On the quantification of the mental image of visual concepts for multi-modal applications

Marc A. Kastner[1], Ichiro Ide[2], Yasutomo Kawanishi[2],
Takatsugu Hirayama[2], Daisuke Deguchi[2], Hiroshi Murase[2]

(1) National Institute of Informatics (2) Nagoya University

Presentation #5 - CVIM-222 / May 14, 2020

About me

- Marc A. Kastner (30), German

- Research interest
  - Language and vision
  - Computational psycholinguistics
  - Emotion and sentiment

- Education and Work
  - TU Braunschweig, Germany
    - 2010 – 2016: B.Sc. and M.Sc.
  - Nagoya University, Japan
    - 2013 – 2014: Student exchange
    - 2016 – 2020: Ph.D.
  - National Institute of Informatics, Japan
    - 2019: Internship
    - 2020 – now: Post-doc
      - With Prof. Shin'ichi Satoh

Mental image of concepts

- Peaceful
  - Different backgrounds
  - Different contents
- Leaf
  - Always same “object”
  - Always in forest

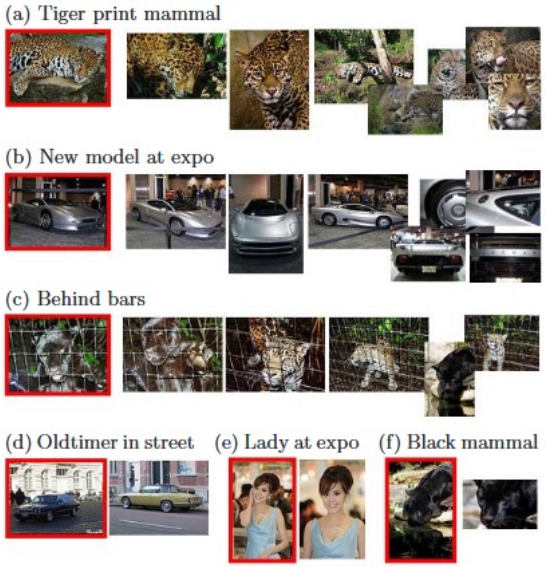
Something Peaceful Morning Breakfast Leaf . . . Hammer

Abstract Concrete

Research motivation

- Mental imagery
  - "Visual experience of a concept […] from memory"
  - In this context, the ***common*** perception across society
- Imagine two concepts like "peaceful" and "leaf"
  - Are they equally hard to visually imagine?
  - Which is more abstract or more concrete?
- Goals
  - Modeling the quantization of a mental image of a concept

Example applications



- Visual diversification [1]
  - Increase variety of image retrieval results

- Multi-modal approaches using text + image [2]
  - Analyzing relationship of slogan and image for advertisements

- Language processing [3]
  - Can help when learning translations in machine translation models

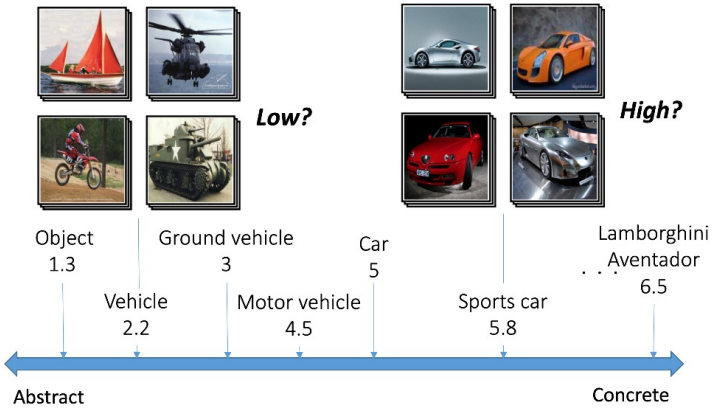
1: Reinier H. van Leuken et al. Visual diversification of image search results. WWW 2009.
2: Zhang et al. Equal But Not The Same: Understanding the implicit relationship between persuasive images and text. BMVC 2018.
3: Hewitt et al. Learning translations via images with a massively multilingual image dataset. ACL 2018.

Core idea

- Assumption
  - Images on the Web somewhat reflect the ***common*** mental image members of our society have of a concept
  - E.g. if crawling many images of ***leaf***, we get a reasonable average mental image of ***leaf***

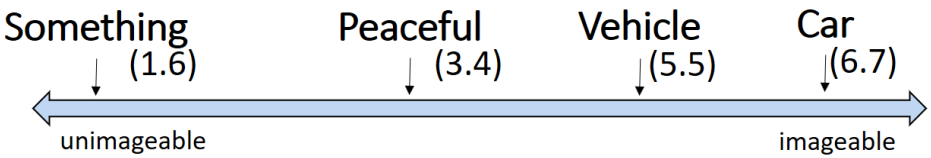
- Possible approaches
  - **Relative** measurement: Estimate relative perception differences of similar words → Research 1
  - **Absolute** measurement: Estimate absolute values for arbitrary concepts → Research 2

Research Target



Research 1

- Estimate **visual variety of concepts** within same domain
  - **Relative** measurement
  - For direct comparison, relative to another
  - E.g. Sports car <-> Car <-> Airplane <-> Vehicle
- **Data-driven** approach creating ideal datasets for the measurement
- Usable for word selection problems
  - But, hard to quantify Vehicle <-> Pizza because of missing reference point



Research Target

Research 2

- Estimate **imageability of words** on a dictionary level
  - **Absolute** measurement
  - For global trend of abstract vs. concrete
  - E.g. Random <-> Peace <-> Airplane <-> Pizza
- **Algorithm-driven** approach using supervised learning
- Usable for text difficulty, abstractness of text, etc.
  - But, scale not granular enough to be used for Sports Car <-> Car

My first Japanese talk at CVIM two years ago!

