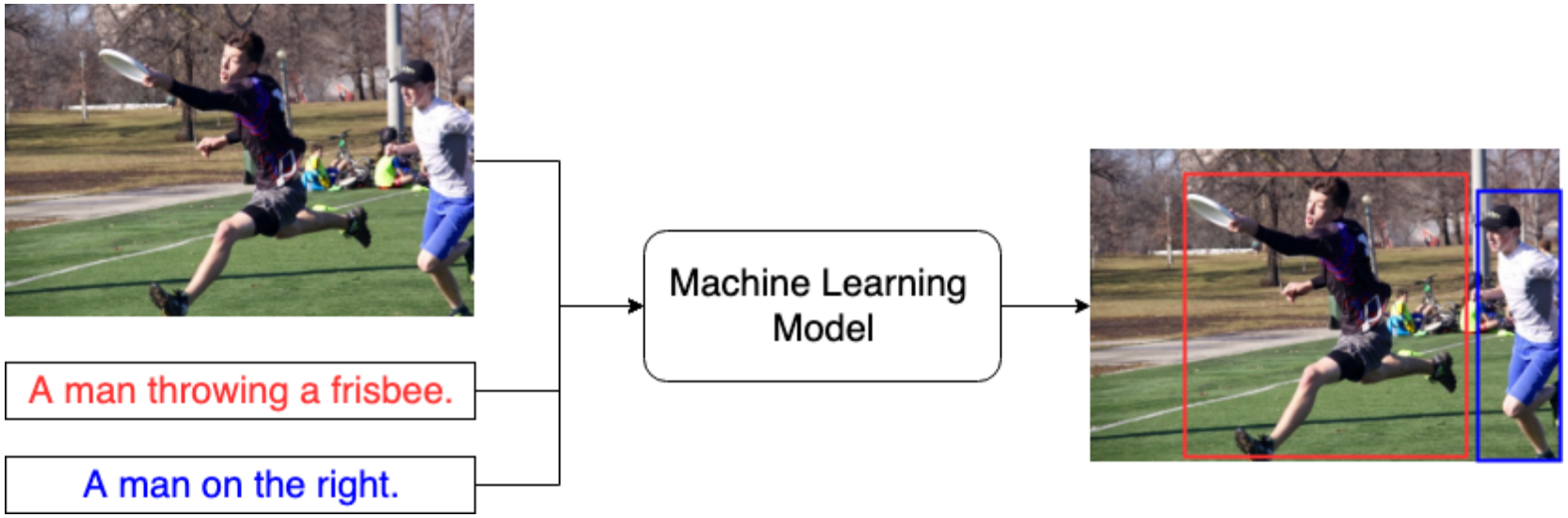




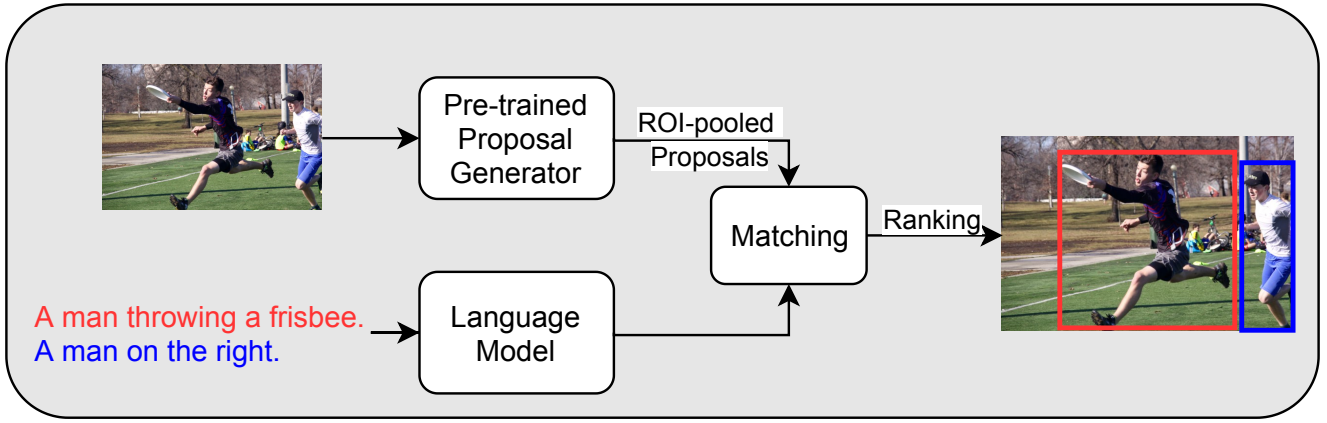


Problem Definition

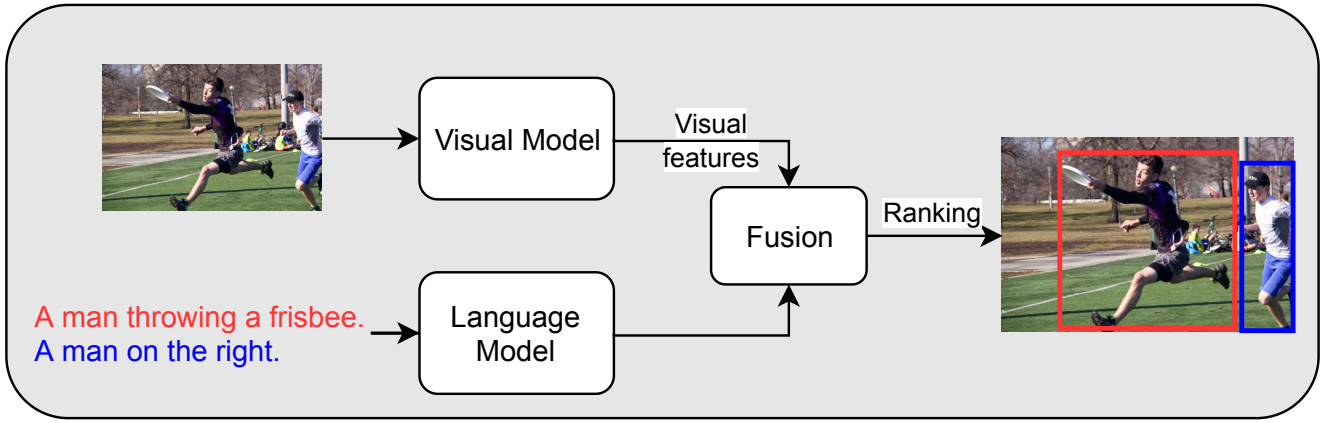
- Goal: Phrase localization
  - Use descriptive phrases to locate one or more objects in the given image.



