Strong Baselines, Evaluation, and the Role of the Humans in Grounded Language Generation

Margaret Mitchell

Shmoogle Research

Previously: Microsoft Research

2016.October

One Path to Publication

1. Implement really cool idea that you have
2. Tinker with it until it does __something sensible__
   a. Implement nonsense baseline that no one would ever use
   b. Implement impoverished version of your cool idea that obviously works less well
3. Half-hearted quick evaluation
4. Scan papers for relevant things to cite, without looking into them
5. Get paper accepted
6. Rinse, Repeat

Paper Priorities

- Get Cool Results
- Ground work within the trajectory of your field
- Ensure your claims are supported by your analyses

Paper Priorities

- Get Cool Results
- Ground work within the trajectory of your field
- Ensure your claims are supported by your analyses

→ Spend more time on scholarship than cool results

→ Teach and encourage making scholarship cool and interesting (awards?)

→ Teach and encourage good evaluation (awards?)

Prenominal modifier ordering: *the big red ball*

Methods:

- 1-pass class-based (2009)
- Multiple Sequence Alignment with Perceptron training (2010)
- HMM with EM training (2011)
- N-gram based (2011)

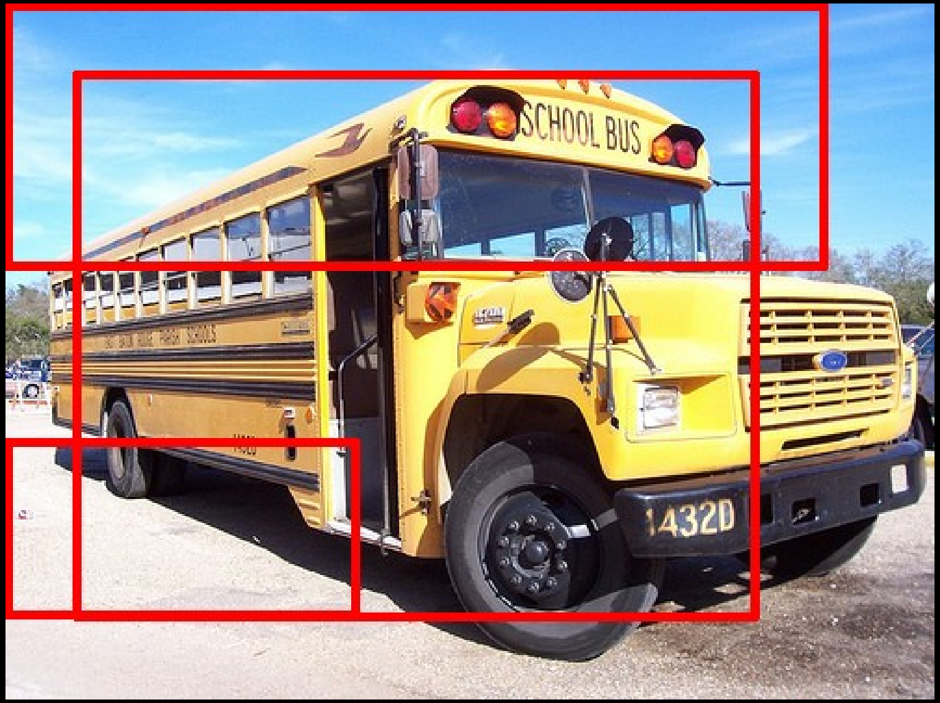
Prenominal modifier ordering: *the big red ball*

Punchline: Sophisticated models do not outperform n-gram language modelling.

Training data	WSJ Accuracy				SWBD Accuracy			
	Ngr	1-cl	HMM	MSA	Ngr	1-cl	HMM	MSA
WSJ manual	88.1	65.7	87.1	87.1	72.9	44.7	71.3	71.8
auto	87.8	64.6	86.7	87.2	72.5	41.6	71.5	71.9
NYT 10%	90.3	75.3	87.4	88.2	84.2	71.1	81.8	83.2
20%	91.8	77.2	87.9	89.3	85.2	72.2	80.9	83.1
50%	92.3	78.9	89.7	90.7	86.3	73.5	82.2	83.9
all	92.4	80.2	89.3	92.1	86.4	74.5	81.4	83.4
NYT+WSJ auto	93.7	81.1	89.7	92.2	86.3	74.5	81.3	83.4

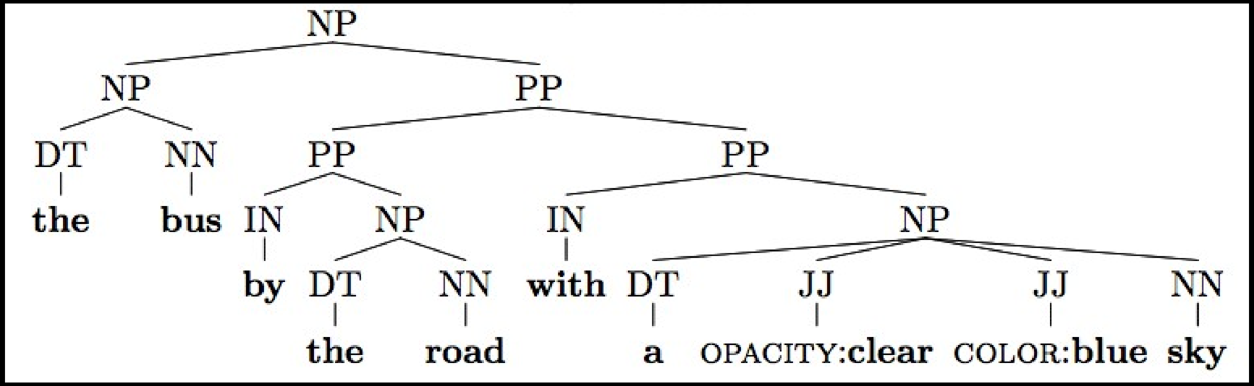
Mitchell et al., 2011, ACL

Image Captioning, 2011: Midge



The bus by the road with
a clear blue sky

stuff:	*sky*	.999	
	id:	1	
	atts:	clear:0.432,	blue:0.945
		grey:0.853,	white:0.501 ...
	b. box:	(1,1 440,141)	
stuff:	*road*	.908	
	id:	2	
	atts:	wooden:0.722	clear:0.020 ...
	b. box:	(1,236 188,94)	
object:	*bus*	.307	
	id:	3	
	atts:	black:0.872,	red:0.244 ...
	b. box:	(38,38 366,293)	



Previous work

Kulkarni et al., 2011
This is a picture of two pottedplants, one dog and one person. The black dog is by the black person, and near the second feathered pottedplant.
Yang et al., 2011
The person is sitting in the chair in the room
Midge
A person in black with a black dog by potted plants

Kulkarni et al., 2011
This is a picture of three persons, one bottle and one diningtable. The first rusty person is beside the second person. The rusty bottle is near the first rusty person, and within the colorful diningtable. The second person is by the third rusty person. The colorful diningtable is near the first rusty person, and near the second person, and near the third rusty person.
Yang et al., 2011
Three people are showing the bottle on the street
Midge
People with a bottle at the table

Evaluation

- 5-point Likert scale, from Strongly Disagree to Strongly Agree, with a neutral middle position (Reiter and Belz, 2009).

Grammaticality:
This description is **grammatically correct**.
Main Aspects:
This description **describes the main aspects** of this image.
Correctness:
This description **does not include extraneous** or incorrect information.
Order:
The objects described are mentioned in **a reasonable order**.
Humanlikeness:
It sounds like a **person wrote** this description.

Likert Scale

- “Distance” between each item category not equivalent (non-parametric), e.g., Wilcoxon Signed-Rank Test
- Composed of several **Likert Items**, which together make a scale

Evaluation

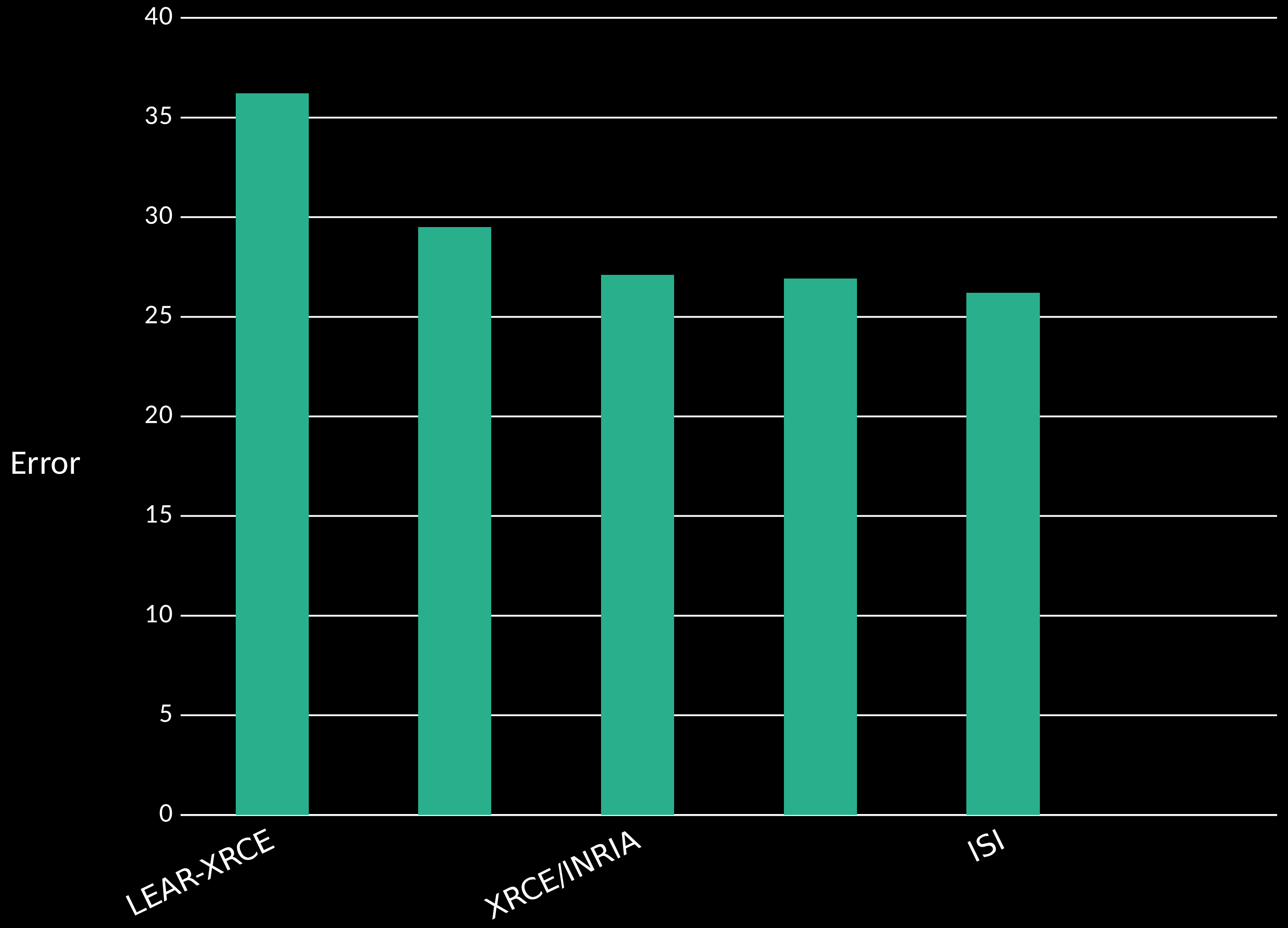
	Grammaticality	Main Aspects	Correctness	Order	Humanlikeness
Human	4 (3.77, 1.19)	4 (4.09, 0.97)	4 (3.81, 1.11)	4 (3.88, 1.05)	4 (3.88, 0.96)
Midge	3 (2.95, 1.42)	3 (2.86, 1.35)	3 (2.95, 1.34)	3 (2.92, 1.25)	3 (3.16, 1.17)
Kulkarni et al. 2011	3 (2.83, 1.37)	3 (2.84, 1.33)	3 (2.76, 1.34)	3 (2.78, 1.23)	3 (3.13, 1.23)
Yang et al. 2011	3 (2.95, 1.49)	2 (2.31, 1.30)	2 (2.46, 1.36)	2 (2.53, 1.26)	3 (2.97, 1.23)

Wilcoxon Signed-Rank Test (non-parametric)

- Midge outperforms on **Correctness** and **Order**
- Outperforms Yang et al. additionally on **Humanlikeness** and **Main Aspects**
- Midge vs. Kulkarni et al. significant at $p < .01$
- Midge vs. Yang et al. significant at $p < .001$.

