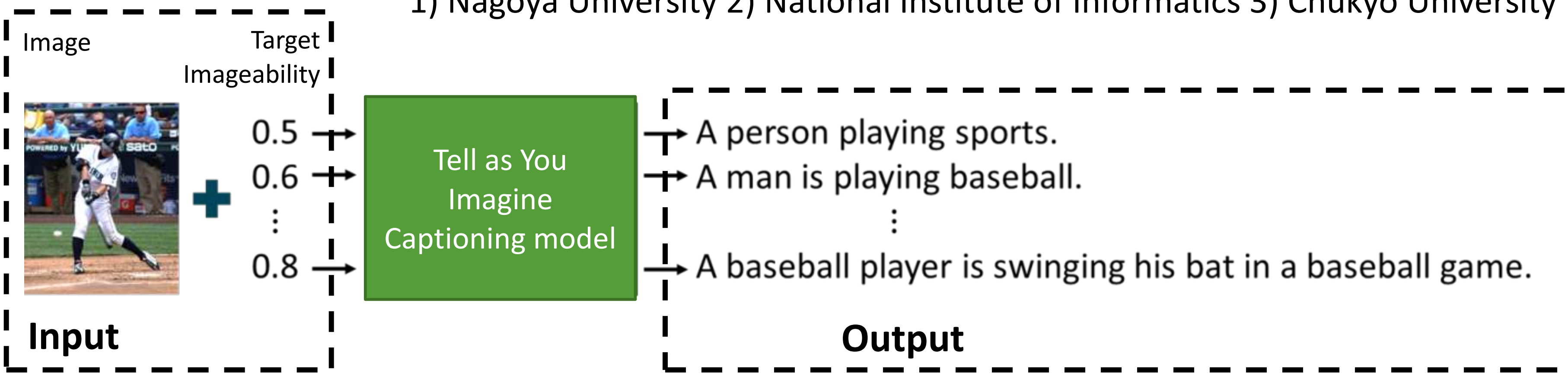
Tell as You Imagine: Sentence Imageability-Aware Image Captioning

Kazuki Umemura[1], **Marc A. Kastner**[2,1], Ichiro Ide[1], Yasutomo Kawanishi[1], Takatsugu Hirayama[1], Keisuke Doman[3,1], Daisuke Deguchi[1], Hiroshi Murase[1]

1) Nagoya University 2) National Institute of Informatics 3) Chukyo University



Background

- Existing image captioning approaches aim for an accurate image content description

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

A giraffe standing in a forest with trees in the background

- However, captions are used in varying applications with different needs and styles

For accessibility

- The advertisement billboard for the movie on the movie theater's building and two walking men.

For news paper article

- A sign for the popular Japanese manga "Demon Slayer" at a Tokyo theater last week.*

*https://www.nytimes.com/2020/10/20/business/demon-slayer-japan-movie.html

Research goal

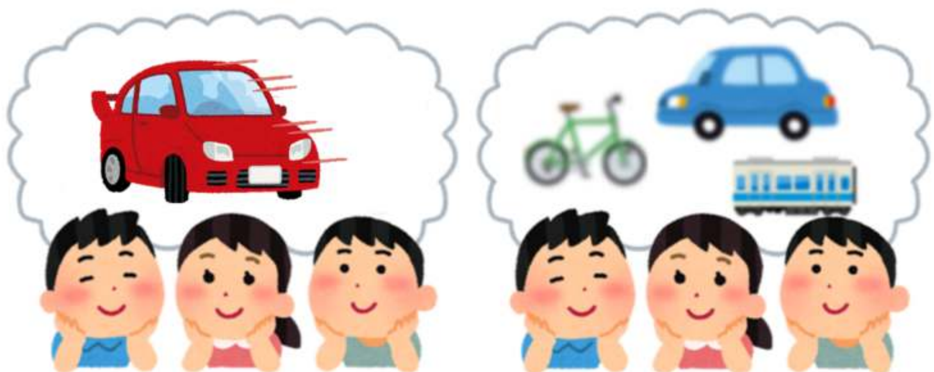
- We aim for diverse captioning with customizable descriptiveness of generated captions

Visual Descriptiveness	High	➔ A boy is riding a snowboard.
	Low	➔ A person is standing on the ground.