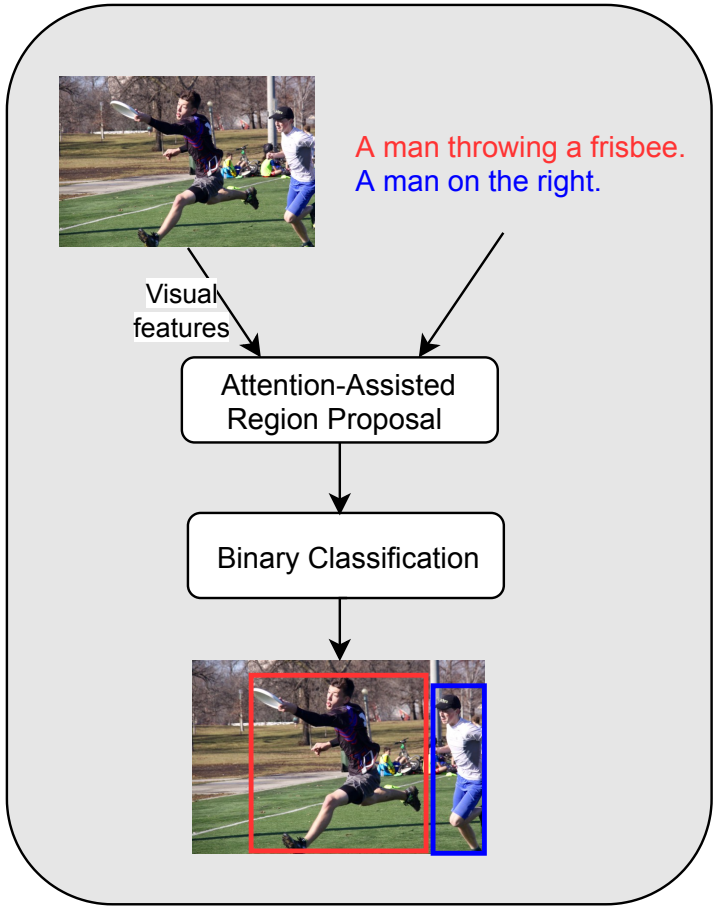
Previous Researches



Two-Stage approach

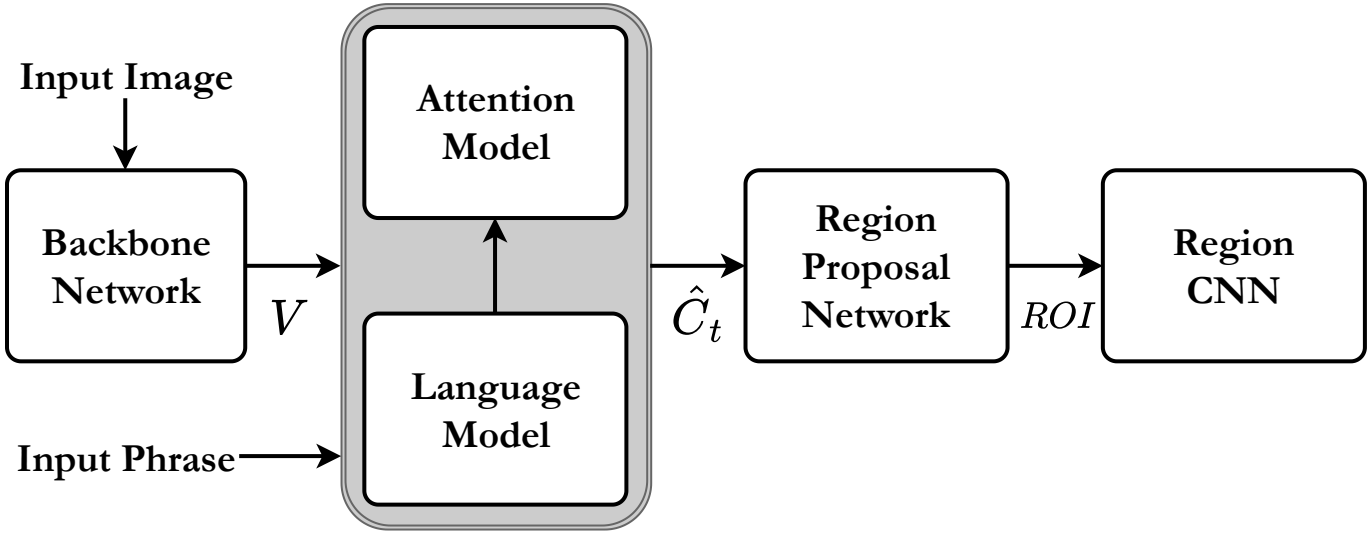


One-Stage approach



Our approach

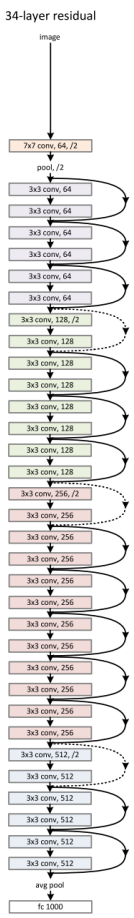
Proposed Methods



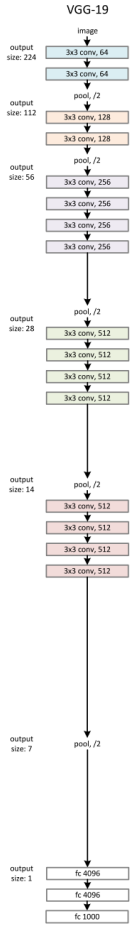
- Backbone Network
  - Understand visual features from the image
  - ResNet, VGG etc
  - Pre-trained on large image datasets (Imagenet, Pascal VOC etc)
- Language model with attention
  - Understand the input phrase in relation to the image
  - Direct attention to the model to relevant parts of the image
- Region Proposal Network
  - Propose multiple areas/objects in the image that is relevant
- Region CNN
  - Classify the proposed regions

Backbone Network

- Understand visual features from the image
- ResNet, VGGNET, etc
- Pre-trained on large image datasets (Imagenet, Pascal VOC etc)
- Features from the top conv layer is utilized
- Choice of the backbone network also depends on the size of the network