カストナーマークアウレル, 井手一郎, 川西康友, 平山高嗣, 出口大輔, 村瀬洋, Web画像の分布に基づく単語概念の視覚的な多様性の推定. CVIM 2018.03

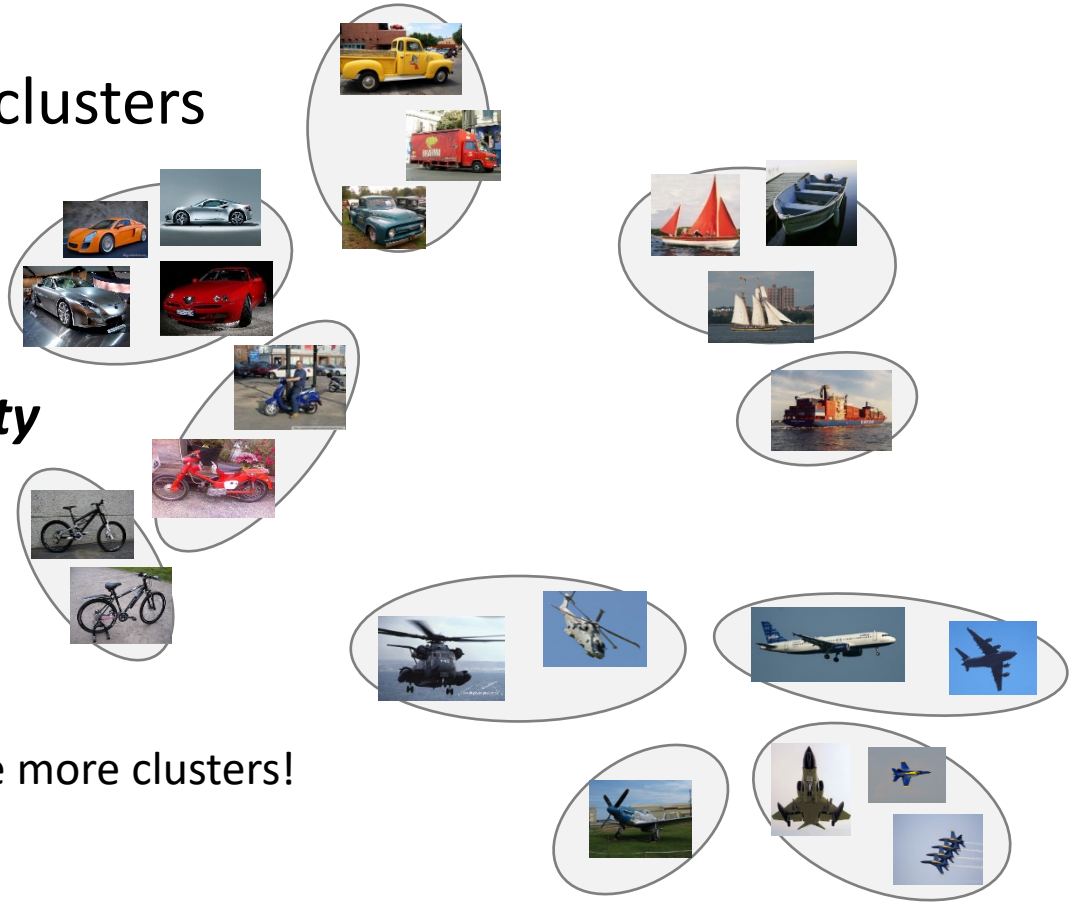
Research 1

Estimation of visual variety of concepts

Target: Relative measurements in same domain

Core approach

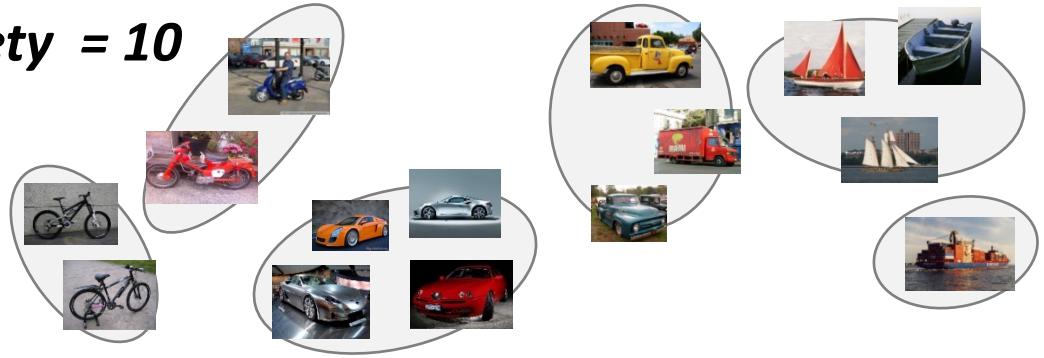
1. Extract visual features of images (Bag-of-Words)
2. Cluster feature space
3. Variety = Number of clusters



Visual variety = 10

Visually variant concepts have more clusters!