- By changing descriptiveness, the output can be adjusted to different applications

Using Imageability

- Imageability is "the ease with which a word gives rise to a sensory mental image"[1]
  - Psycholinguistic measure
  - Available as dictionaries[2] or estimation[3]

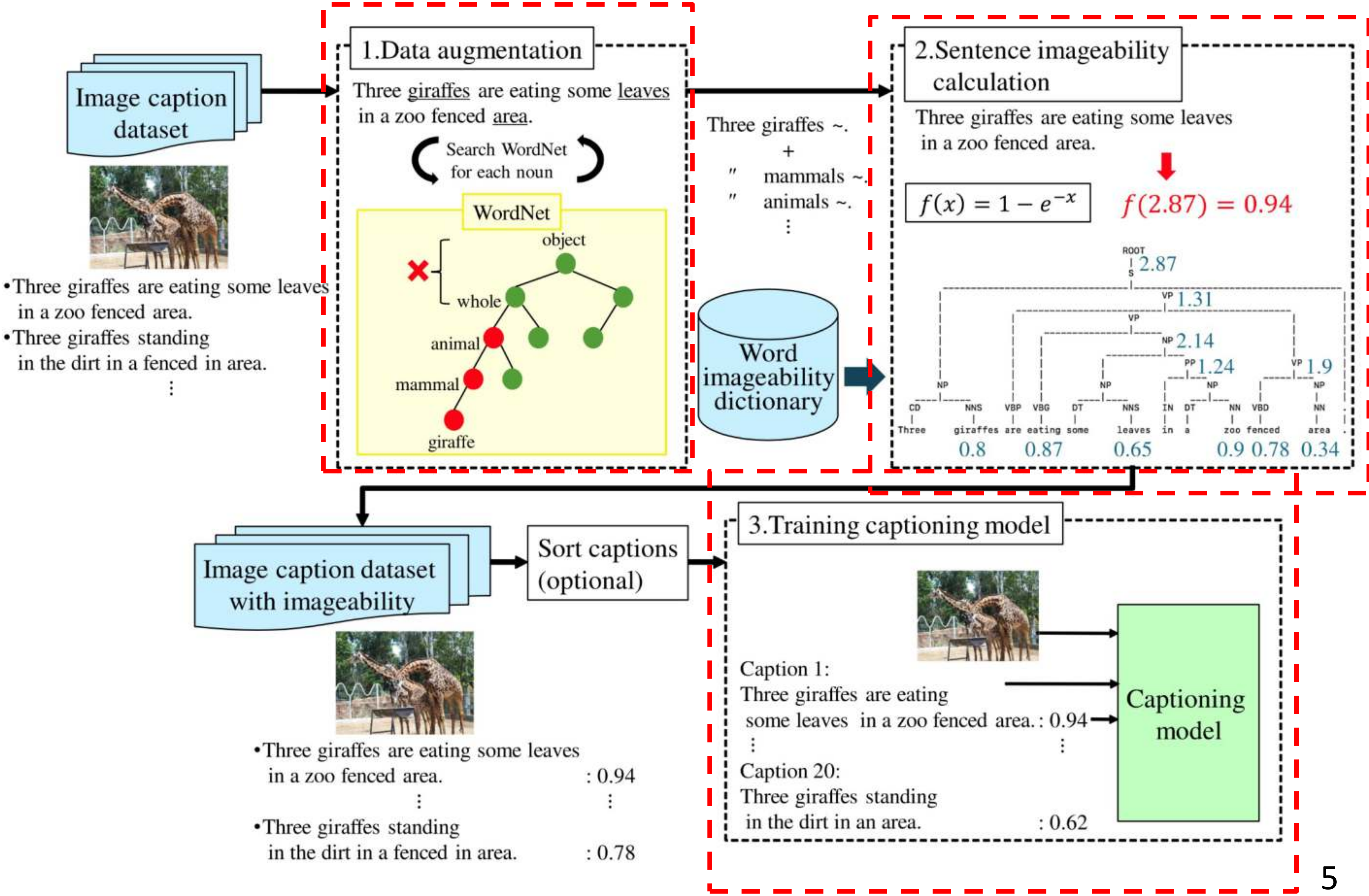


	"sports car"		"vehicle"
Imageability	0.6	>	0.3

→ Use imageability as an approximation for a captions' descriptiveness

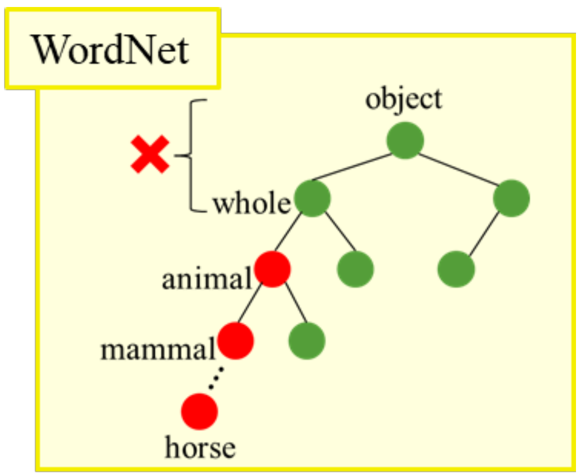