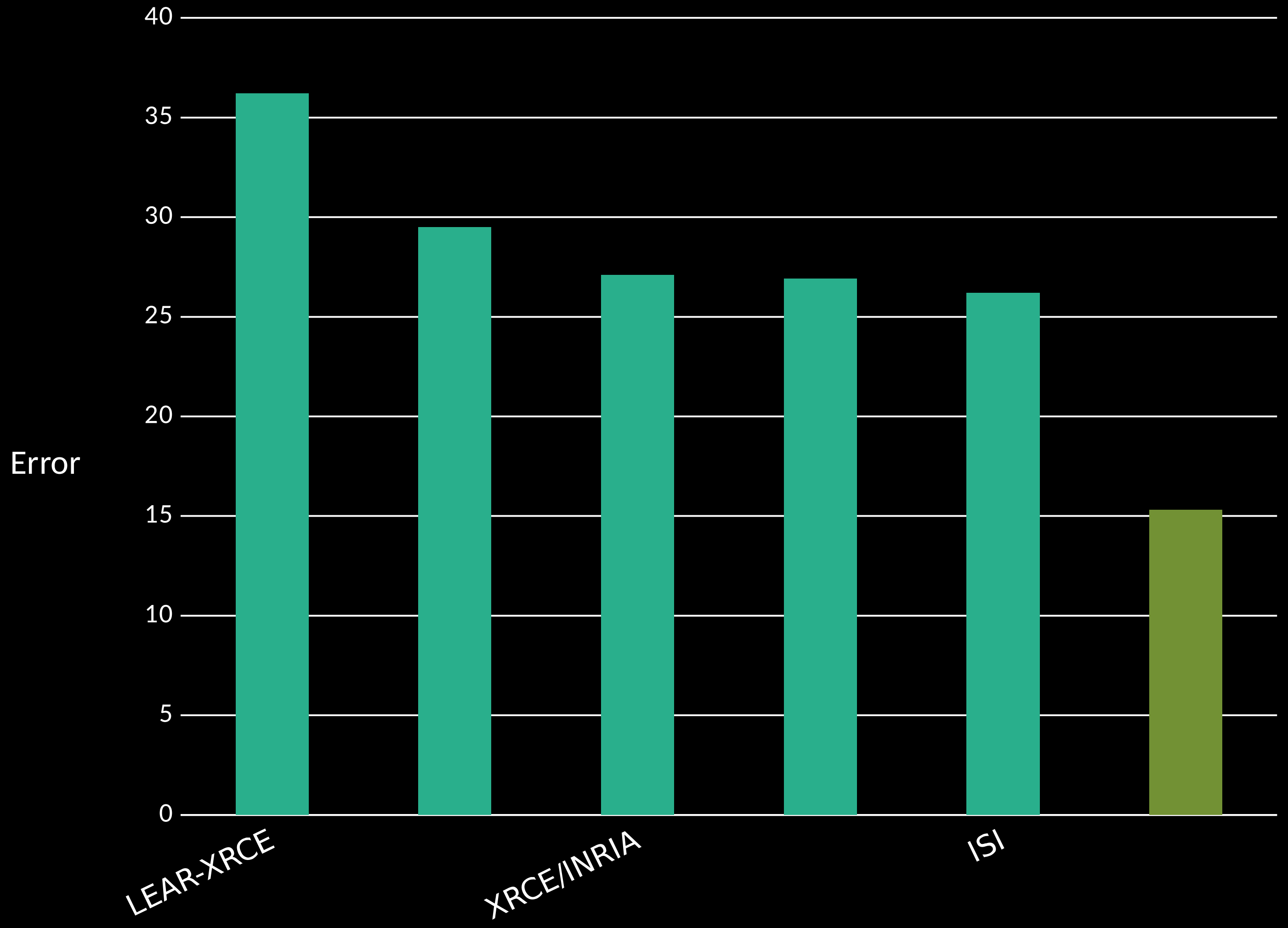
From Describing Objects to Describing Scenes 2011-2015: In the background

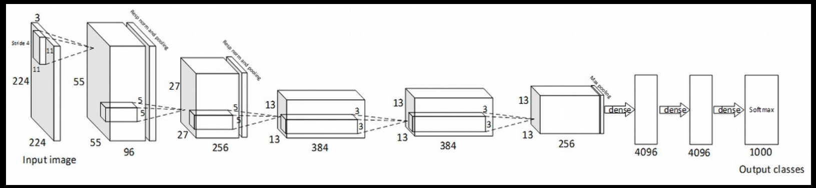
2012 ImageNet 1K Challenge



From Describing Objects to Describing Scenes 2011-2015: In the background

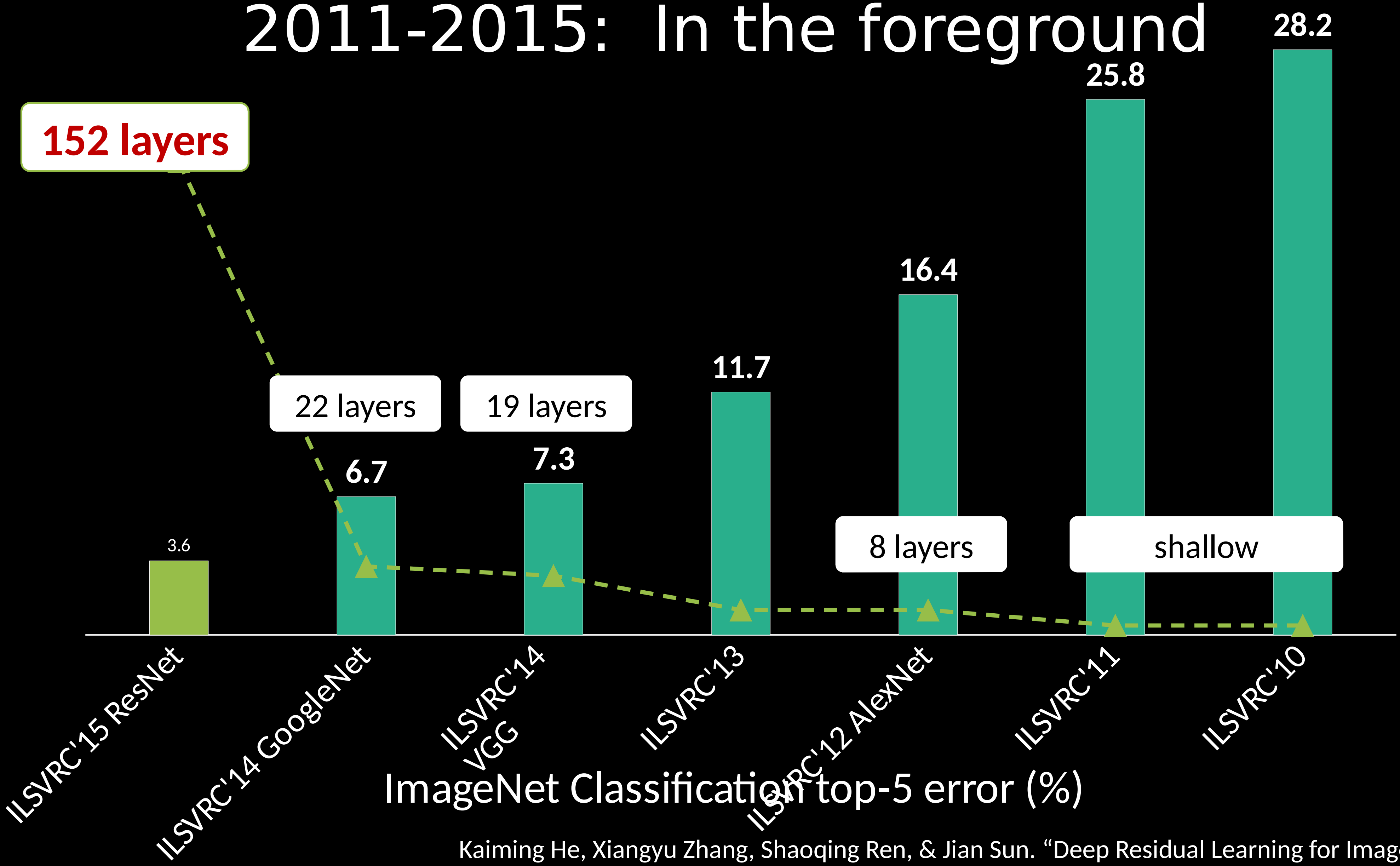
2012 ImageNet 1K Challenge





Krizhevsky, A., Sutskever, I., and Hinton, G. E. NIPS, 2012

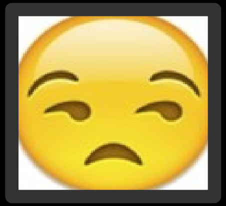
From Describing Objects to Describing Scenes 2011-2015: In the foreground



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition", 2015

Image Captioning, 2014

- Key idea: Generation of **each word** can be seen as a function of the visual scene.
  - Just use ngrams for ordering
  - (Cynical Meg)



- Why not logistic regression for each word?
  - Multinomial? == Maximum Entropy
- With tons of other MSR researchers: Combination of CNN for vision + maximum entropy + “blackboard” of detections to-be-used.
  - ➢ World’s best image captioning system.
  - ➢ Closest to human performance when evaluated by humans.