Visual variety = 10



Biased data

- Naïve:
  - Extract visual features
  - Cluster the visual feature space

  →The number of clusters express variety!

- Looking closer at datasets like ImageNet[6]
  - Very biased composition
  - *Vehicle* largely consists of *military vehicles*
  - Does not reflect reality!

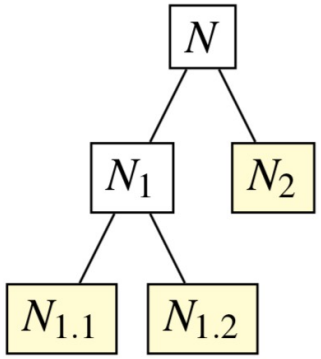
- We need to create a *"proper"* dataset!

6: J.D.J. Deng et al. ImageNet: A large-scale hierarchical image database. CVPR 2009.

Ideal dataset

- Of course, a mental image is very subjective…
- But in general, the dataset composition should probably reflect the distribution in real life:
  - Rare sub-concepts → Few images
    (For *vehicles,* few *tanks*)
  - Common sub-concepts → Many images
    (For *vehicles*, many *cars*)
- Recreate datasets based on this idea!
  - Compose abstract concepts' datasets using images of their sub-concepts reflecting *"natural"* distribution

Dataset creation

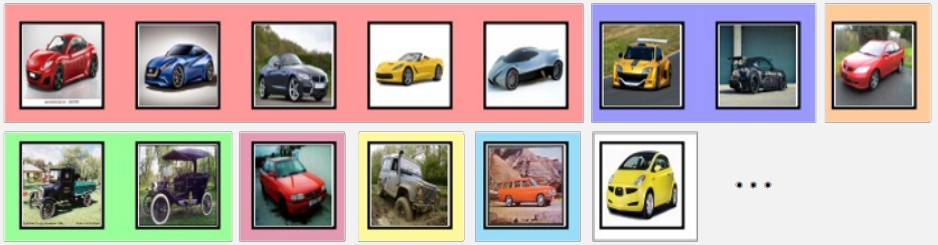


- Recompose dataset N' from sub-concept images p using Web popularity weighting w

$$N' = w_1 p(N_{1.1}) + w_2 p(N_{1.2}) + w_3 p(N_2)$$

Comparative

$w_1 = w_2 = \cdots$



Ideal dataset

$w_i = ?$

So, what is w_i?

Popularity weighting

Idea:

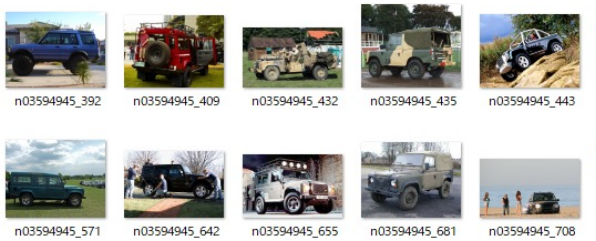
- Retrievable with Google API
  - Number of search results for a term
  - Separate API and results for *Google Text* and *Google Image*

- Use this to decide ratio for dataset composition

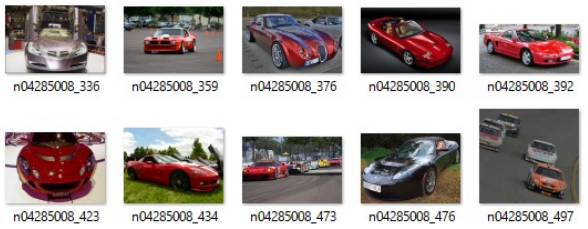
Dataset creation (2)

	Google Image Results
sports car	27.4%
racer	9.2%
Model T	8.8%
coupe	6.9%
used-car	6.7%
jeep	5.0%
beach w.	4.8%
compact	4.5%
cab	3.9%
convertible	3.5%
hatchback	2.7%
minivan	1.3%
ambulance	1.4%