[1] Paivio et al., "Concreteness, imagery, and meaningfulness values for 925 nouns.," J. Exp. Psychol, 1968.
[2] Scott et al., "The Glasgow Norms: Ratings of 5,500 Words on Nine Scales.", Behav. Res. Meth, 2018.

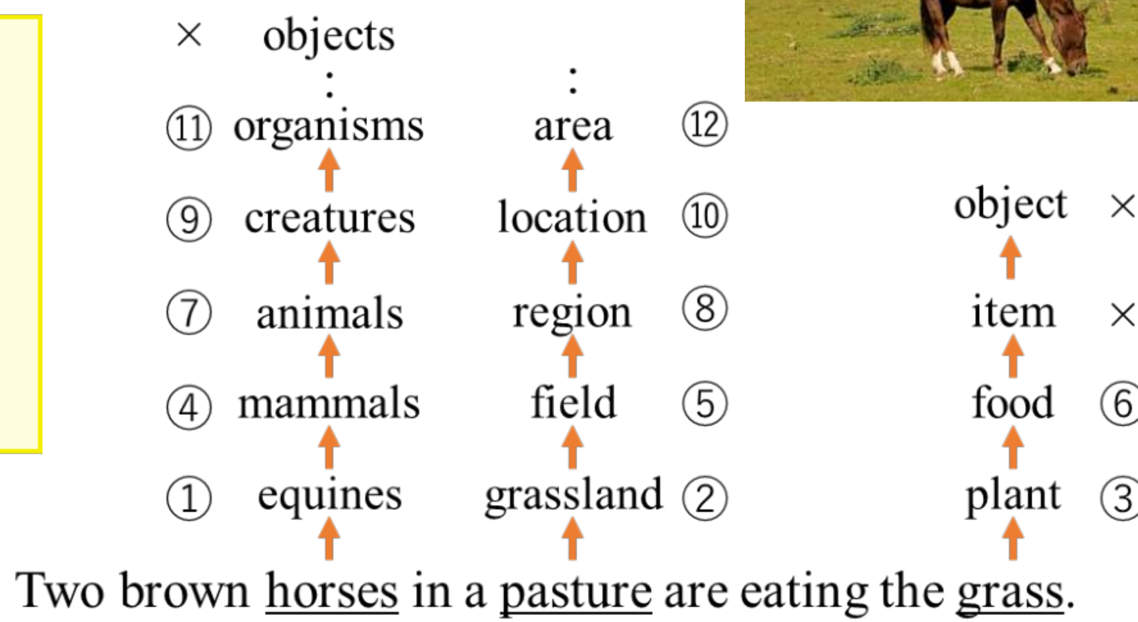
Proposed framework



1. Data augmentation

- Increase caption variety on an existing dataset[4]
  - For each noun, we add extra captions by replacing it with more abstract words
    - Using WordNet[5] for replacement