Image Captioning, 2014: More straightforward approach

- Use `fc7` as initial state in recurrent neural network language model

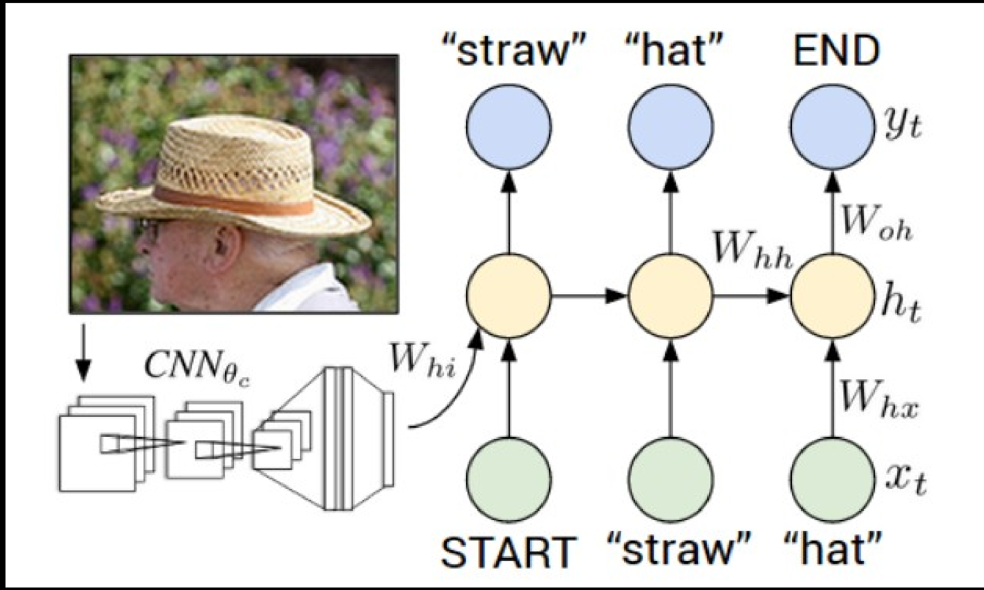


Image Credit: Karpathy and Fei-Fei 2015

Long Short Term Memory

(LSTM)

Gated Recurrent Neural Network

(GRNN)

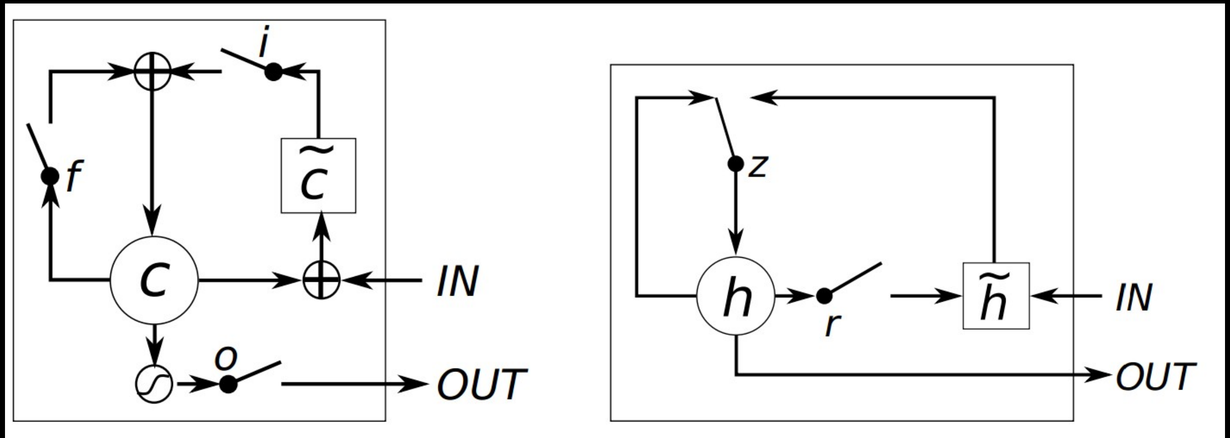


Image Credit: Cho et al. 2015

Image Captioning, 2014: Baseline -- Nearest Neighbor

1-nearest neighbor:

1. Find nearest training image based on `fc7` cosine distance
2. Output random caption from nearest neighbor

***k*-nearest neighbor**:

3. Find *k* (e.g., *k*=90) nearest training images based on `fc7` cosine distance
4. Find *consensus* caption based on *n*-gram overlap in nearest neighbor caption set

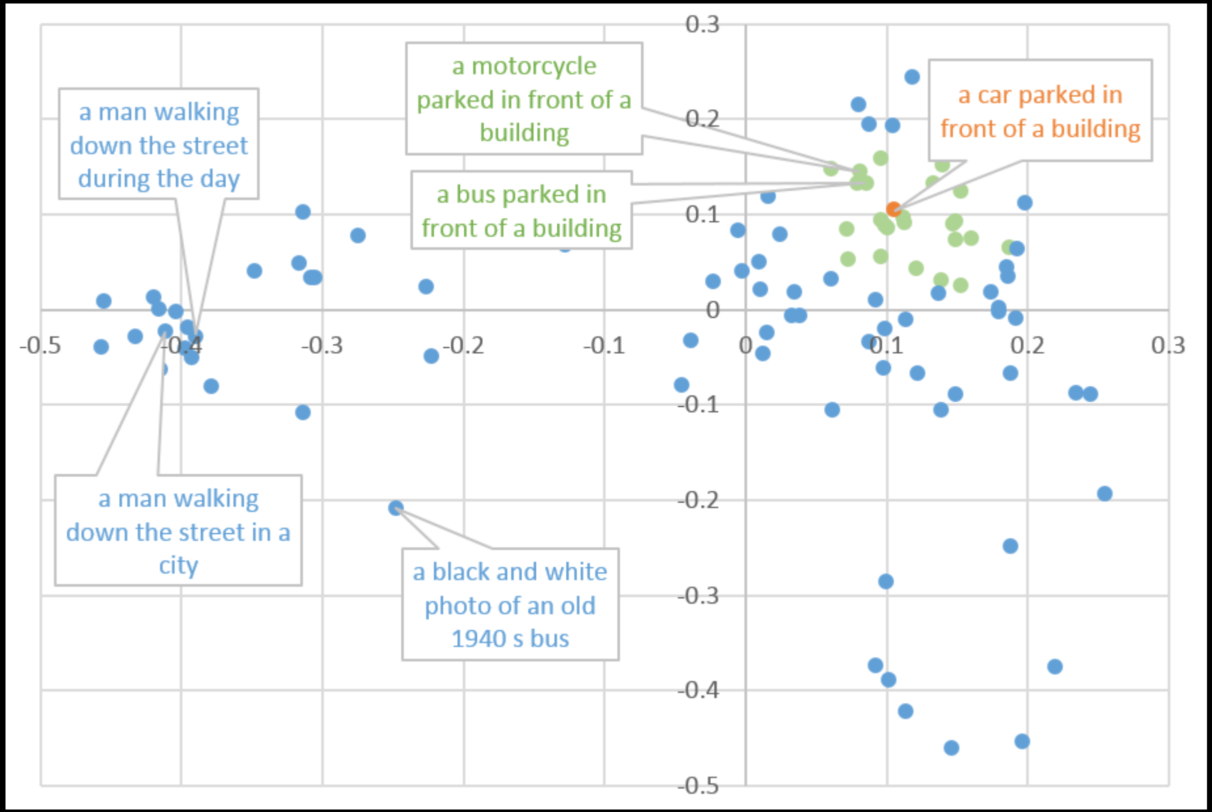


Image Captioning, 2014: More straightforward approach

- Use `fc7` as initial state in recurrent neural network language model

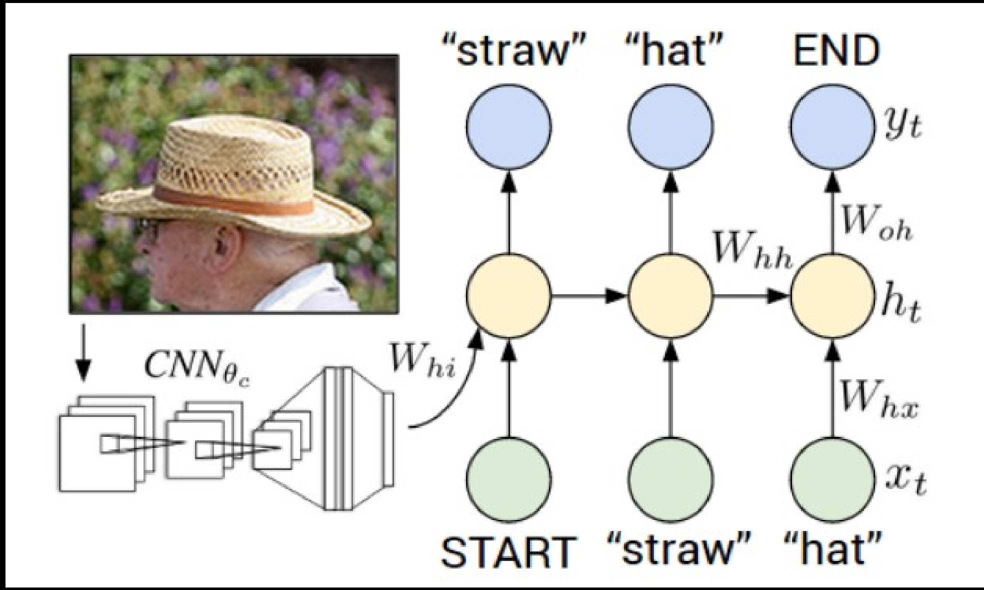


Image Credit: Karpathy and Fei-Fei 2015

Long Short Term Memory (LSTM)

Gated Recurrent Neural Network (GRNN)

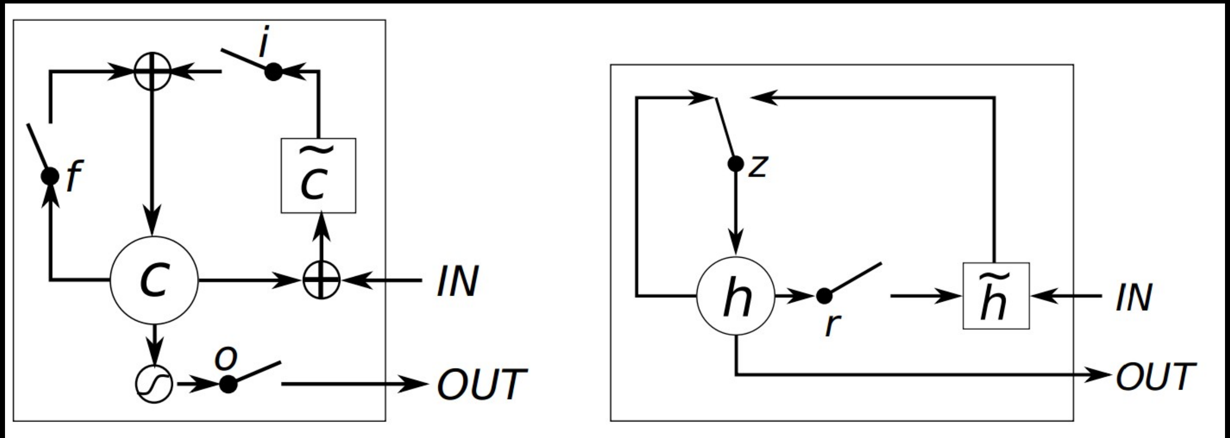
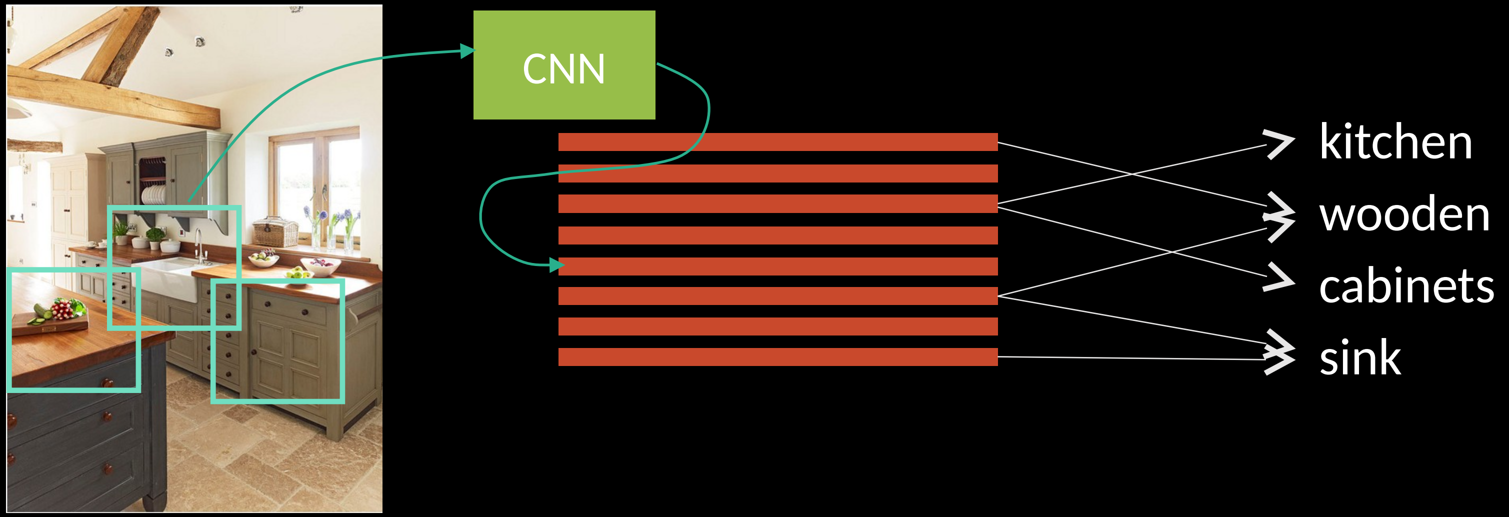