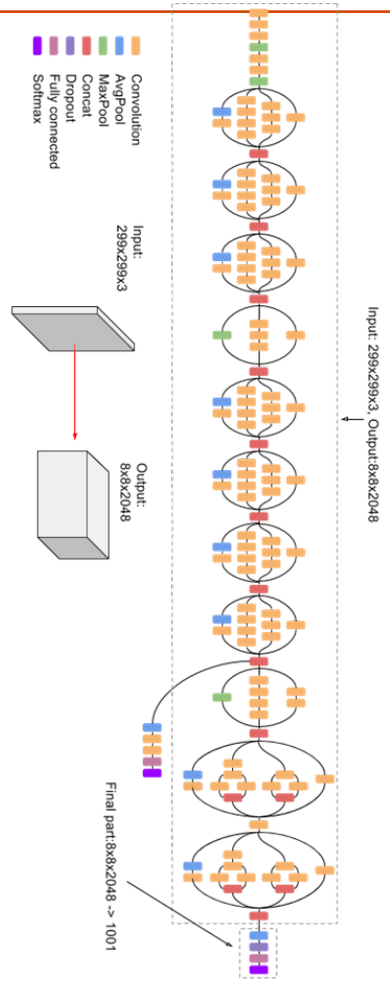


ResNet



VGGNet

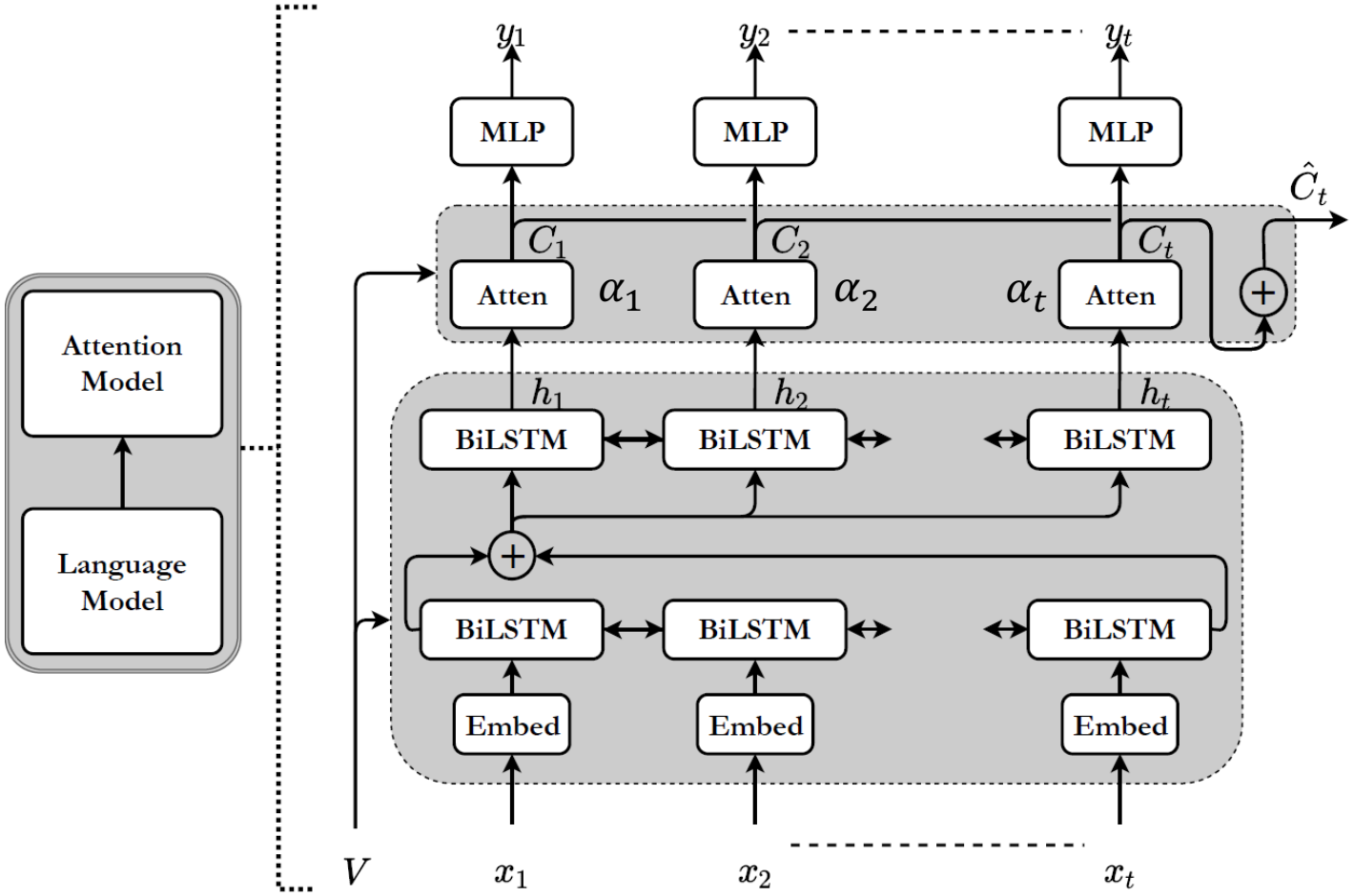
GoogleNet



Inception v3

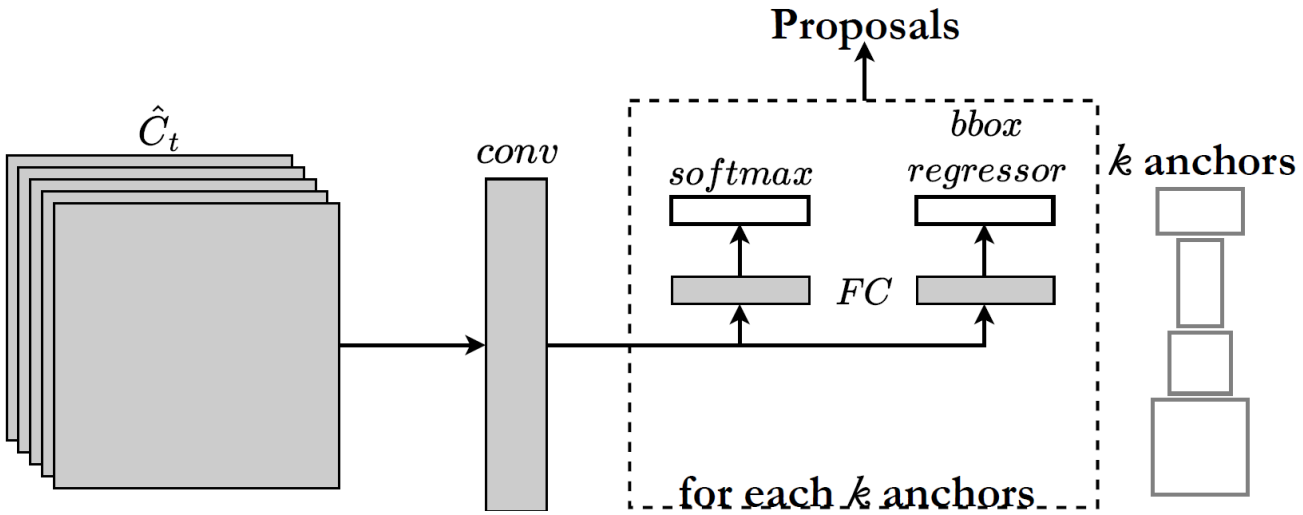
Language Model with Attention

- Understand the input phrase in relation to the image
- Direct attention of the model to relevant parts of the image