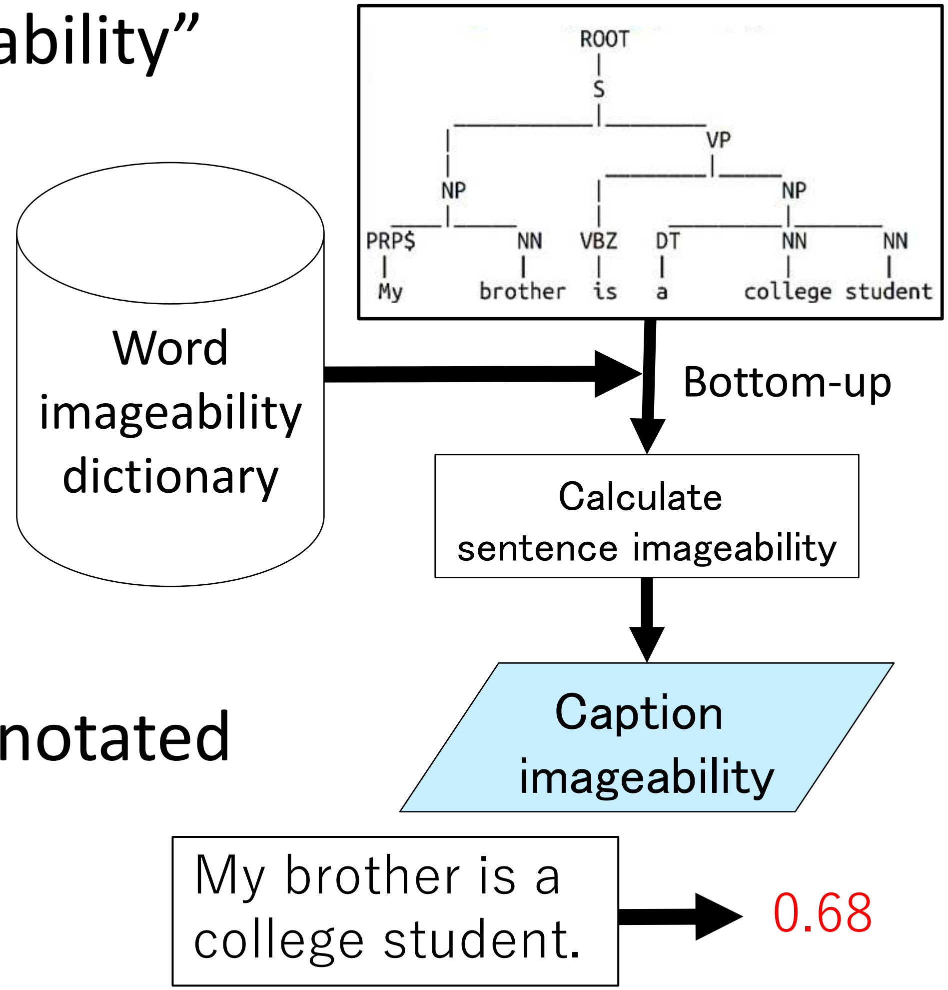




[4] Lin et al., "Microsoft COCO Common Objects in Context.", ECCV, 2014.
[5] Miller., "WordNet: A lexical database for English.", Commun. ACM, 1995.