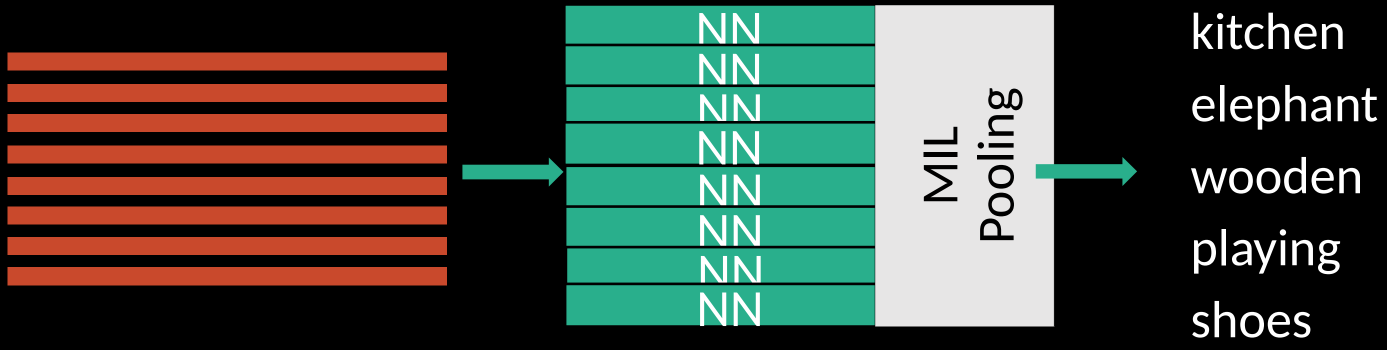
Image Credit: Cho et al. 2015

Image Analysis

Train to predict words in captions



Which words should be detected? Let a neural network figure it out



Vocabulary = the 1000 most common words in the training captions (92% of data)

(1) Enumerate regions

- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- = “bag of boxes” b_i

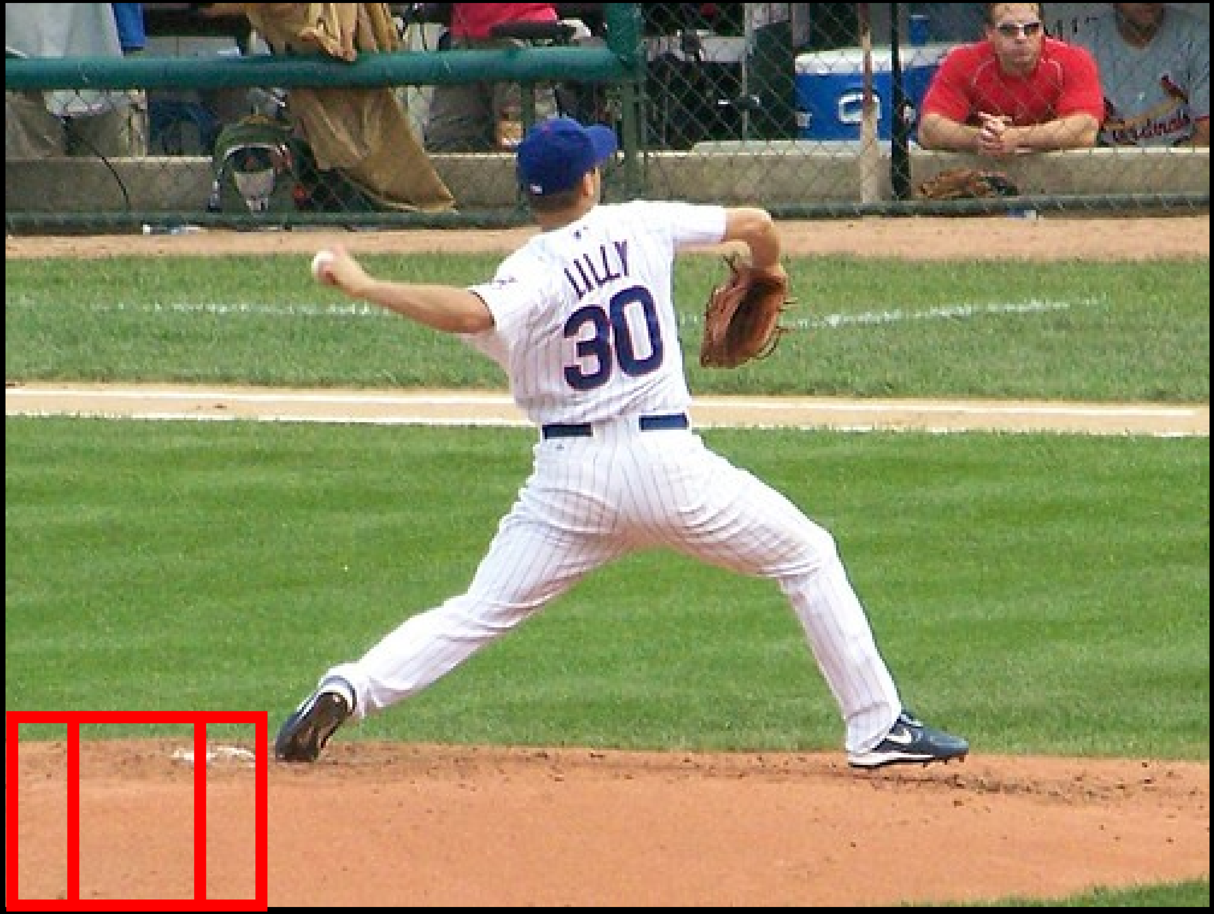
(1) Enumerate regions

- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- ≡ “bag of boxes” b_i

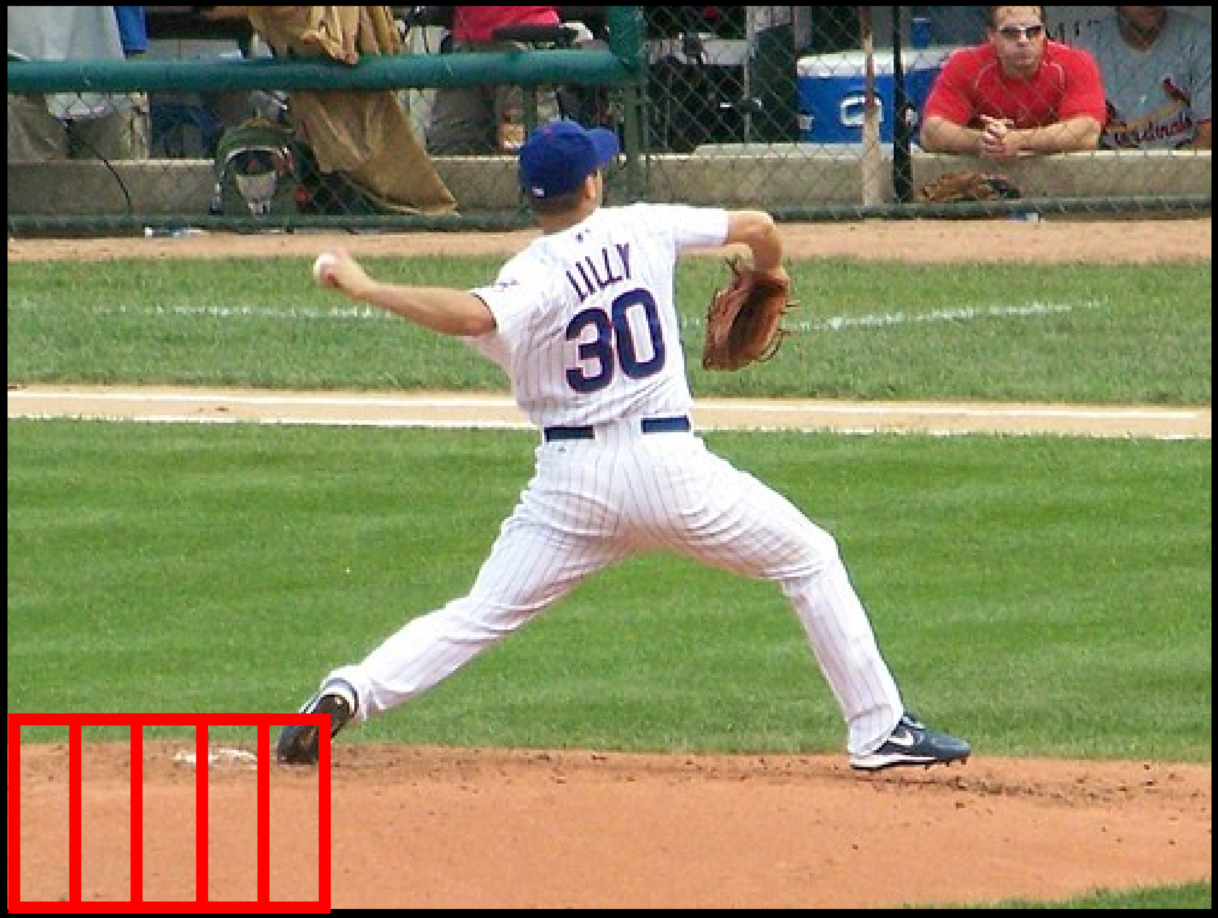
- 190 boxes per image
- “bag of boxes”

(1) Enumerate regions



- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1
- 190 boxes per image
- ≡ “bag of boxes” b_i

