SemCLIP: A Semantic Memory-Aligned Vision Language Model

Tanveer Syeda-Mahmood[2,1], Niharika D'Souza[1], Ken C.L. Wong[1], Razi Mahmood[3], Luyao Shi[1], Ashutosh Jadhav[1], Satyananda Kashyap[1], David Beymer[1]

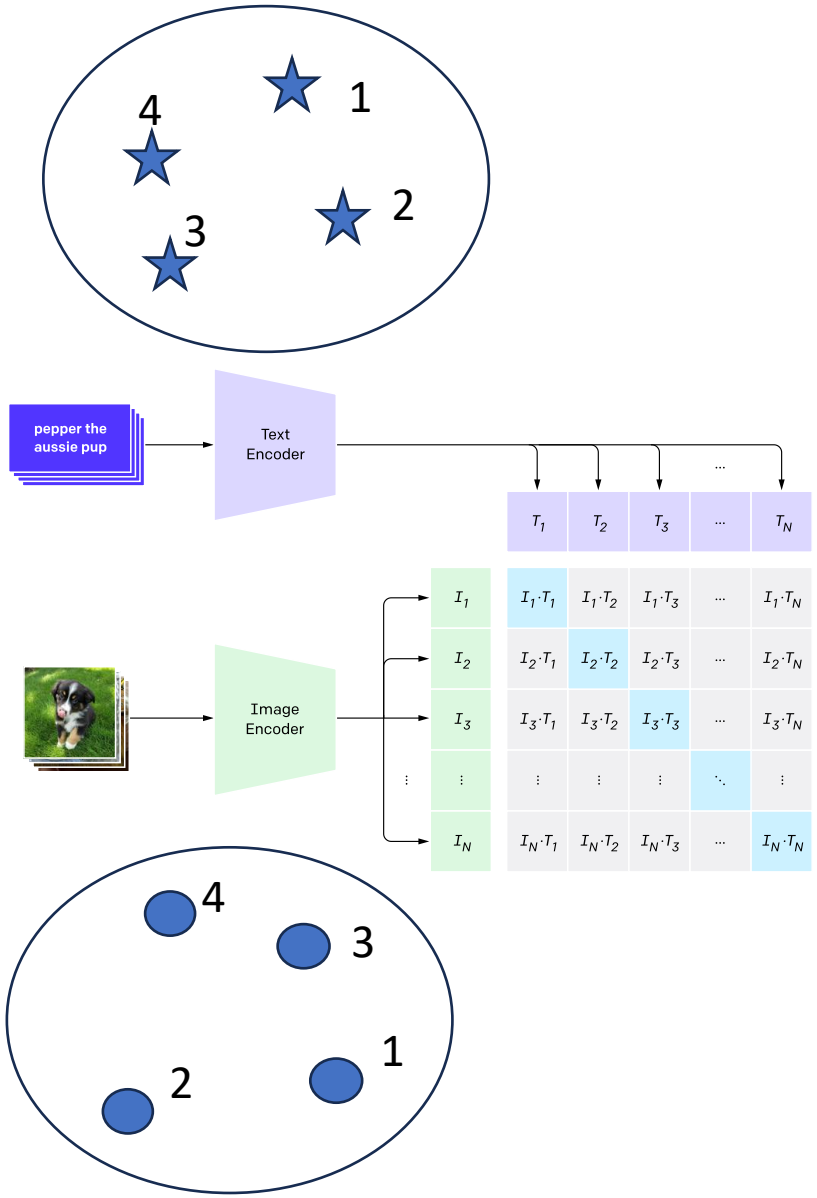
[1]IBM Research, [2]Stanford University, [3]Rensselaer Polytechnic Institute