2. Caption imageability calculation

- Calculate a “caption- imageability” score for each caption
  - Using word imageability in existing dictionaries[2,7]
  - In a bottom-up way using parsing tree
    - Rule-based approach to decide imageability for upper nodes (Details in paper)
- Resulting in imageability-annotated captions



[6] Manning et al. , “The Stanford CoreNLP natural language processing toolkit”, ACL, 2014.
[7] Ljubešić et al., “Predicting concreteness and imageability of words within and across languages via word embeddings.”, Workshop on RL for NLP, 2018.

3. Training the captioning model

- Extending LSTM-based architecture by Xu et al.[8]
- For a caption $c = \{\boldsymbol{w}_0, \boldsymbol{w}_1, \dots, \boldsymbol{w}_N\}$
  - w_i: i-th word vector
- Training 512-dim. vectors
  - $\boldsymbol{x}_t$: Language features
    - $\boldsymbol{x}_t = W_e \boldsymbol{w}_{t-1}$, where $t \in \{1, \dots, N\}$
  - $\boldsymbol{I}_t$: Attention-based visual features
    - $\boldsymbol{I}_t = \mathrm{Att}(\boldsymbol{h}_{t-1}, \boldsymbol{I}_f)$
  - $\mathbf{Imag}$: Imageability vector
    - $\mathbf{Imag} = [\mathit{Caption\ imageability}, \dots, \mathit{Caption\ imageability}]$

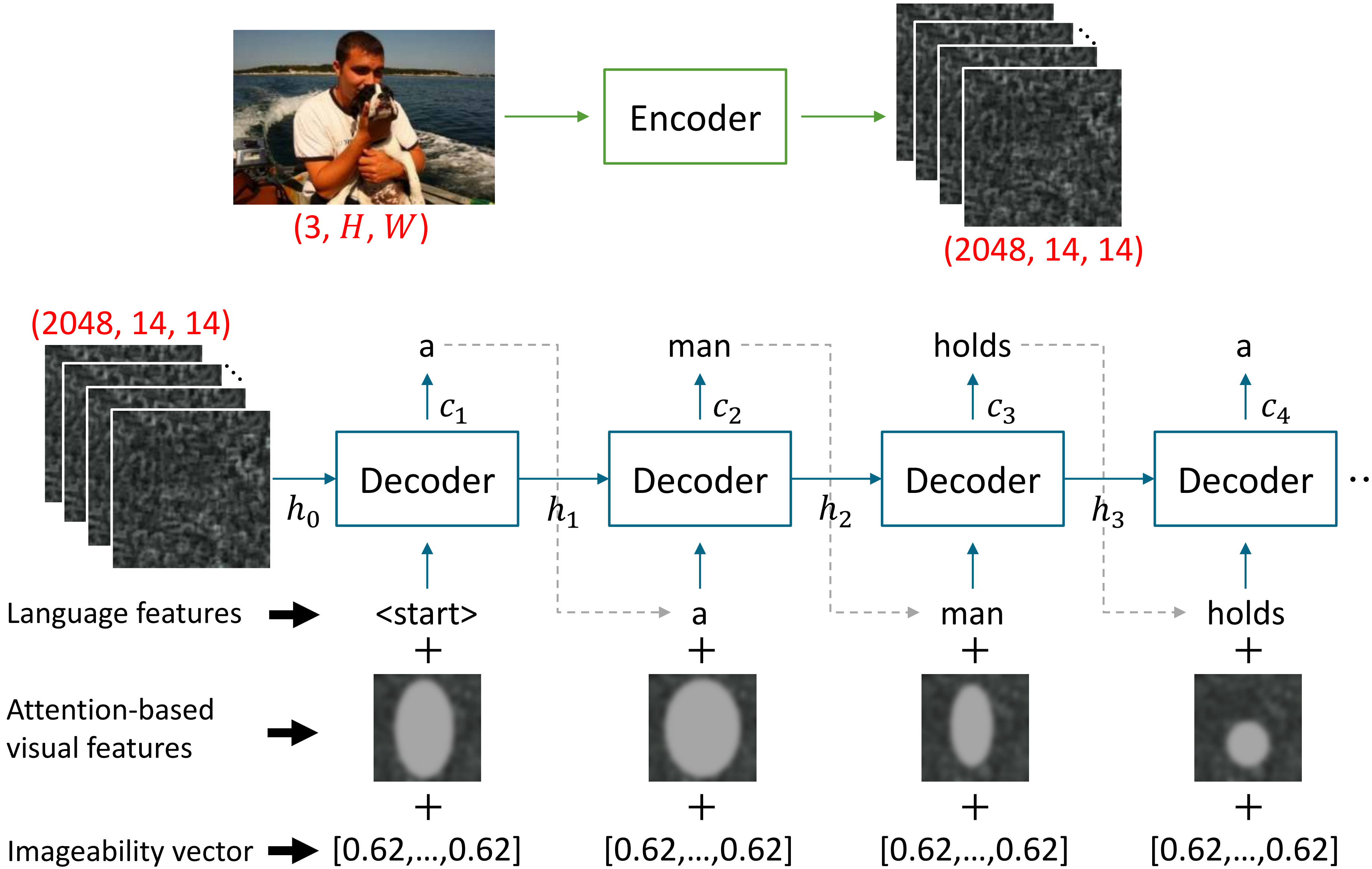
$\boldsymbol{I}_f$: Visual features from the attention network
$\boldsymbol{h}_{t-1}$: Hidden features from the previous step