(1) Enumerate regions

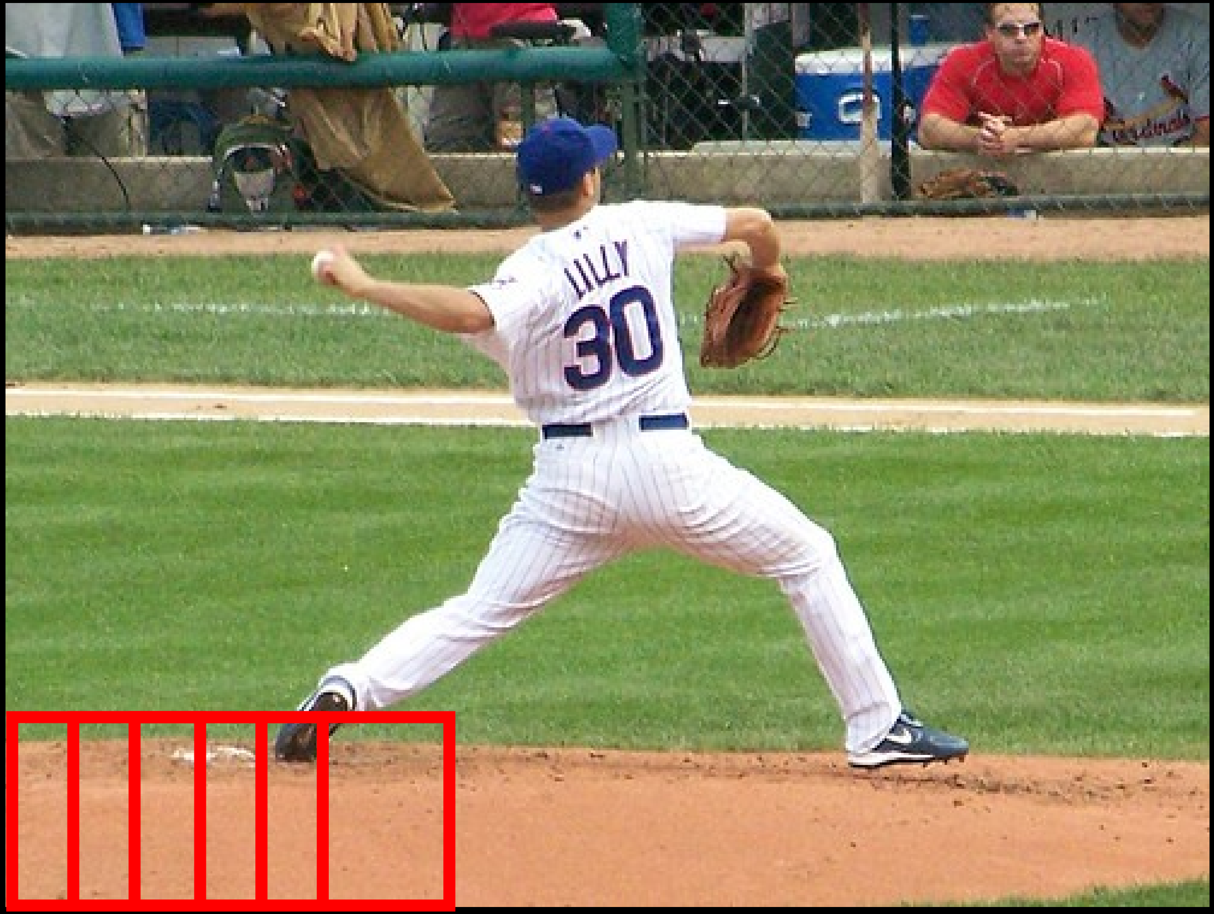


- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- ≡ “bag of boxes” b_i

190 boxes per image
“bag of boxes”