Pictures of: *Jeep*



Pictures of: *Sports car*



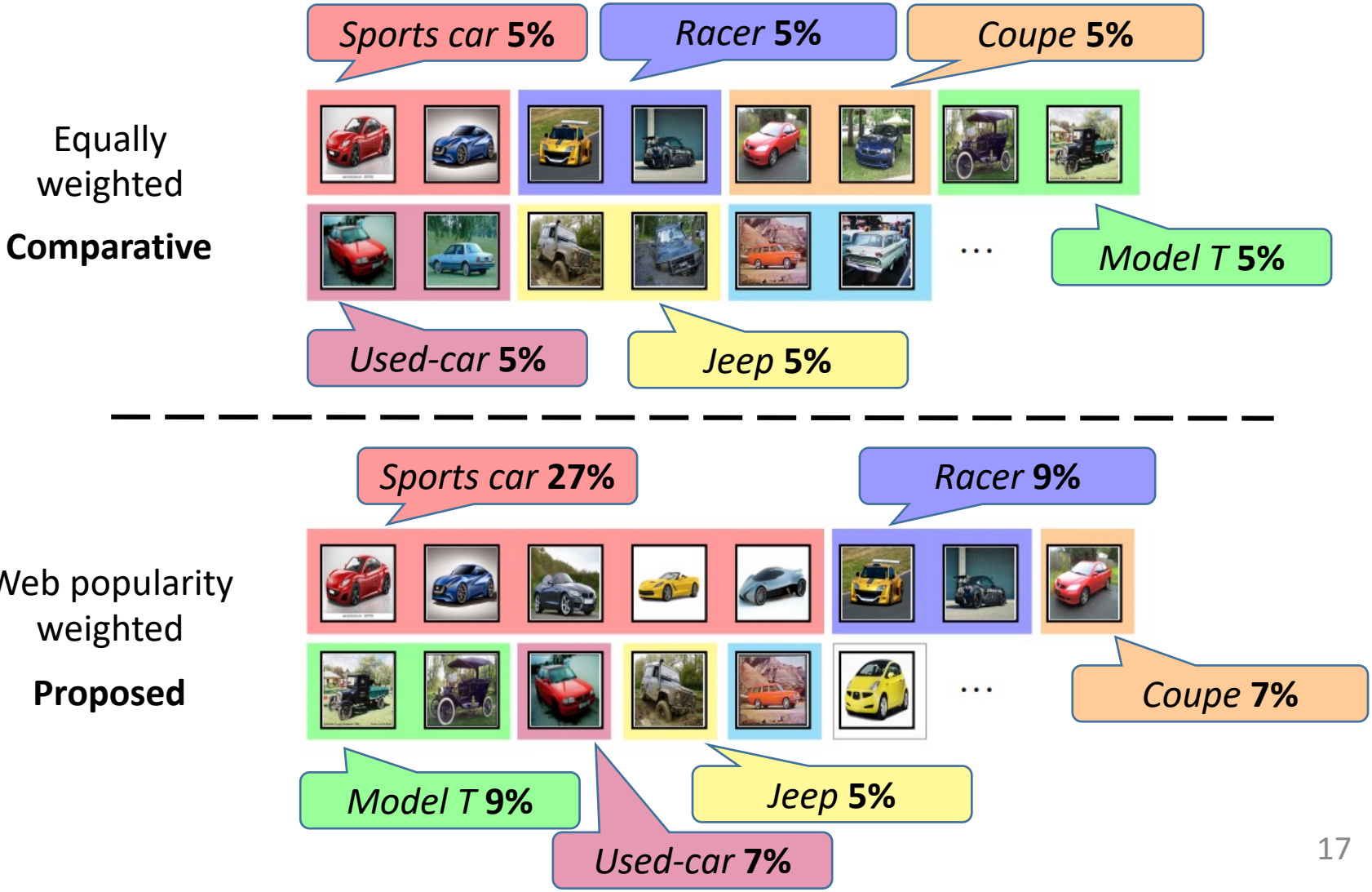
Re-composited dataset for *car*

Experiment

- Implementation
  - Visual features: Bag-of-Visual-Words (SURF)
  - Clustering: Mean-Shift
    - Results → Normalized number of clusters
- Ground-truth
  - Through crowdsourced survey: 4,529 pair comparisons [7] from 158 people via social media
- Evaluation metrics
  - Spearman's Rank Correlation
  - Mean Squared Error (MSE)

7: L. L. Thurstone. The method of paired comparisons for social values. J Abnorm Psychol 1968.

Dataset for *car* (Example)



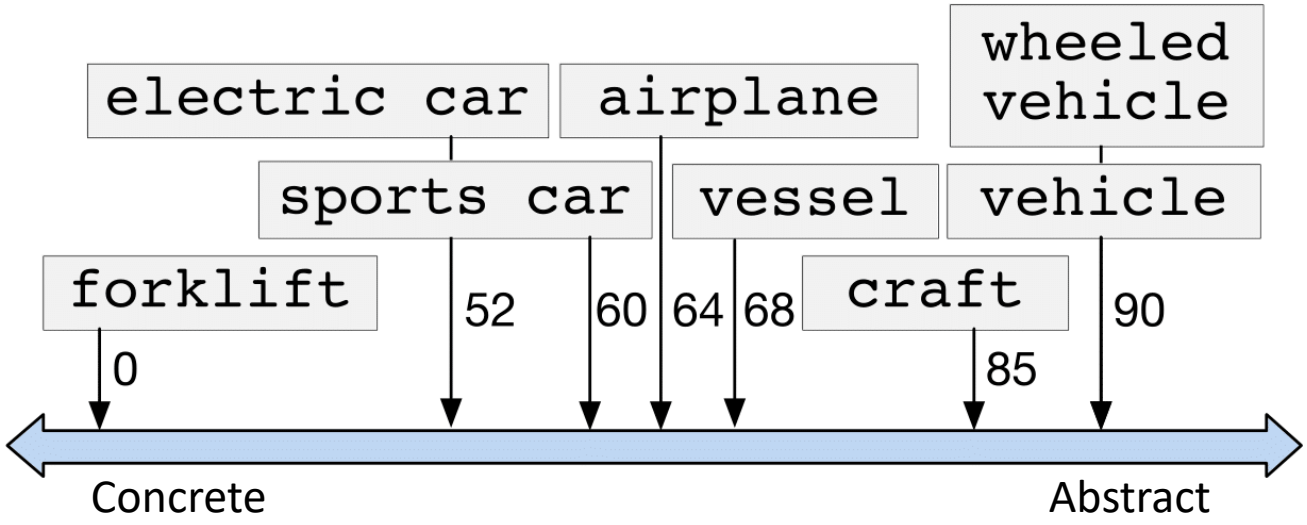
Results

Dataset	Rank correlation	Mean Squared Error (MSE)
Baseline (**Original** ImageNet)	0.25	10.54
Comparative (Re-compose with **equal** weighting)	0.62	9.23
Proposed 1 (Re-composed with Google **Text** weighting)	0.56	14.89
Proposed 2 (Re-composed with Google **Image** weighting)	**0.73**	**9.01**