- Mixes context with the visual features used for RPN
- Attention network map C_t at each word. produces context feature
- Trained through categorical cross-entropy



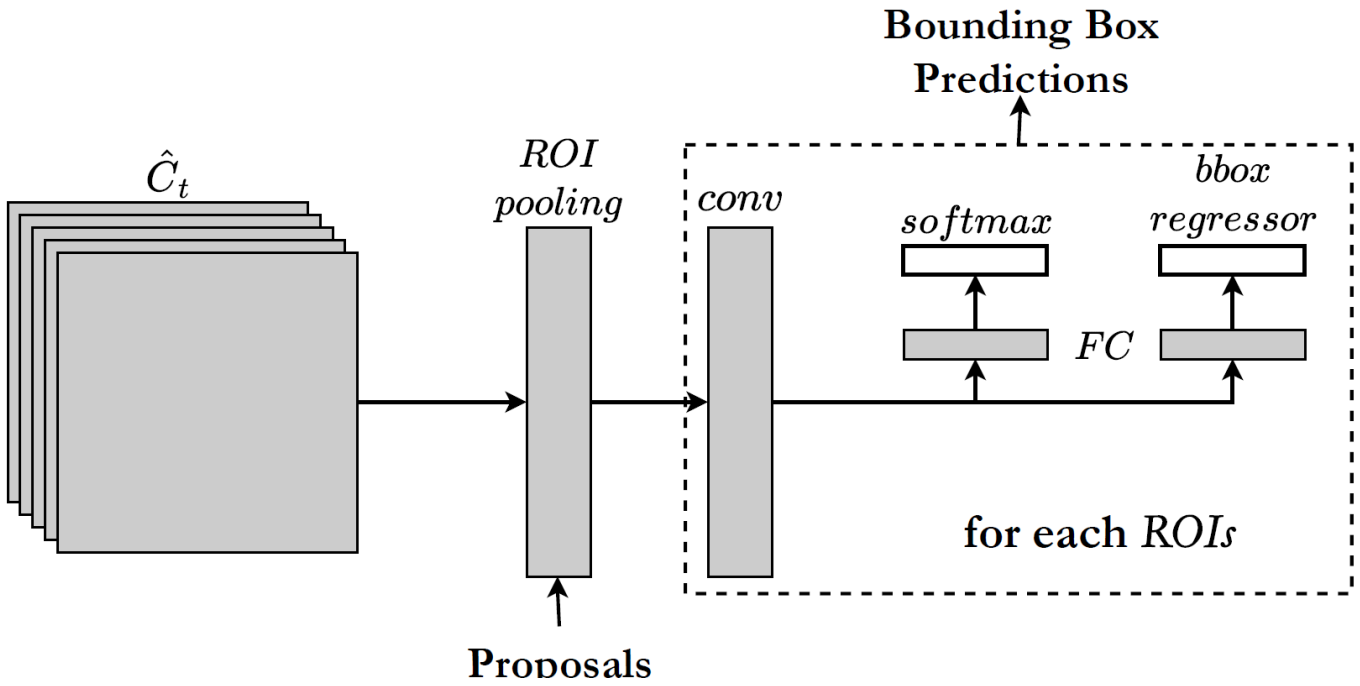
Region Proposal Network

- Proposals are generated by a small network moving a sliding window over a context feature map $\hat{C}_t$
- RPN has a classifier and a regressor. Anchor is the central point of the sliding window.
- Classifier determines the probability of a proposal having the target object.
- Regression regresses the coordinates of the proposals.