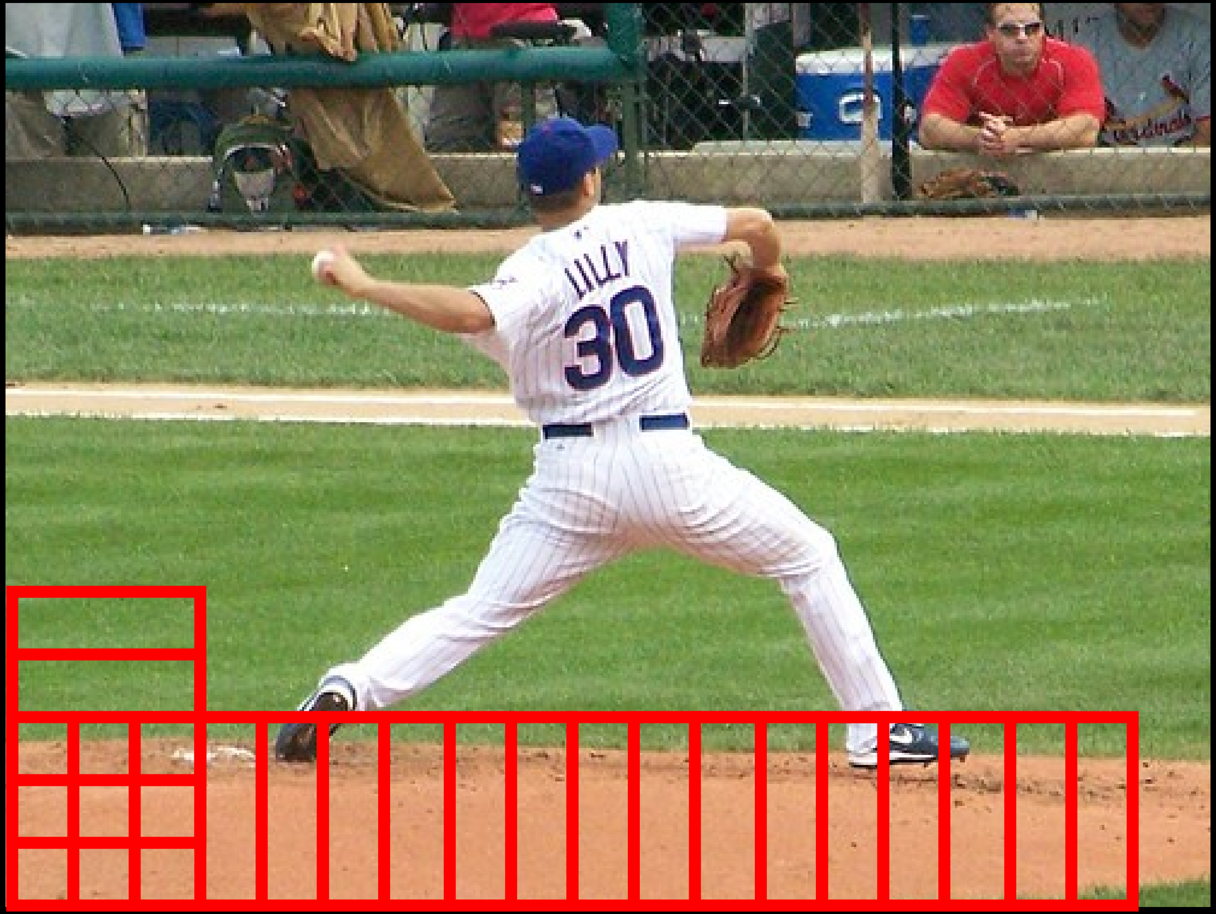
(1) Enumerate regions



- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- ≡ “bag of boxes” b_i

(1) Enumerate regions

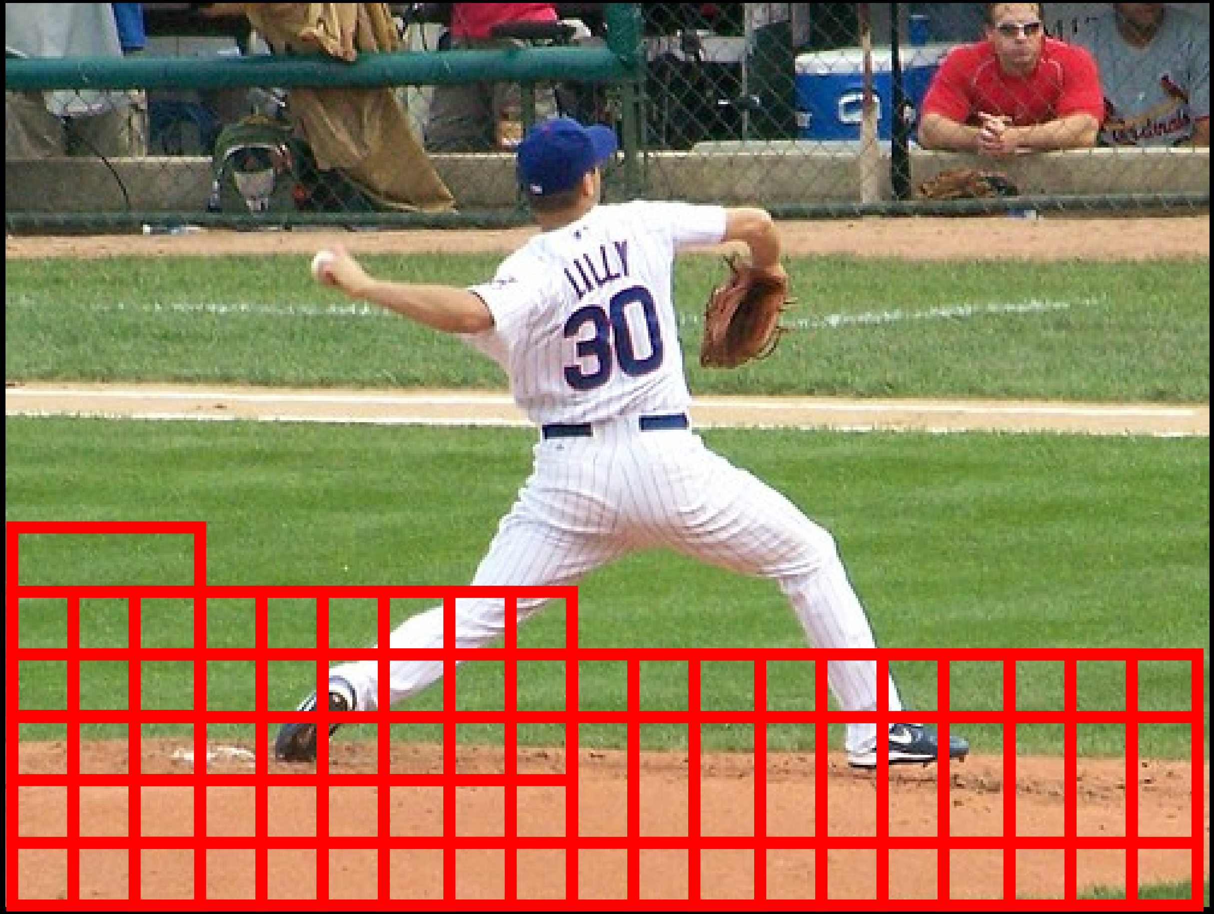


- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- ≡ "bag of boxes" b_i

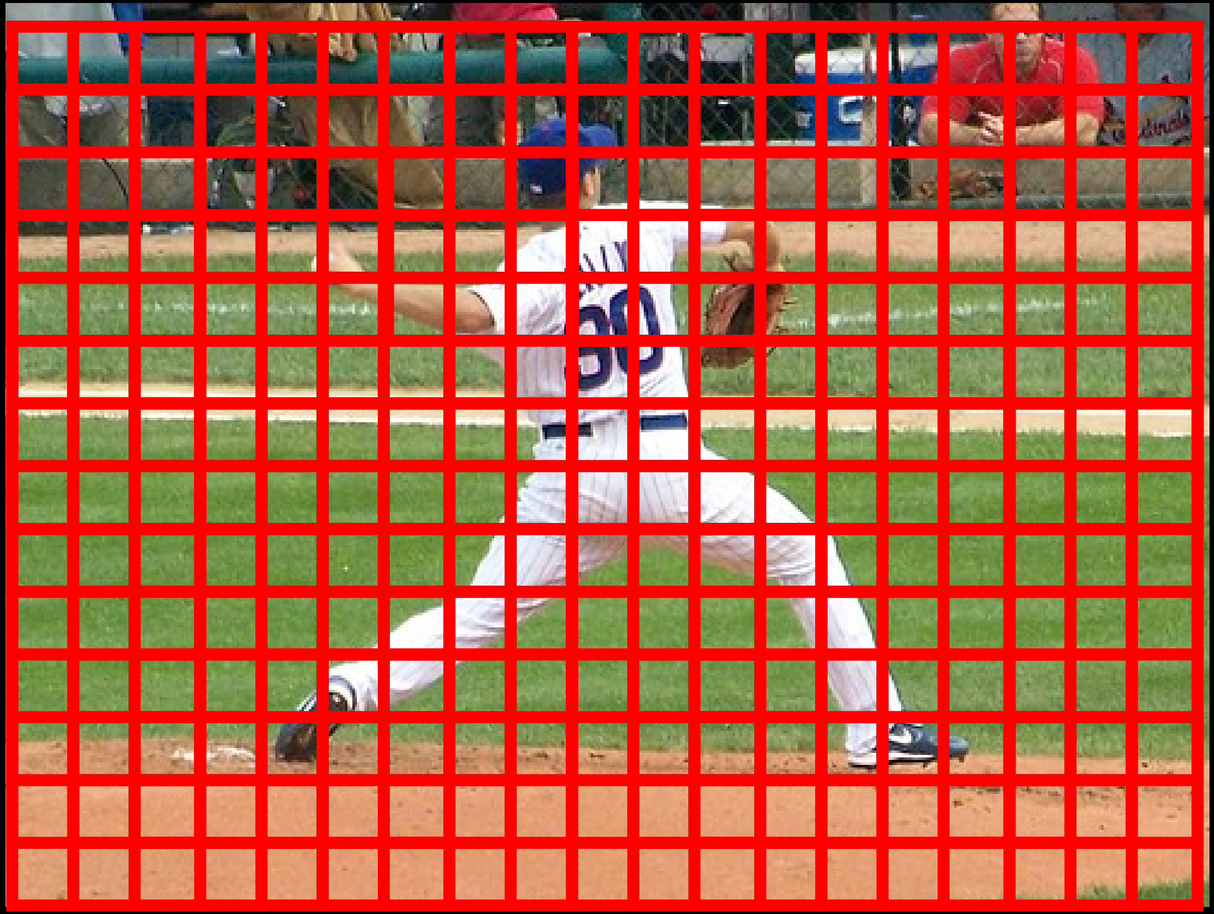
190 boxes per image
bag of boxes

(1) Enumerate regions



- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1
- 190 boxes per image
- ≡ “bag of boxes” b_i

(1) Enumerate regions

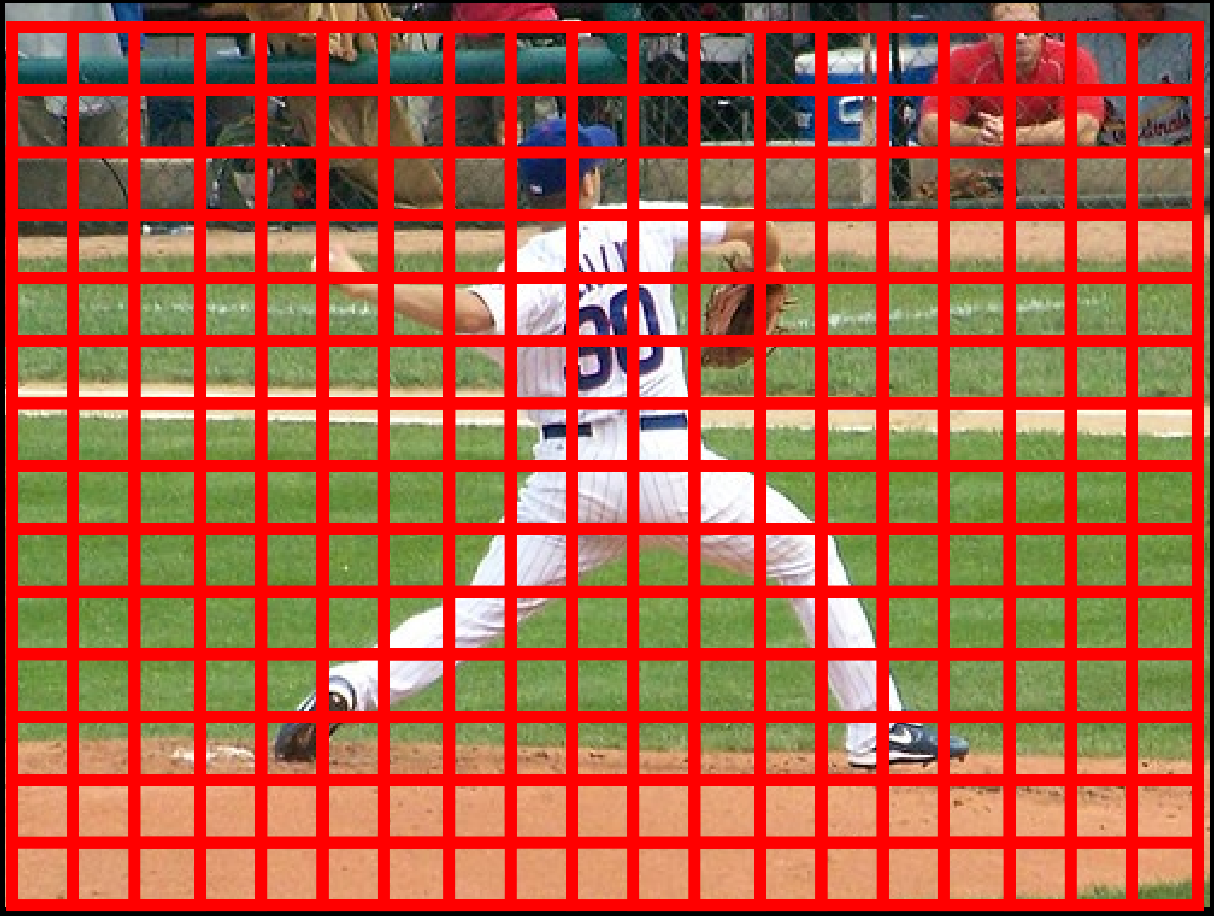


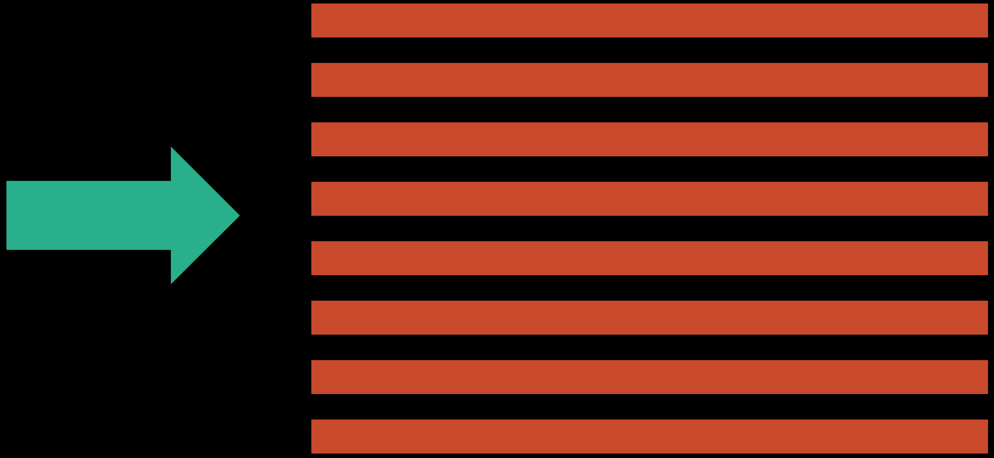
- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1

- 190 boxes per image
- ≡ “bag of boxes” b_i

(1) Enumerate regions

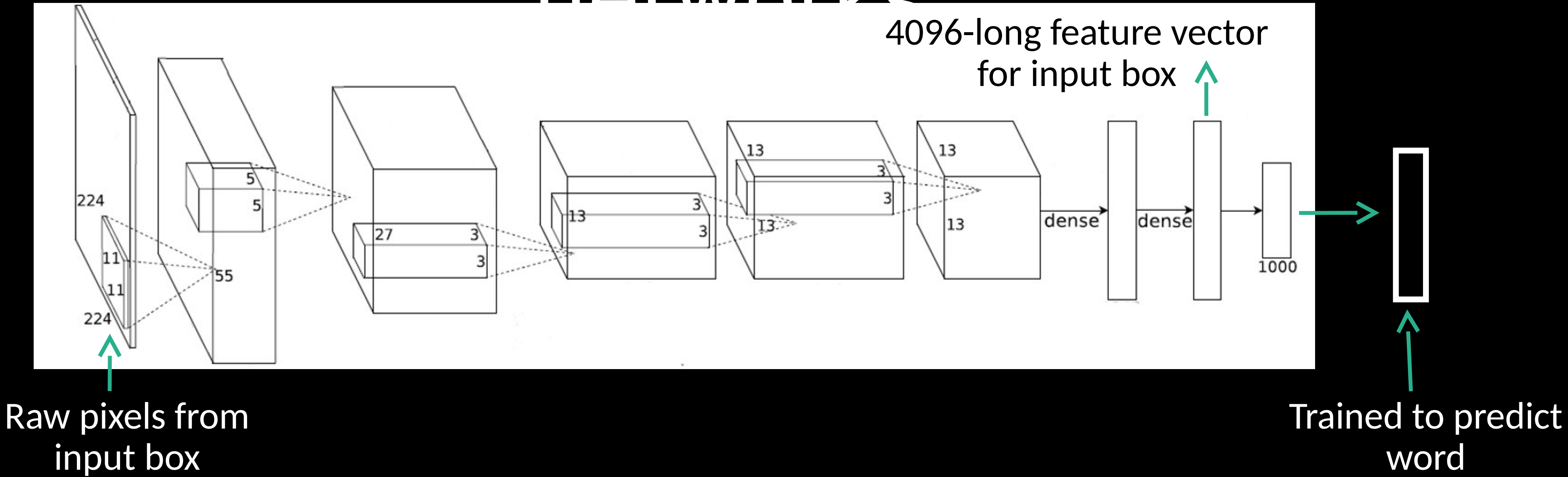
- Brute force enumeration
- Image made into a 565x565 square and fixed-size boxes run over the image
  - Sampled at different scales
  - 12x12, 6x6, 3x3, 1x1