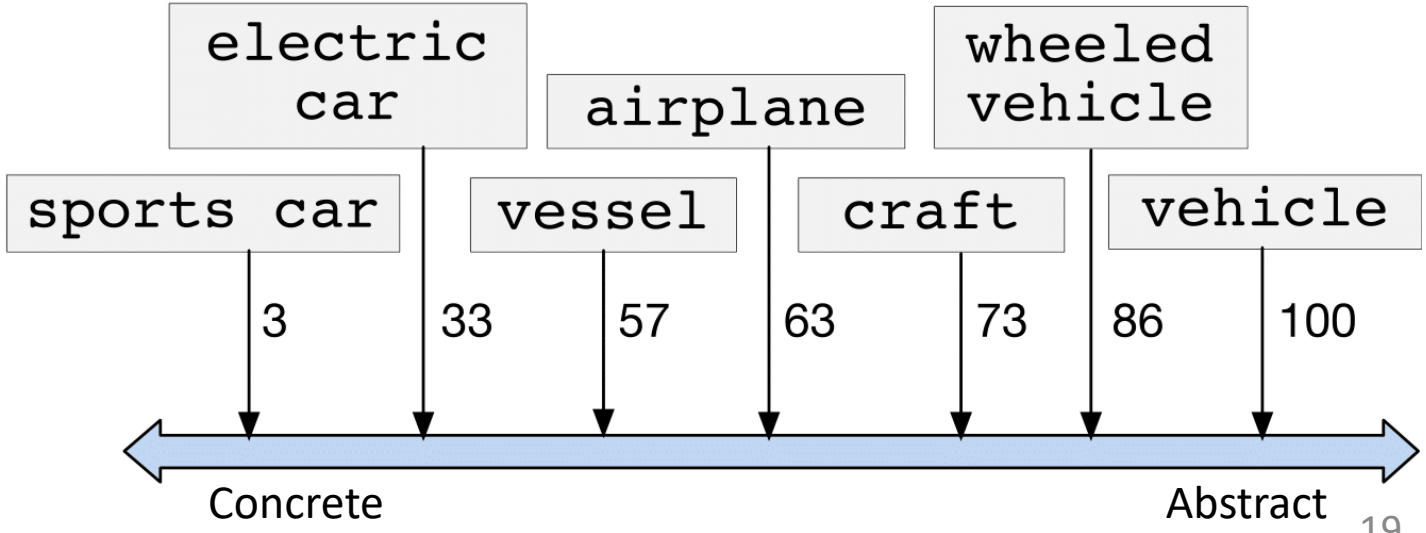
- Baseline does not correlate at all
- Proposed method 2 improves correlation by 192% over the baseline and 17.7% over the comparative method.

Examples

Ground truth



Proposed method 2



Summary

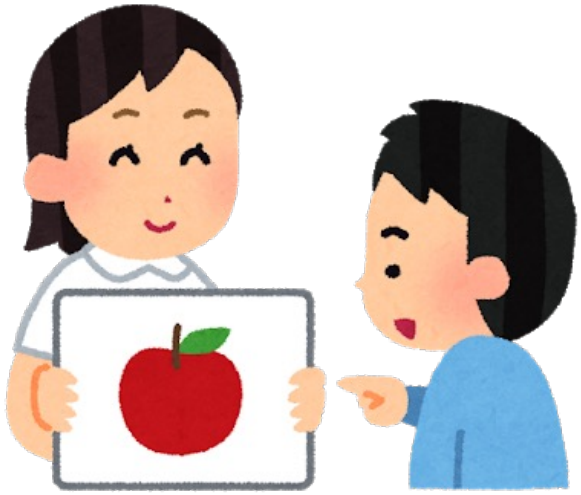
- Analyzing local visual variety differences of words within the same domain
  - Using a dataset-driven approach to estimate human perception of visual concepts
  - Established ground-truth results with a crowd-sourced survey (*n* = 158)
- Proposed method improves correlation by 192% over the baseline and 17.7% over the comparative method