Region CNN

- After RPN, we get proposed regions with different sizes.
- Region of Interest (ROI) Pooling reduces the feature maps into the same size.
- Pooled area goes through CNN and two FC branches for softmax and bounding box regressor.
- Softmax is to classify whether it's the correct area/object given the phrase or not.



Dataset

- Flickr30
  - 31,783 images focusing mainly on people and animals, and 158,915 captions, 275,775 bounding boxes.
  - A dog on wood floor is staring at a yellow ball that is lying on orange carpet
- ReferIt
  - containing 130,525 expressions, referring to 96,654 distinct objects, in 19,894 photographs of natural scenes.
  - Right rocks, rocks along right side, stone right side of stair.
- Visual Genome
  - 108,077 Images with 5.4 Million Region Descriptions.
  - Girl feeding large elephant. A man taking a picture behind girl.

Experimental Results

	Methods	Backbone Network	Flickr30k Entities R@1	ReferItGame R@1	Visual Genome R@1
2-Stage	SCRC[6]	VGG16	27.80	17.93	11.00
	GroundeR[19]	VGG16	47.81	26.93	-
	CCA[1]	VGG19	50.89	-	-
	Similarity Net[2]	VGG19	51.05	-	-
	MSRC[28]	VGG	57.53	32.31	-
	QRN[9]	VGG	60.21	43.57	-
	QRC[9]	VGG	65.14	44.07	-
	CITE[10]	VGG16	61.89	34.13	24.43
	PIRC Net[18]	ResNet	**72.83**	59.13	-
1-Stage	IGOP[11]	VGG16	53.97	34.70	-
	SSG[12]	VGG	-	54.24	-
	ZSGNet[13]	ResNet50	63.39	59.63	-
	[14]	DarkNet53	68.69	59.30	-
	MAGNet(Ours)	ResNet50	60.20	**71.60**	**28.85**

- Flickr30k: The phrase queries extracted from the image caption ignore the context in the original sentence.
  - Our model can detect all matching objects in the image, and the one mentioned in the image caption may not necessarily have the highest score.
- ReferIt and Visual Genome: The queries are self-sufficient with specific positional cues and thus less ambiguous.

a baseball player

the lady

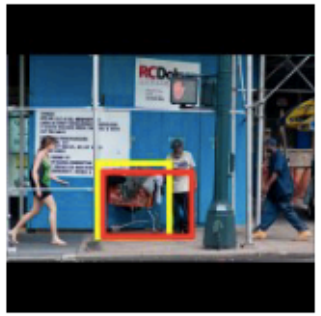
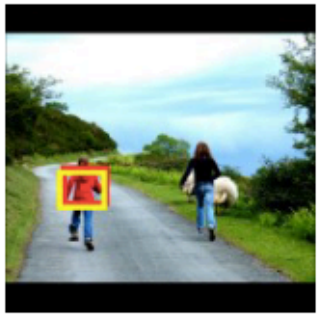
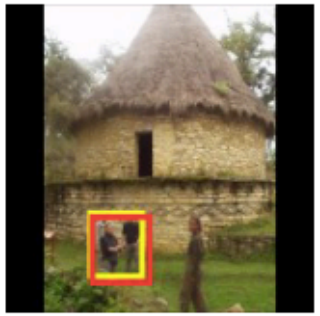
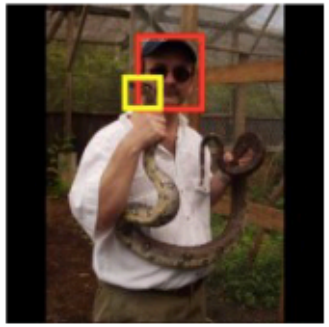
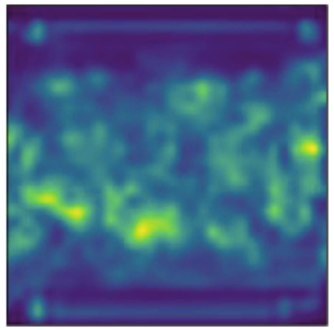
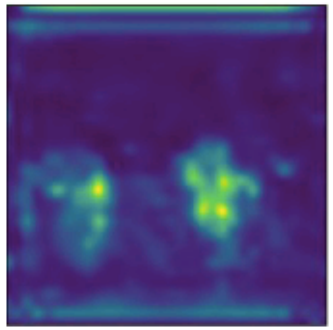
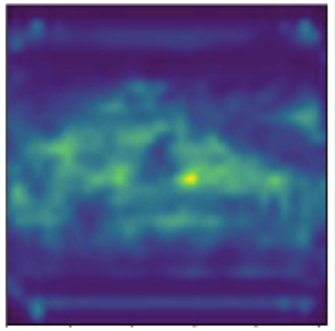
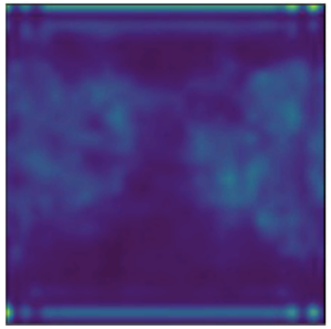
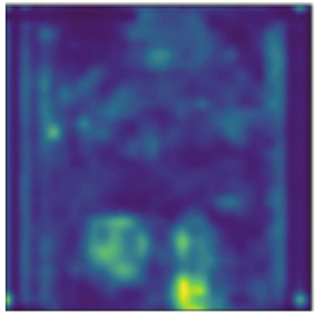
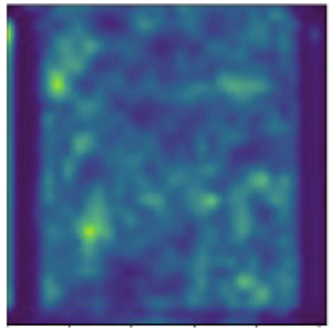
two boys

