Cogsci and AI - Assignment 4 report

Souvik Ghosh - 2024701022

April 2025

1 Encoder Results

Encoder	ROI	Alpha	2V2 Accuracy	Pearson Correlation
CLIP	Language	0.01	0.0226	0.7535
	Vision	0.01	0.0416	0.7939
	Task	0.01	0.0354	0.6812
	DMN	0.01	0.0289	0.6451
Small BERT (all-MiniLM-L6-v2)	Language	0.01	0.7767	0.6094
	Vision	0.01	0.7831	0.6751
	Task	0.01	0.7120	0.4054
	DMN	0.01	0.6872	0.3926
Large BERT	Language	0.01	0.8123	0.6610
	Vision	0.01	0.7985	0.6884
	Task	0.01	0.7450	0.5021
	DMN	0.01	0.7201	0.4789
T5	Language	0.01	0.7902	0.6288
	Vision	0.01	0.7750	0.6495
	Task	0.01	0.7304	0.4507
	DMN	0.01	0.7015	0.4192

Table 1: Cross-validation results for different encoders and ROIs (Alpha = 0.01). CLIP has relatively low accuracy but high correlation in language and vision ROIs.

1.1 Interpretability

we take a sentence Sentence = I hesitantly skied down the steep trail that my buddies convinced me to try. Nouns = trail buddies Verbs = skied convinced try Adjectives = steep

POS Category	Language ROI	Vision ROI	Task ROI	DMN ROI
Nouns	0.5355	**0.5889**	0.4942	0.4902
Verbs	0.5720	**0.7075**	0.2898	0.2352
Adjectives	0.2614	**0.5526**	0.1297	-0.0942

Table 2: Correlation between model prediction scores and ROI activity for different parts of speech.

Observations:

- Across all parts of speech (nouns, verbs, and adjectives), the **vision ROI consistently shows the highest correlation**, suggesting that visual features are strongly aligned with word category representations in the brain.

- **Nouns** show moderate-to-high correlations across all ROIs, with the highest in vision (0.5889), indicating that noun processing likely engages multimodal cortical areas, particularly visual and language.
- **Verbs** exhibit the strongest correlation of all results in the vision ROI (**0.7075**), implying that action-related words may evoke strong visual simulation or grounding.
- **Adjectives** have generally lower correlations, especially in the DMN ROI where the value is negative (-0.0942). The highest correlation for adjectives is again in the vision ROI (0.5526), indicating some visual alignment but weaker engagement with language or task-related regions.

2 Decoder Results

Decoder	**2V2 Acc.**	**Pearson Corr.**	**Median Rank**	**Top-1 Acc.**	**Top-5 Acc.**	**Top-10 Acc.**
CLIP	0.0000	**1.0000**	63.2000	0.0080	0.0399	0.0797
Large BERT	**0.9945**	0.3771	**2.0000**	**0.4434**	**0.7448**	**0.8581**
Small BERT	0.9933	0.3482	2.2000	0.4178	0.7240	0.8420
T5	0.9908	0.6457	3.2000	0.3238	0.6379	0.7784

Table 3: Average decoder metrics across cross-validation folds. Best value for each metric is highlighted in bold.

Observations:

- CLIP achieves a perfect Pearson correlation of **1.0** but fails on all other retrieval metrics, indicating overfitting or collapse to a degenerate solution.
- Large version of BERT shows the strongest overall performance across nearly all ranking-based metrics, with the best 2V2 accuracy (**0.9945**), median rank (**2.0**), and top-1/top-10 accuracies.
- T5 has the second-best Pearson correlation (**0.6457**) after CLIP, suggesting that its predictions are relatively smooth and correlated with ground truth embeddings, but ranking metrics are lower than BERT variants.
- Small version of BERT performs consistently well, close to large BERT on most metrics but slightly behind in correlation and top-k accuracies.