$$\begin{cases} \boldsymbol{h}_t = \mathrm{LSTM}\big(\mathrm{concat}(\boldsymbol{x}_t, \boldsymbol{I}_t, \mathbf{Imag})\big) \\ \boldsymbol{w}_t = \mathrm{softmax}(W_l \boldsymbol{h}_t) \end{cases}$$

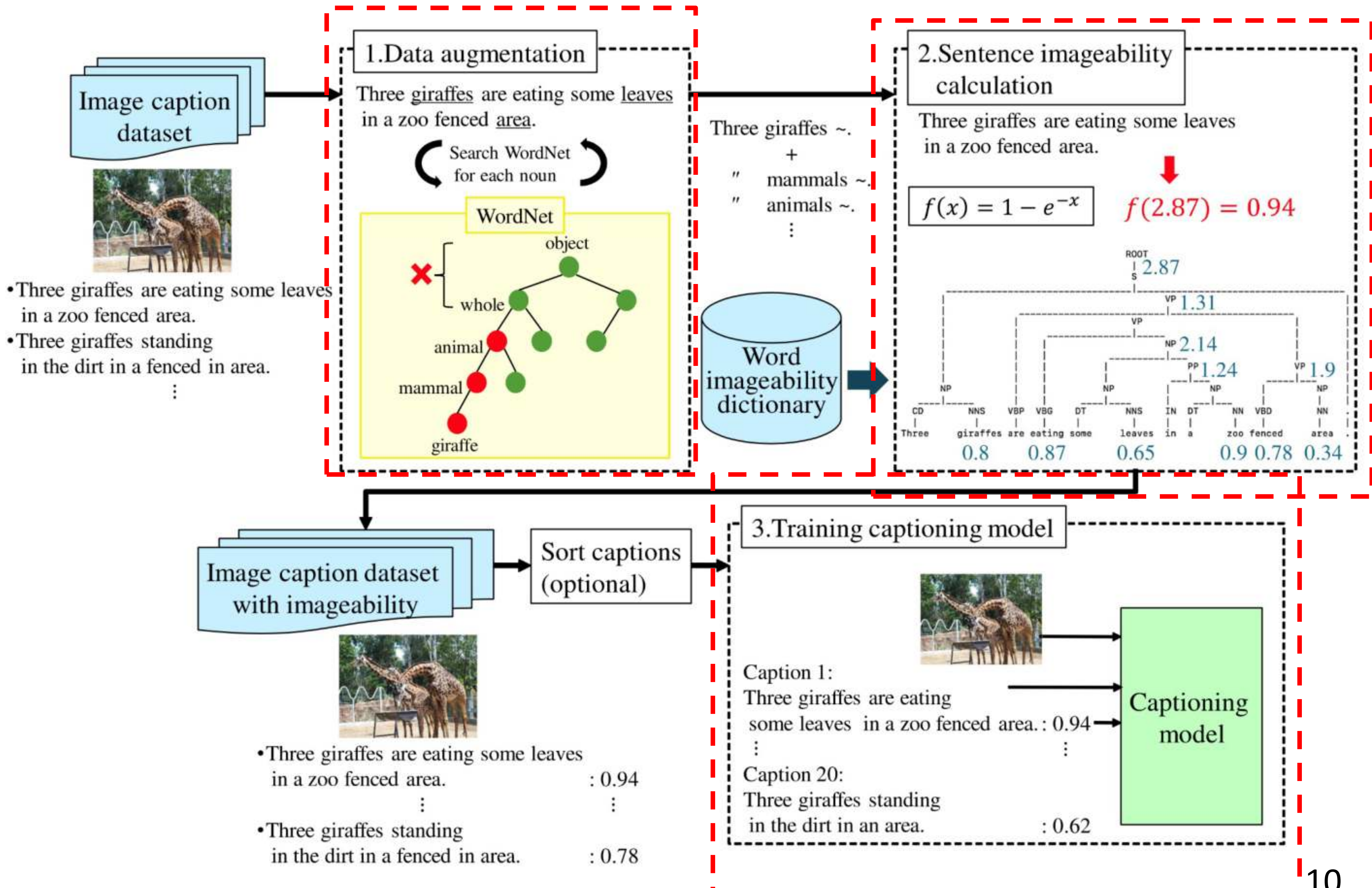
W_e, W_l: Training parameters

[8] Xu et al., “Show, attend and tell: neural image caption generation with visual attention”, ICML, 2015

Captioning model (extended from [8])



Proposed framework



Caption generation

- Input: Image＋Target imageability in [0,1]
- Output: Caption with customized visual descriptiveness

1. Generating output candidates based on beam-size
2. Calculating caption imageability for each output
3. Select the best candidate

+
0.8

CapA. A dog sitting in front of a red door.	→ 0.59
CapB. A brown and white dog sitting on a leash.	→ 0.72
CapC. A brown and white dog laying next to a bike.	→ 0.77
CapD. A brown and white dog standing next to a red container.	→ 0.81
CapE. A white dog standing on the ground.	→ 0.63

Environment (1/2)

- Training setting for the proposed method
  - Parameters
    - 9 levels of target imageability: 0.1, 0.2, ..., 0.9
    - Beam Size: 5
  - Sampling for training
    - w/o sorting: Order of augmentation
    - w/ sorting: Alternate between lowest/highest imageability

①	0.45 : An organism laying on…
③	0.46 : An animal sitting on ...
	:
④	0.82 : A dog sitting in ...
②	0.89 : A brown and white dog standing ...

- Comparison method
  - Train with imageability-annotated dataset
  - Select the first generated caption without selecting the best candidate

Environment (2/2)

- Baseline dataset: MS COCO[4]
- Ground-truth for word imageability
  - Combining Scott et al.[2] + Ljubešić et al.[7]
- Extending dataset as discussed before
  - Removing images which cannot be diversified
  - Ending up with (#imageability-annotated images):
    - Training: 109,114
    - Validation: 4,819
    - Test: 4,795
- Experiments
  1. Target Imageability
  2. Image captioning
  3. Crowd-sourced user study

Experiment 1: Imageability

- **Metrics**
  - Diversity of generated captions (avg. # generable captions)
  - Span of generated imageability (for targets between [0,1])
  - MSE between GT imageability and generated imageability
  - RMSE between GT imageability and generated imageability