Research 2

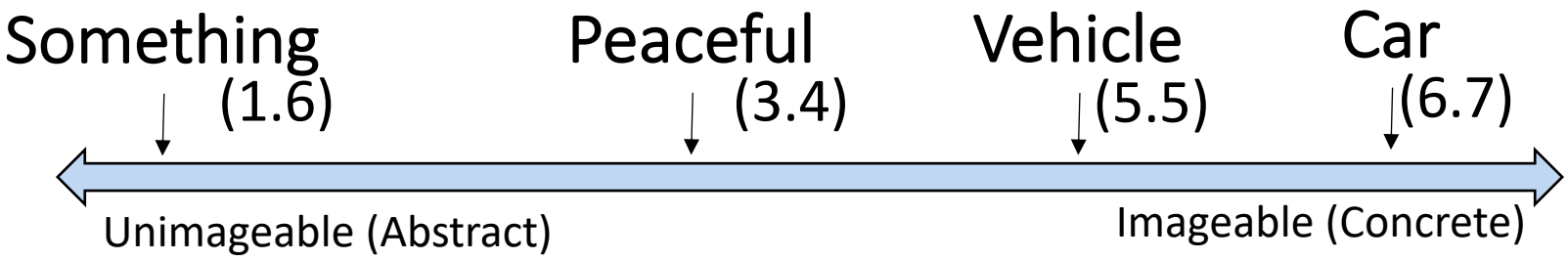
Estimation of word imageability

Target: Absolute measurements on dictionary-level

心像性
Imageability of words



- Concept from Psycholinguistics [1]
  - Quantize the perception of words
  - Often described on Likert scales
    - *Unimageable <-> Imageable* or *Abstract <-> Concrete*
  - Is a concept imageable? Do you have a mental image when thinking of a concept?



1: Pavio et al. Concreteness, imagery, and meaningfulness values for 925 nouns. J Exp Psych 1968.

Motivation

- While there are existing imageability dictionaries
  - Datasets are small (< 6000 words)
  - Most dictionaries are created by hand
    - Extension is very labor intensive
    - Data often republished or reshuffled, but rarely increased
- Idea: **Estimate imageability scores** to extend existing dictionaries by analyzing visual data
  - Use core assumption of research 1
  - Train a model for imageability estimation

Purpose

- Estimate an imageability score for a word based on its visual characteristics
  - Analyze images crawled for each word
  - Train regression model to estimate the score based on visual features

Analysis

$I_{\text{leaf}} \in [1,7]$

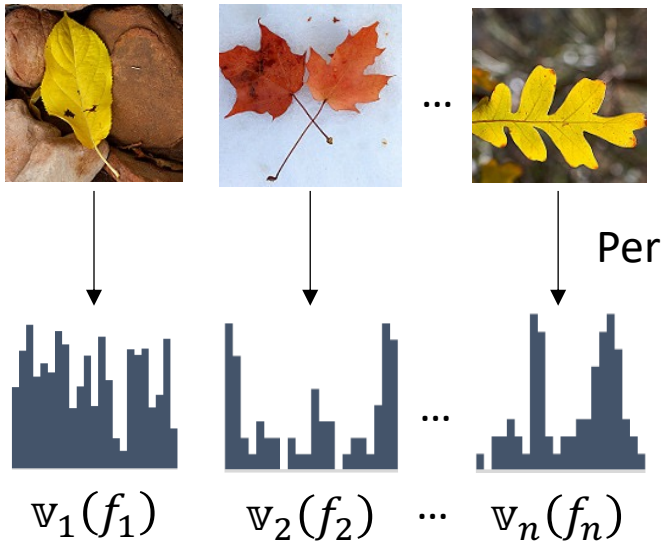
Input:
Images for “leaf”

Output:
Imageability score for “leaf”

Approach

Extracting visual features

- For each word, crawl image data from social media
- Then, extract visual features from each image
  - E.g. Color histograms, Bag-of-Visual-Words histograms, ...



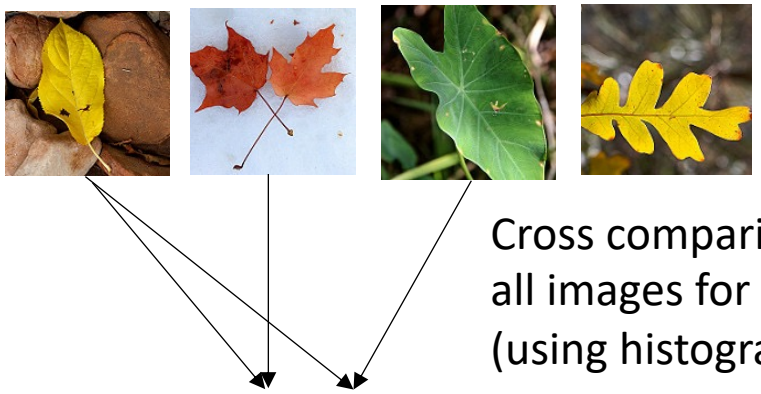
Crawled images for “leaf”

Feature vector for each image $\mathbb{v}_i$

Approach

Cross comparison of images

- Cross-compare all images of same word
- Create similarity matrix containing similarity between all image pairs



Cross comparison between all images for “leaf” (using histogram similarity)

Using visual feature histograms

$$s_i = d\left(\mathbb{v}_i(f_i), \mathbb{v}_j(f_i), \ldots, \mathbb{v}_n(f_i)\right) = \begin{bmatrix} 1.0 & 0.3 & \cdots \\ \vdots & \ddots & \vdots \\ 0.7 & \cdots & 1.0 \end{bmatrix}$$

Similarity matrix per visual feature f_i

Approach

Regression model

- Random forest based on visual characteristics
  - Train model on eigenvalues of similarity matrix
  - Use imageability dictionary as ground-truth

$$s_i = \begin{bmatrix} 1.0 & 0.5 & \cdots \\ \vdots & \ddots & \vdots \\ 0.1 & \cdots & 1.0 \end{bmatrix} \qquad s_j = \begin{bmatrix} 1.0 & 0.3 & \cdots \\ \vdots & \ddots & \vdots \\ 0.7 & \cdots & 1.0 \end{bmatrix}$$