Stanford MEDICINE

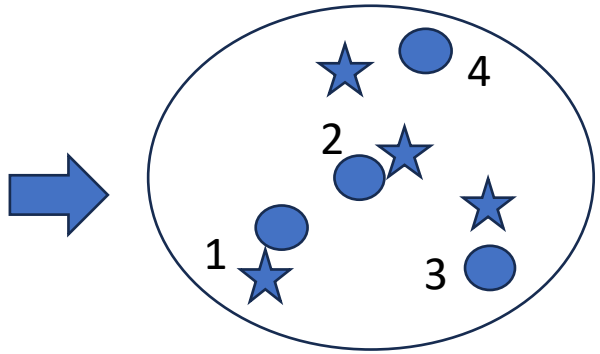
NEURAL INFORMATION PROCESSING SYSTEMS

IBM Research

Vision-Language Model – A Joint Embedding Space



VLM joint embedding

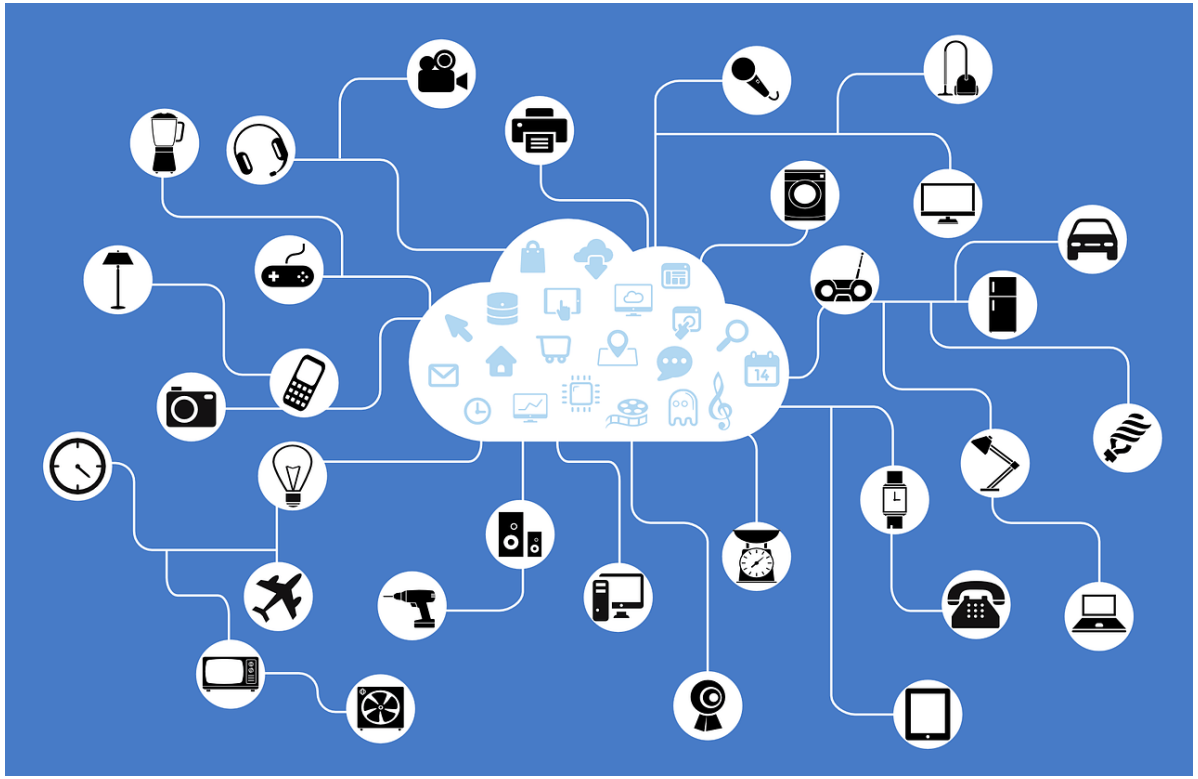


$$\mathcal{L}^{self} = \sum_{i \in I} \mathcal{L}_i^{self} = -\sum_{i \in I} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_{j(i)} / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}$$

Alignment using contrastive pre-training

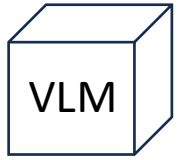
Concept general enough to beyond images to other multimodal data

