SEED: Towards More Accurate Semantic Evaluation for Visual Brain Decoding


Juhyeon Park


Peter Yongho Kim

Equal Contribution


Jiook Cha


Shinjae Yoo


Taesup Moon

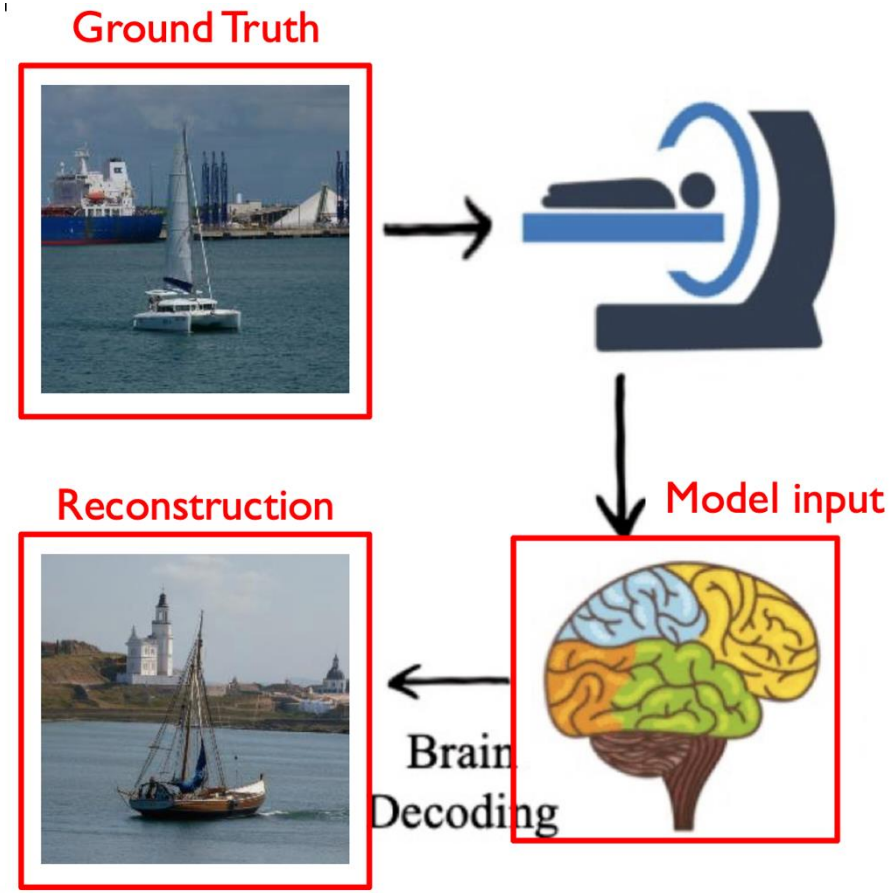
Machine Intelligence & Data science LAB

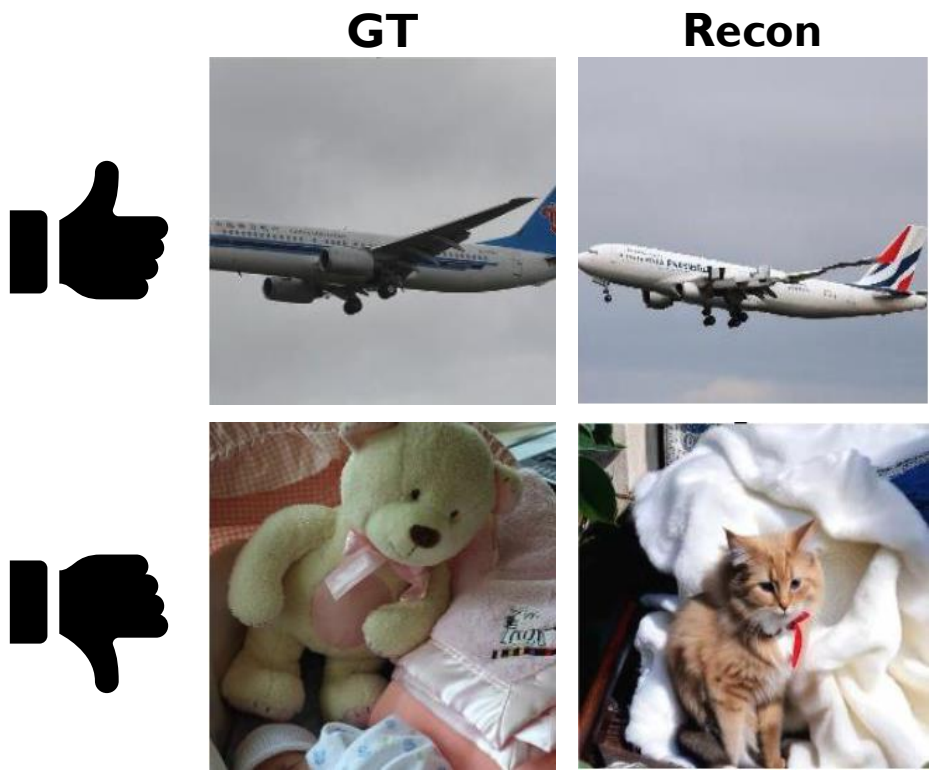
Contents

- Research motivation
- New semantic evaluation methods
- Experimental results

Research motivation

- fMRI-based visual brain decoding





How should we **evaluate** these images?

Research motivation

- For the evaluation of visual decoding models, the following 8 metrics are widely used

Pixel-level **2-way identification** **Feature Correlation**

Method	Low-Level				High-Level			
	PixCorr ↑	SSIM ↑	Alex(2) ↑	Alex(5) ↑	Incep ↑	CLIP ↑	Eff ↓	SwAV ↓
MindEye2	**0.322**	**0.431**	**96.1%**	98.6%	95.4%	93.0%	**0.619**	0.344
MindEye2 (unrefined)	0.278	0.328	95.2%	**99.0%**	**96.4%**	**94.5%**	0.622	**0.343**
MindEye1	0.319	0.360	92.8%	96.9%	94.6%	93.3%	0.648	0.377
Ozcelik and VanRullen (2023)	0.273	0.365	94.4%	96.6%	91.3%	90.9%	0.728	0.421
Takagi and Nishimoto (2023)	0.246	0.410	78.9%	85.6%	83.8%	82.1%	0.811	0.504