Similarity matrix for each visual feature

Train on eigenvalues

X

Imageability dictionary

Y

Random Forest

Y'

$I_{\text{leaf}} \in [1,7]$

Output:
Imageability for "leaf"

Input:
Images for “leaf”

For each visual feature f_i

$\mathbb{v}_i$

Feature vector for f_i

Cross comparison between
all images for “leaf”

$$s_i = \begin{bmatrix} 1.0 & 0.3 & \cdots \\ \vdots & \ddots & \vdots \\ 0.7 & \cdots & 1.0 \end{bmatrix}$$

Similarity matrix

Train on eigenvalues

Imageability dictionary → Random Forest

Regressor

$I_{\text{leaf}} \in [1,7]$

Output:
Imageability for “leaf”

Visual features

- **Low-level** (Traditional computer vision)

1. Color histogram (HSV)
2. Bag-of-Visual-Words (SURF)
3. GIST descriptors

- **High-level** (YOLO-based)

1. Image theme (e.g. Indoor, Architecture, Park, …)
2. Image content (e.g. 2 people, 1 dog, 2 signs, ..)
3. Image composition (e.g. 3 objects in edges, 1 in center)

Experiment

- Objective: Predict ***imageability*** for a set of words
- Using dataset of 587 words and 5,000 images each
  - Ground truth -> Imageability dictionary [8, 9]
  - Images -> YFC100M [10]

- Evaluation metrics
  - Pearson Correlation
  - Mean Absolute Error (MAE)
- Methods:
  - Proposed: Visual data mining
  - Comparative: Text data mining [11]

8: Cortese et al. Imageability ratings for 3,000 monosyllabic words. Behav Res Method 2004.
9: Reilly et al. Formal distinctiveness of high- and low-imageability nouns: analyses and theoretical implications. Cogn Sci 2007.
10: Thomee et al. YFCC100M: The new data in multimedia research. CACM 2016.
11: N. Ljubesic et al. Predicting concreteness and imageability of words within and across languages via word embeddings. 3rd Workshop on Representation Learning for NLP 2018.