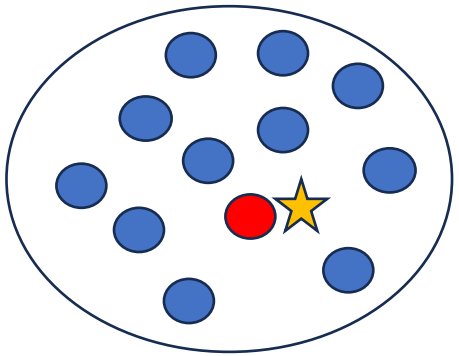
VLM as Episodic Memory



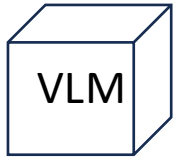
https://blog.rsisecurity.com/is-it-safe-to-store-personal-data-on-the-cloud/

Cloud store





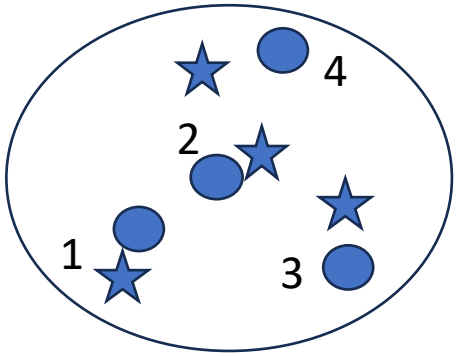
Search query

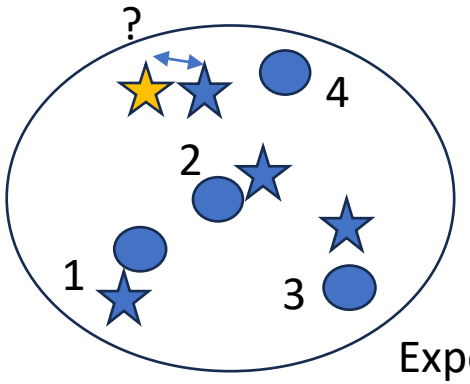


Episodic memory

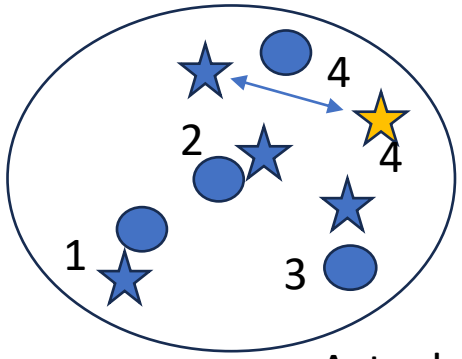
Retrieved Image

How stable is the VLM embedding?