person on the left side

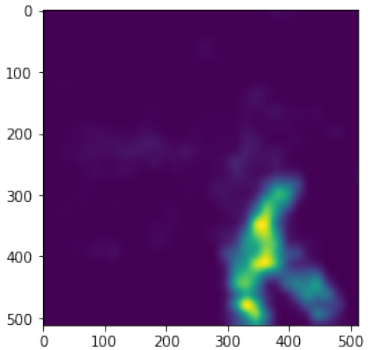
person on the right side

person in the middle

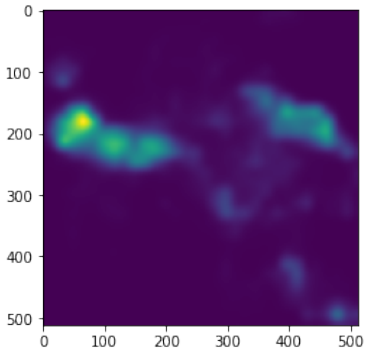
Examples

	Flickr30K			ReferItGame		
	Positive		Negative	Positive		Negative
Exps	a shopping cart	a red shirt	the load	the beige building on the front right	people to the left	damn that snake head
Predictions						
Attention						

Examples

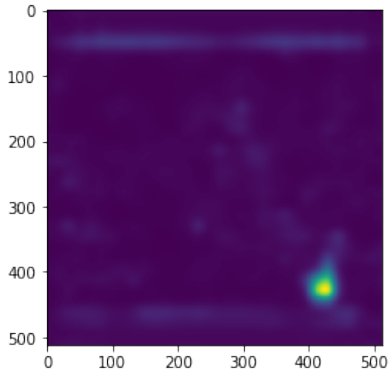


black pants



His arms

A man



Conclusion

- MAGNet framework
  - Utilize spatial attention networks for image-level visual-textual fusion preserving local (word) and global (phrase) information to refine region proposals with an in-network Region Proposal Network (RPN) and detect single or multiple regions for a phrase query.
- We can achieve respectable results in Flickr30k entities and 12% improvement over the state-of-the-art in ReferIt game.
- Our model is capable of grounding multiple regions for a query phrase, which is more suitable for real-life applications.