MindEye2 (low-level)	0.399	0.539	70.5%	65.1%	52.9%	57.2%	0.984	0.673
MindEye2 (1 hour)	0.195	0.419	84.2%	90.6%	81.2%	79.2%	0.810	0.468

- RQ: *"Is the current framework to evaluate visual decoding models* ***aligned with human intuition****?"*

Machine Intelligence & Data science LAB

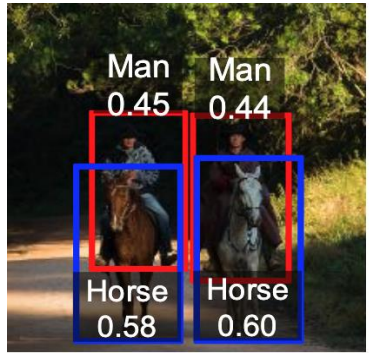
Contents

- Research motivation
- New semantic evaluation methods
- Experimental results

New semantic evaluation methods

- Object Recall, Object Precision, Object F1
  - See if "**key objects**" are detected in the reconstruction
  - For a fixed detection threshold $t \in [0,1]$

GT



Recon

$$\text{Object Recall}_t := \frac{\text{\# of categories in both GT and recon}}{\text{\# of categories in GT}}$$

$$\text{Object Precision}_t := \frac{\text{\# of categories in both GT and recon}}{\text{\# of categories in recon}}$$

Threshold t	GT Objects	Recon Objects	Recall_t	Precision_t
0.3	{Man, Horse}	{Woman, Man}	0.5	0.5
0.4	{Man, Horse}	{Woman, Man}	0.5	0.5
0.5	{Horse}	{Man}	0	0
0.6	{Horse}	∅	0	None
0.7	∅	∅	None	None