- 190 boxes per image
- ≡ "bag of boxes" b_i

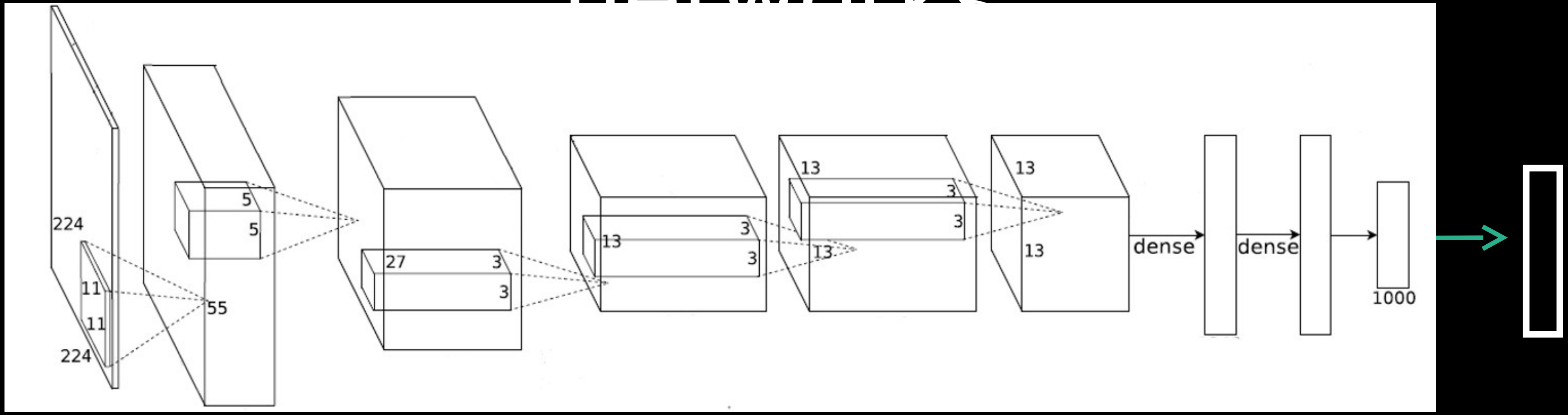
(2) Features from convolutional networks



- Pretrained from ImageNet (Krizhevsky et al., 2012) & finetuned
- For each word *w*, box *j*, image *i*, compute $p_{ij}(w)$:

$$p_{ij}(w) = \frac{1}{1 + \exp\left(-v_w \phi(b_{ij}) - u_w\right)}$$

(2) Features from convolutional networks



- Pretrained from ImageNet (Krizhevsky et al., 2012) & finetuned
- For each word *w*, box *j*, image *i*, compute $p_{ij}(w)$:

$$p_{ij}(w) = \frac{1}{1 + \exp\left(-v_w \phi(b_{ij}) - u_w\right)}$$

weights box bias

(3) Map features to likely image words

- Train with Noisy-OR Multiple Instance Learning (MIL)
- For each word w, MIL uses positive and negative bags of bounding boxes
  - For each image i:
    - We have the "bag of boxes", b_{ij}
    - b_{ij} is **positive** if w in i's description
    - b_{ij} is **negative** if w not in i's description
  - Probability that image i manifests word w, $p_{ij}(w)$↓↑:

- Train with Noisy-OR Multiple Instance Learning (MIL)
- For each word w, MIL uses positive and negative bags of bounding boxes
  - For each image i:
    - We have the "bag of boxes", b_{ij}
    - b_{ij} is **positive** if w in i's description
    - b_{ij} is **negative** if w not in i's description
  - Probability that image i m

$$1 - \prod_{j \in b_i} \left(1 - p_{ij}^{w}\right)$$

Each bounding box in image

Calculated from CNN (last slide)

(3) Map features to likely image words

- We use p_{ij}^{w} to compute global precision threshold τ on held-out training subset
  - Output all words $\tilde{V}$ with precision of τ or higher

- We use to compute global precision threshold τ on held-out training subset
- Output all words with precision of τ or higher

cat

baseball

red

	Metric	Noun	Verb	Adjective
Human Agreement	PHR	63.8	35.1	35.9
Classification	PHR	45.3	31.0	37.1
MIL NOR	PHR	**51.6**	**33.3**	**44.3**

	Metric	Noun	Verb	Adjective
Chance	AP	2.0	2.3	2.5
Classification	AP	37.0	19.4	22.5
MIL NOR	AP	**41.4**	**20.7**	**24.9**

Language generation

Language models learn to babble

Probability of a word depends on previous words

Probability model

Language model

Nay, I know not:
Is by a sleep to say we end
The ratifiers and props of every word,
They are not the trail of policy so sure
As hush as death, anon the dreadful thunder
Doth all the days i' the church.

Language models learn to babble

Probability of a word depends on previous words

Probability model

Malkovich

Language model

The Malkovich of Malkovich is not Malkovich: it droppeth like the gentle Malkovich from Malkovich

Keep track of what you want to say

Previous words

New word

Probability model

A kitchen with wooden cabinets and a sink

wooden

kitchen

sink

cabinets

remove

Maximum Entropy Language Model

Word probability:

$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup \text{</s>}} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1})\right]}$$

Sentence end

Sentence start

Feature	Type	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram +	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is in the attribute set.
N-gram -	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \notin \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is not in the attribute set.
End	0/1	$\bar{w}_l = \kappa$ and $\tilde{\mathcal{V}}_{l-1} = \emptyset$	The predicted word is κ and all attributes have been mentioned.
Score	$\mathbb{R}$	score($\bar{w}_l$) when $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	The log-probability of the predicted word when it is in the attribute set.

Maximum Entropy Language Model

Word probability:

$$\Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1}) = \frac{\exp\left[\sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1})\right]}{\sum_{v \in \mathcal{V} \cup \text{</s>}} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \cdots, \bar{w}_1, \text{<s>}, \tilde{\mathcal{V}}_{l-1})\right]}$$

Feature	Type	Definition	Description
Attribute	0/1	$\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	Predicted word is in the attribute set, i.e. has been visually detected and not yet used.
N-gram +	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is in the attribute set.
N-gram -	0/1	$\bar{w}_{l-N+1}, \cdots, \bar{w}_l = \kappa$ and $\bar{w}_l \notin \tilde{\mathcal{V}}_{l-1}$	N-gram ending in predicted word is κ and the predicted word is not in the attribute set.
End	0/1	$\bar{w}_l = \kappa$ and $\tilde{\mathcal{V}}_{l-1} = \emptyset$	The predicted word is κ and all attributes have been mentioned.
Score	$\mathbb{R}$	score($\bar{w}_l$) when $\bar{w}_l \in \tilde{\mathcal{V}}_{l-1}$	The log-probability of the predicted word when it is in the attribute set.

All sentences

Sentence length

Objective:

$$L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{\#(s)} \log \Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \cdots, \bar{w}_1^{(s)}, \text{<s>}, \tilde{\mathcal{V}}_{l-1}^{(s)})$$

Generation Process

- Perform left-to-right beam search (Ratnaparkhi, 2000)
  - Maintain stack of *l* partial hypotheses
  - Extend with likely words, prune to top (*k*=200) paths
  - Generate until </s> is generated
    - Give up once you hit sentence length *L*=20
- Form a M-best list (*M*=500)
  - Add all sequences covering at least *T*=10 concepts
  - If less than M sequences, decrement T; repeat until M sequences

Linear regression based ranker

- Minimum Error Rate Training (MERT) uses linear combination of features
- Trained on M-best lists using BLEU

1. The log-likelihood of the sequence.
2. The length of the sequence.
3. The log-probability per word of the sequence.
4. The logarithm of the sequence's rank in the log-likelihood.
5. 11 binary features indicating whether the number of mentioned objects is x ($x = 0, \ldots, 10$).
6. The DMSM score between the sequence and the image.

Test metrics

Test on held-out set

- Images + captions unseen by training algorithms

Three different metrics

- **BLEU**
  - Machine translation quality metric
  - Measures overlap between system-produced captions and human-written ones
- **METEOR**
  - Quality metric similar to BLEU
  - Found to correlate better with human-perceived quality metrics
- **Human preference**
  - Ask Mturkers blind taste test: system better, human caption better, or are they of equal quality?

Results

System	PPLX	BLEU	METEOR	≈human	>human	≥human
1. Unconditioned	24.1	1.2%	6.8%			
2. Shuffled Human	–	1.7%	7.3%			
3. Baseline	20.9	16.9%	18.9%	9.9% (±1.5%)	2.4% (±0.8%)	12.3% (±1.6%)
4. Baseline+Score	20.2	20.1%	20.5%	16.9% (±2.0%)	3.9% (±1.0%)	20.8% (±2.2%)
5. Baseline+Score+DMSM	20.2	21.1%	20.7%	18.7% (±2.1%)	4.6% (±1.1%)	23.3% (±2.3%)
6. Baseline+Score+DMSM+ft	19.2	23.3%	22.2%	–	–	–
7. VGG+Score+ft	18.1	23.6%	22.8%	–	–	–
8. VGG+Score+DMSM+ft	18.1	25.7%	23.6%	26.2% (±2.1%)	7.8% (±1.3%)	**34.0%** (±2.5%)
Human-written captions	–	19.3%	24.1%			

* we use 4 references when measuring BLEU and METEOR, while the official COCO eval server uses 5 references.

- Compared to human, our system is better or equal 34% of the time.
- DMSM gives additional 2.1 pt BLEU (8 vs. 7) over a strong system.

Turns out this works **really well**.

- COCO server hosted evaluation on unseen data
- 15 competing systems (Berkeley, Stanford, Google, Baidu, Toronto…)

	CIDEr	Meteor	ROUGE-L	BLEU1	BLEU2	BLEU3	BLEU4
MSR Captivator	0.937	0.339	0.68	0.907	0.819	0.71	0.601
Google	0.946	0.346	0.682	0.895	0.802	0.694	0.587
Baidu/UCLA m-RNN	0.896	0.32	0.668	0.89	0.801	0.69	0.578
MSR	0.925	0.331	0.662	0.88	0.789	0.678	0.567
MSR Nearest Neighbor	0.916	0.318	0.648	0.872	0.77	0.655	0.542
Berkeley LRCN	0.891	0.322	0.656	0.871	0.772	0.653	0.534
Montreal/Toronto	0.878	0.323	0.651	0.872	0.768	0.644	0.523
Human	0.91	0.335	0.626	0.88	0.744	0.603	0.471
Stanford NeuralTalk	0.692	0.28	0.603	0.828	0.701	0.566	0.446
Brno University	0.536	0.252	0.509	0.716	0.541	0.392	0.278

Turns out this works **really well**.

- 1st place at CVPR image captioning challenge
- Evaluated by humans

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Turns out this works **really well**.

- 1st place at CVPR image captioning challenge
- Evaluated by humans

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	**Percentage of captions that pass the Turing Test.**
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	**Percentage of captions that are similar to human description.**

Turns out this works **really well**.

- Also, 2nd place at CVPR image captioning challenge
  - When we add in GRNN forced decoding

Metric	Description
M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	**Average correctness of the captions on a scale 1-5 (incorrect - correct).**
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

Language Analysis

- **GRU-NN weakness:** Long-distance language modelling

MELM + DMSM	GRU-NN
a slice of pizza sitting on top of it	a bed with a red blanket on top of it
a black and white bird perched on top of it	a birthday cake with candles on top of it

- **GRU-NN weakness:** Repeated emissions

MELM + DMSM	GRU-NN
a large bed sitting in a bedroom	a bedroom with a bed and a bed
a man wearing a bow tie	a man wearing a tie and a tie

Devlin, J. and Cheng, H. and Fang, H. and Gupta, S. and Deng, L. and He, X. and Zweig, G. and Mitchell, M. (2015). Language Models for Image Captioning: The Quirks and What Works. *Proceedings of ACL 2015*.