- **Results**

Method	Sampling	Diversity	Imag. range	MSE			RMSE		
				Low [0.1, 0.3]	Mid [0.4, 0.6]	High [0.7, 0.9]	Low [0.1, 0.3]	Mid [0.4, 0.6]	High [0.7, 0.9]
Prop.	w/ sorting	**4.68**	0.083	0.405	0.118	**0.011**	0.632	0.334	**0.098**
	w/o sorting	4.63	**0.182**	**0.338**	**0.089**	**0.014**	**0.573**	**0.276**	0.107
Comp.	w/ sorting	3.50	0.070	0.434	0.131	0.015	0.655	0.354	0.117
	w/o sorting	3.26	0.164	0.378	0.103	0.022	0.607	0.300	0.142

Experiment 2: Image captioning

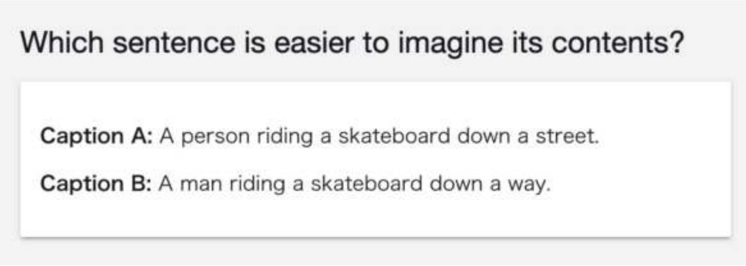
- Metrics
  - BLEU-4, CIDEr, ROUGE, METEOR, SPICE
  - Average across all imageability ranges
- Results

Method	Sampling method	BLEU-4	CIDEr	ROUGE	METEOR	SPICE
Proposed	w/ sorting	0.258	0.620	0.497	0.231	0.089
	w/o sorting	0.267	0.676	0.501	0.236	0.090
Comparison	w/ sorting	**0.267**	**0.636**	**0.501**	**0.233**	**0.090**
	w/o sorting	**0.277**	**0.706**	**0.506**	**0.240**	**0.091**

  - Comparison method slightly better, but does not consider imageability
    - ***To be expected***: BLEU-4 etc. intrinsically disadvantageous for style-changes as targeted in research goal.

Experiment 3: User study

- Using Amazon Mechanical Turk (AMT)
  - Evaluating 200 images with 278 English-speaking participants



- Experiment
  - Paired comparisons to decide descriptiveness of captions
    - Do they match the intended order (= low/mid/high descriptiveness)?
- Tested method
  - Proposed method w/ sorting
    Generating three captions per image: {0.5, 0.7,0.9}
- Results
  - "Correct" answers for pair-comparisons: **65.8%**
  - Spearman correlation between
    AMT order and intended order: **0.37**

Generated captions examples

Target score	Generated caption	
0.6	A placental is laying on a keyboard on a desk.	Prop. Method
0.7	A vertebrate is laying on a keyboard on a desk.	
0.8	A feline is laying on a keyboard on a desk.	
0.9	A cat is laying on a computer keyboard.	
0.6~0.9	A placental is laying on a keyboard on a desk.	Comp. Method

Target score	Generated caption	
0.6	A white and blue medium sitting on a runway.	Prop. Method
0.7	A white and blue medium on a runway.	
0.8	A small white and blue craft on a runway.	
0.9	A small craft sitting on top of an airport tarmac.	
0.6~0.8	A white and blue craft sitting on a runway.	Comp. Method
0.9	A small craft sitting on top of a runway.	

Conclusion

- Novel diverse image captioning framework
  - Allow for customizing visual descriptiveness to create captions for different purposes
  - Use word imageability to express and train descriptiveness
- Proposed framework
  - Augmenting existing dataset for diversity
  - Calculate caption imageability score for each caption
  - Train on {image, caption, imageability}
- Results promising, validated by crowd-sourced user study