Averaged over different thresholds

$$\text{Object Recall} := \frac{1}{t_{\text{valid}}^{\text{recall}}} \int_0^{t_{\text{valid}}^{\text{recall}}} \text{Object Recall}_t \, dt$$

$$\text{Object Precision} := \frac{1}{t_{\text{valid}}^{\text{precision}}} \int_0^{t_{\text{valid}}^{\text{precision}}} \text{Object Precision}_t \, dt$$

$$\text{Object F1} := \frac{2}{\text{Object Recall}^{-1} + \text{Object Precision}^{-1}}$$

New semantic evaluation methods

- Cap-Sim
  - Generate captions using an image captioning model
  - Compare similarities between caption embeddings

  - Evaluate semantic factors that are hard to identify through the existence of objects, such as the background information or attributes of the detected object
    - e.g., pose, color

New semantic evaluation methods

- SEED
  - Integrates the three different metrics

$$\text{SEED} := (\text{Object F1} + \text{Cap-Sim} + \overline{\text{EffNet}}) \,/\, 3$$

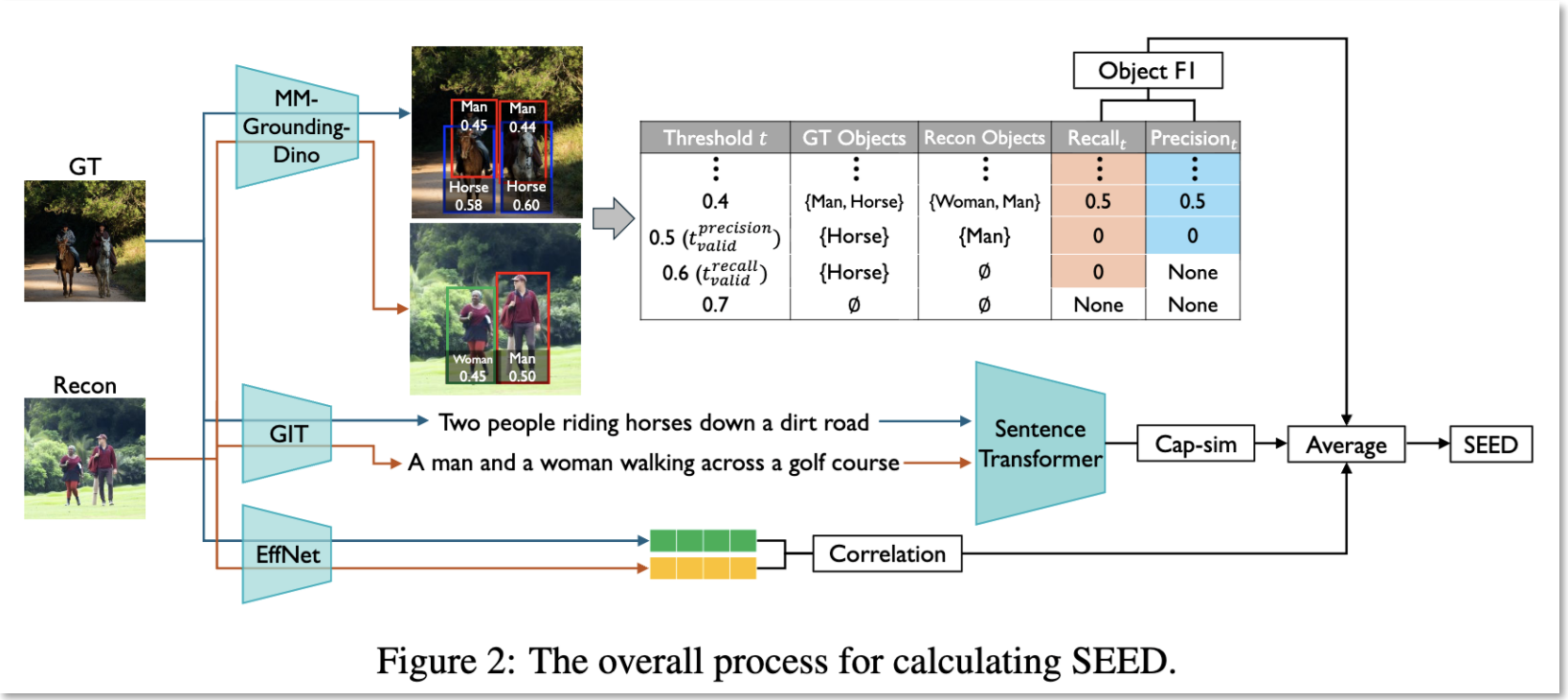


Figure 2: The overall process for calculating SEED.

Contents

- Research motivation
- New semantic evaluation methods
- Experimental results

Experimental results

- Meta-evaluation results
  - We collected 5-point Likert scale ratings from 22 human evaluators to assess the semantic similarity (GT, Recon) pairs
  - We meta-evaluated each metric with the human judgments using three metrics
    - Pairwise accuracy
    - Kendall's Tau
    - Pearson correlation

Table 1: The meta-evaluation results on NSD with MindEye2. The best results are **bolded**. SwAV was calculated similarly to Eq. 6.