Language Analysis

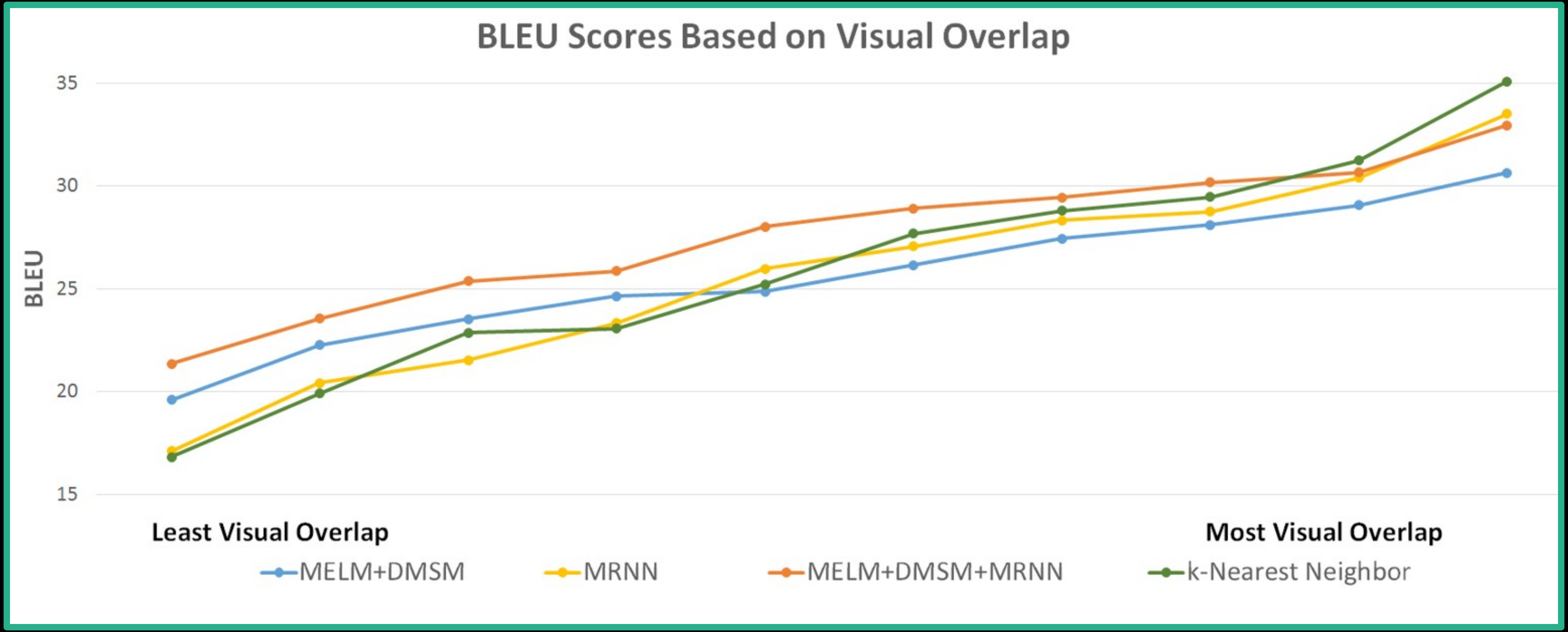
- **MRNN & *k*-NN weakness**: Repeated captions

MELM + DMSM	MRNN
a plate with a sandwich and a cup of coffee	a close up of a plate of food

System	Unique Captions	Seen In Training
Human	99.4%	4.8%
MELM + DMSM	47.0%	30.0%
MRNN	33.1%	60.3%
MELM + DMSM + MRNN	28.5%	61.3%
***k*-Nearest Neighbor**	36.6%	100%

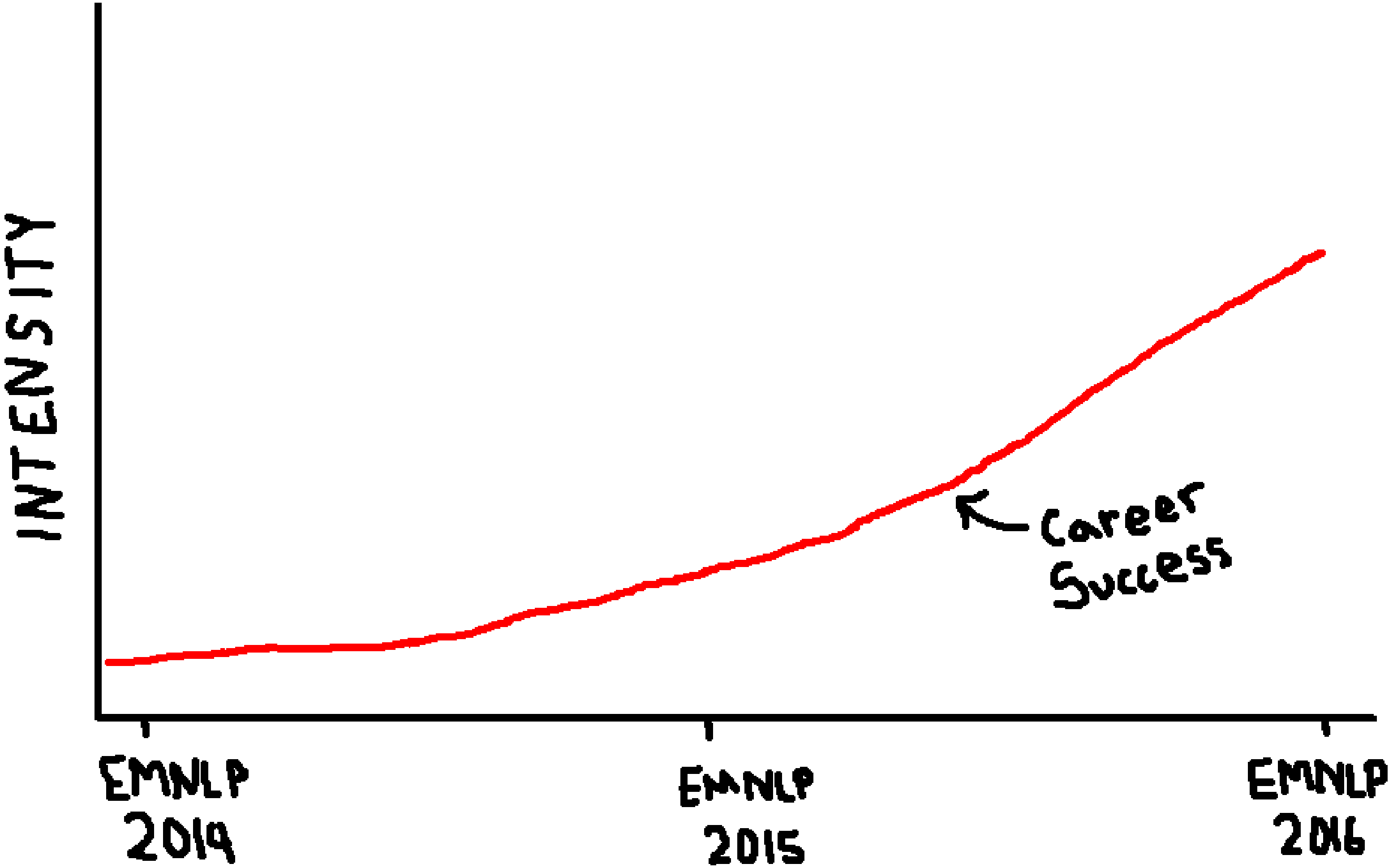
Image Diversity

- Bin test images based on visual overlap with training
- MELM + DMSM does well on images with low overlap
- MRNN/*k*-Nearest Neighbor does well on images with high overlap

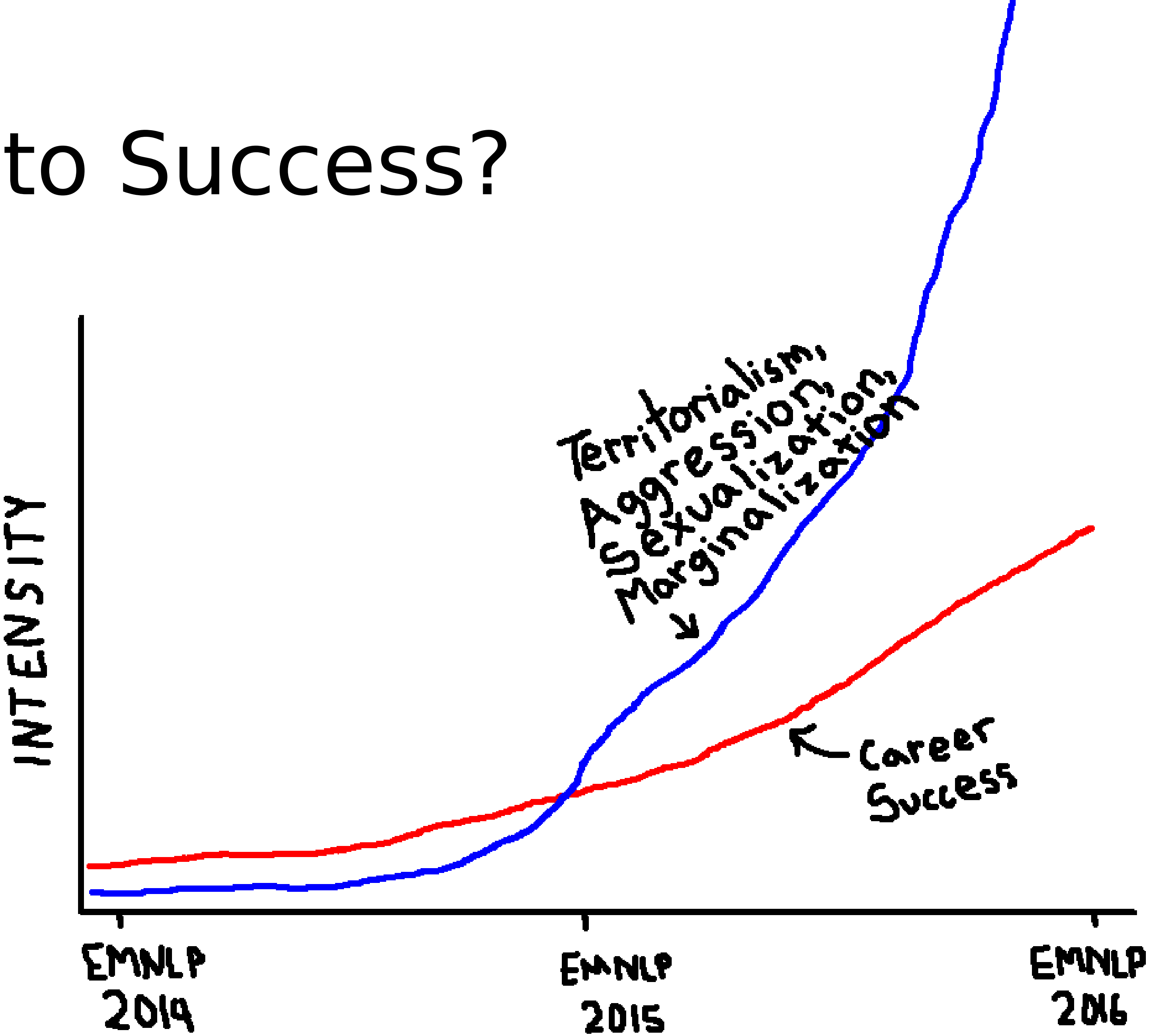


Meta (?) Uphill Battle

Path to Success?



Path to Success?



Thanks!

- E-mail: margarmitchell@gmail.com
- Webpage: m-mitchell.com