Evaluation: Results

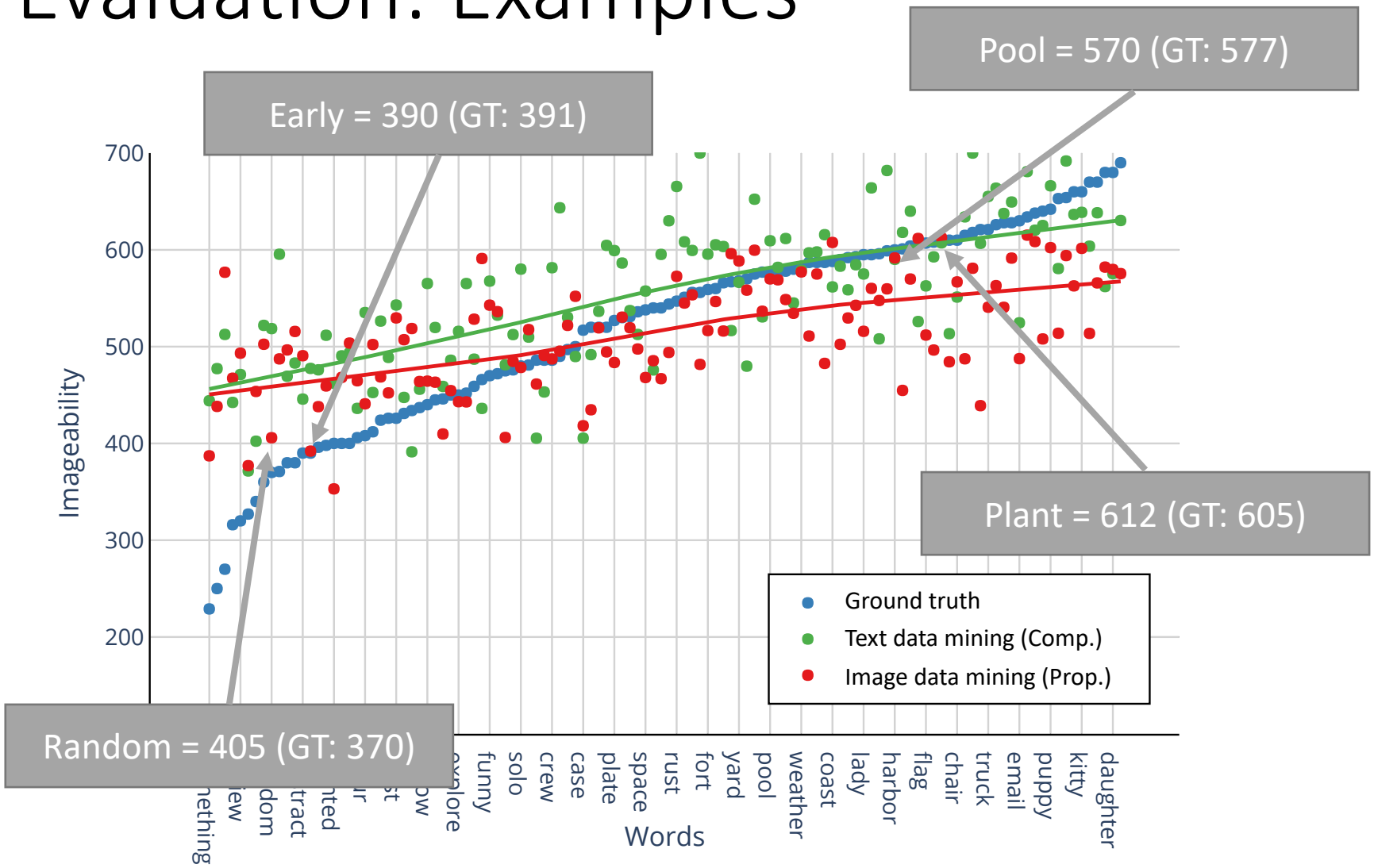
- Estimating imageability for test data

Feature	Correlation	MAE
Combined (Low)	0.60	11.03
Combined (High)	0.61	10.18
Combined (All)	0.63	**10.14**
Comparative (Text data mining [8])	**0.70**	10.39

- Splitting test-data in abstract vs. concrete words

Feature	Abstract only		Concrete only	
	Corr.	MAE	Corr.	MAE
Combined (Low)	**0.32**	**10.90**	0.16	11.37
Combined (High)	0.27	11.31	**0.10**	**9.10**
Combined (All)	0.26	10.79	0.17	10.11

Evaluation: Examples



Summary

- Proposed a method to estimate the imageability of words on dictionary level
  - By analyzing the visual characteristics of Web-crawled images from social media
- Estimated imageability with an error of 10.14% and a correlation of 0.63
  - Results are similar to solely text-based approaches, but fusing both might further improve results towards better correlation

Research summary

- During my doctoral studies, I quantized the perception and mental image of visual concepts

Research 1

- Local visual variety on the same domain
  - Verified a high correlation on 21 terms related to vehicles
  - Sports car <-> Car <-> Ground vehicle <-> Vehicle

Research 2

- Global imageability on dictionary-level
  - Low error and good correlation for a dataset with 587 words
  - Interesting results when playing with low-level vs. high-level features for abstract vs. concrete words.

Summary of my Doctoral studies

- Proposed idea of using visual variety for mental image modeling
- Main assumption: Connection between mental image and Web-crawled image sets
  - Verified by results of both Research 1 and 2
  - Image sets can be used to both replicate and augment results from Psycholinguistics
- Different visual characteristics contribute to mental image of different concepts

Publications during Ph.D.

Journal

1. **M.A. Kastner**, I. Ide, F. Nack, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, Estimating the imageability of words by mining visual characteristics from crawled image data. MTAP 2020.
2. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, Estimating the visual variety of concepts by referring to Web popularity. MTAP 2019.

International conferences

1. C. Matsuhira, **M.A. Kastner,** I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, K. Doman, H.Murase, Imageability estimation using visual and language features. ICMR 2020. (Accepted for publication in October 2020.)
2. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, Browsing visual sentiment datasets using Psycholinguistic Groundings. MMM 2020.
3. **M.A. Kastner**, On quantizing the mental image of concepts for visual semantic analyses. ACMMM 2019 Doctoral Symposium.

Domestic conferences

1. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, On the Quantification of the Mental Image of Visual Concepts for Multi-modal Applications. CVIM 2020.05 Presentation.
2. 松平茅隼, **カストナーマークアウレル**, 井手一郎, 川西康友, 平山高嗣, 道満恵介, 出口大輔, 村瀬洋, 視覚特徴と言語特徴を用いた単語の心像性推定の検討. ANLP 2020 Poster.
3. 梅村和紀, **カストナーマークアウレル**, 井手一郎, 川西康友, 平山高嗣, 道満恵介, 出口大輔, 村瀬洋, 心像性に基づく画像キャプショニングの検討. MVE 2020 (**MVE賞受賞**).
4. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, On visualizing psycholinguistic groundings for sentiment image datasets. MIRU 2019 Demonstration.
5. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, A preliminary study on estimating word imageability labels using Web image data mining. ANLP 2019 Presentation.
6. 梅村和紀, **カストナーマークアウレル**, 井手一郎, 川西康友, 平山高嗣, 道満恵介, 出口大輔, 村瀬洋, 画像キャプションの質的評価に向けた文の心像性推定手法の検討. ANLP 2019 Presentation.
7. **M.A. Kastner**, I. Ide, Y. Kawanishi, T. Hirayama, D. Deguchi, H. Murase, On understanding visual relationships of concepts by visualizing bag-of-visual-words models. MIRU 2018 Poster.
8. **M.A. Kastner**, Y. Tsuchiya, K. Osumi, R. Akiyama, L. Kawai, H. Ikeda, A Survey on Psychology - Connecting Perception, Language, and Memory studies with Computer Vision. MIRU 2018 Wakate Poster + Presentation.
9. **カストナーマークアウレル**, 井手一郎, 川西康友, 平山高嗣, 出口大輔, 村瀬洋, Web画像の分布に基づく単語概念の視覚的な多様性の推定. CVIM 2018.03 Poster + Presentation.

Thank you for your attention!

Questions?