Metric	Pairwise Acc.	Kendall	Pearson
PixCorr	53.8%	.075	.117
SSIM	54.5%	.090	.112
AlexNet(2)	55.0%	.185	.187
AlexNet(5)	49.5%	.236	.258
Inception	63.8%	.330	.475
CLIP	66.4%	.368	.436
EffNet	78.0%	.559	.748
SwAV	69.7%	.394	.576
Object F1	75.8%	.516	.708
Cap-Sim	73.8%	.477	.666
SEED	**81.0%**	**.621**	**.813**

Table 2: The meta-evaluation results of reconstructions of the GOD dataset with Mind-Vis. The best results are **bolded**.

Metric	Pairwise Acc.	Kendall	Pearson
PixCorr	51.3%	.029	.078
SSIM	49.2%	-.013	-.103
AlexNet(2)	66.0%	.377	.492
AlexNet(5)	65.8%	.423	.445
Inception	62.6%	.324	.356
CLIP	63.2%	.338	.309
EffNet	72.5%	.453	.661
SwAV	68.6%	.376	.498
Object F1	66.0%	.322	.431
Cap-Sim	68.7%	.376	.577
SEED	**73.7%**	**.477**	**.706**

Experimental results

- SEED enables failure mode discovery
  - Current decoding models still fail
    - to correctly capture fine-grained object categories while correctly capturing their supercategory
    - to reconstruct semantic details such as color and background

Figure 5: Examples of the semantic near-miss phenomenon.

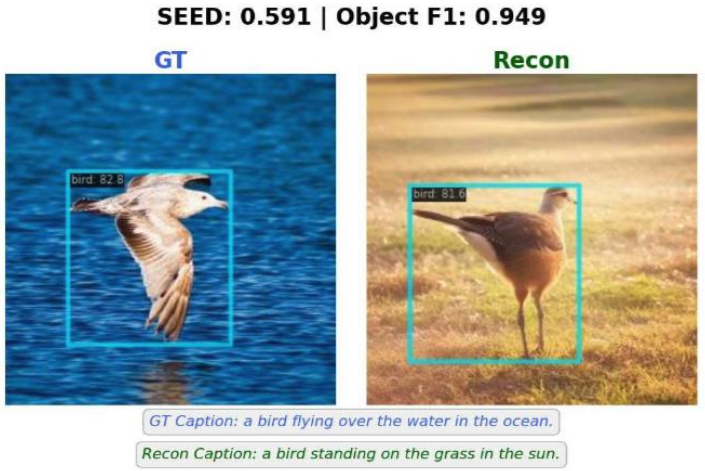


Figure 6: An example of reconstruction which captures objects correctly but misses semantic details.

Conclusion

- We proposed a **new evaluation metric, SEED**, which integrates three metrics, to **achieve the best alignment with human judgments**

- Building on this, we **discovered failure modes of current visual decoding models**, suggesting future research directions

Thank you!