Expected



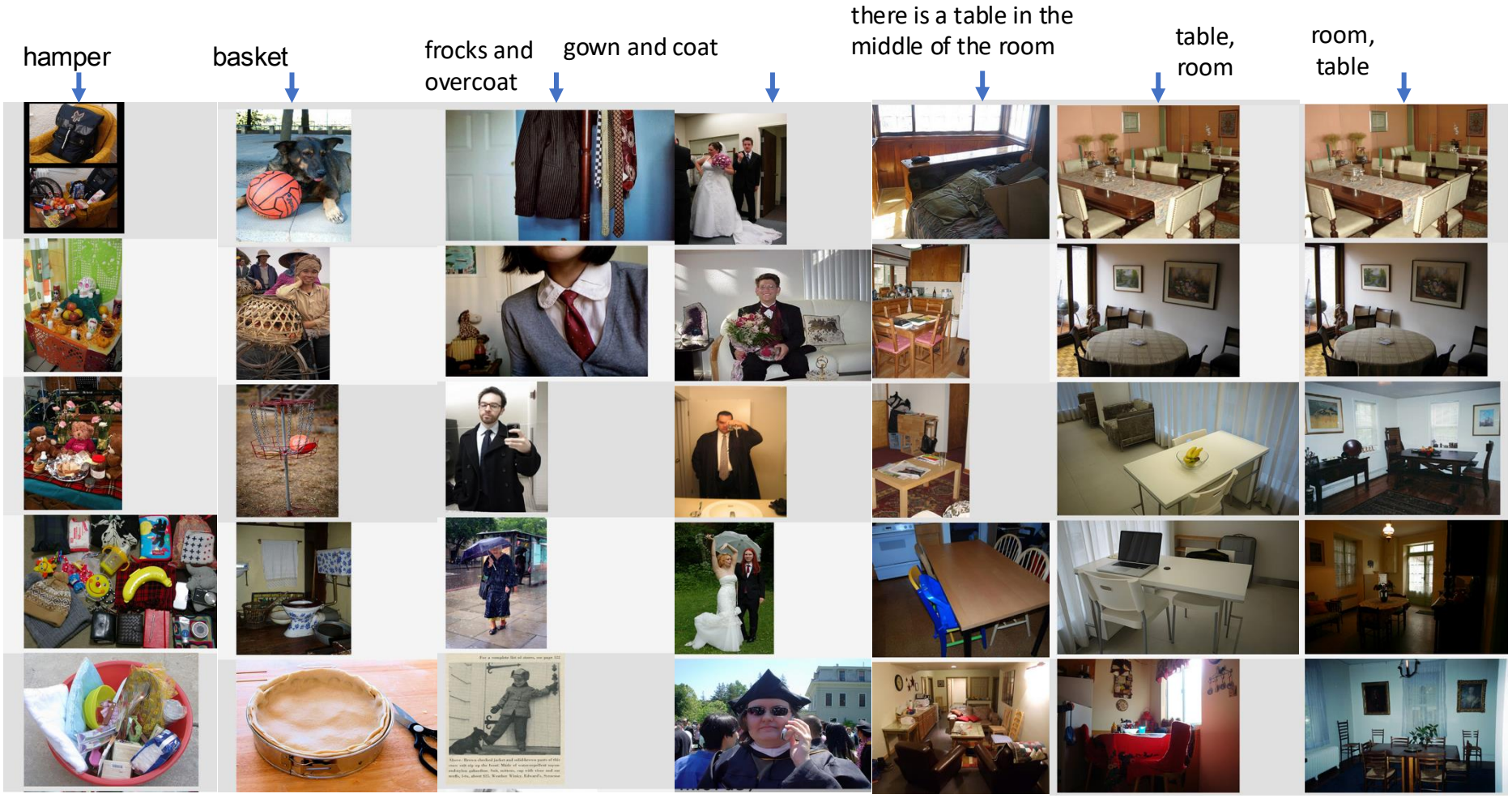
Actual

4 = Related concept to 4

Similar meaning text is neither close to the text nor its associated images

Embedding	# Queries	Synonyms in Top10	%age synonyms covered
CLIP [15]	71895	28070	49.27%
SBERT [17]	71895	37888	52.7%

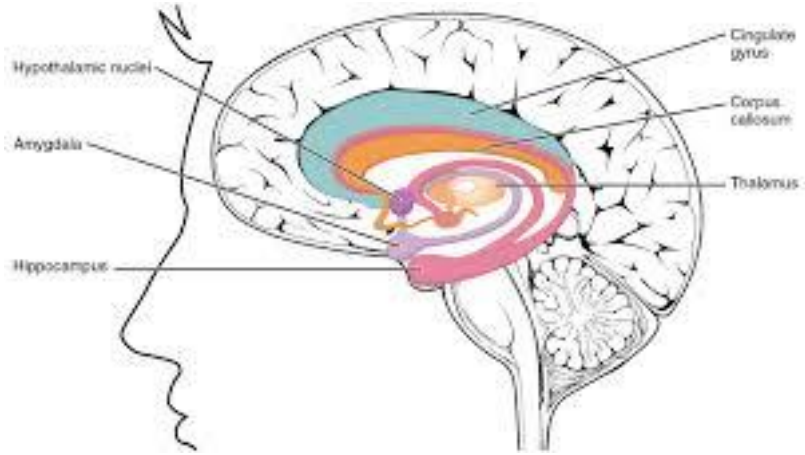
How stable are VLMs?

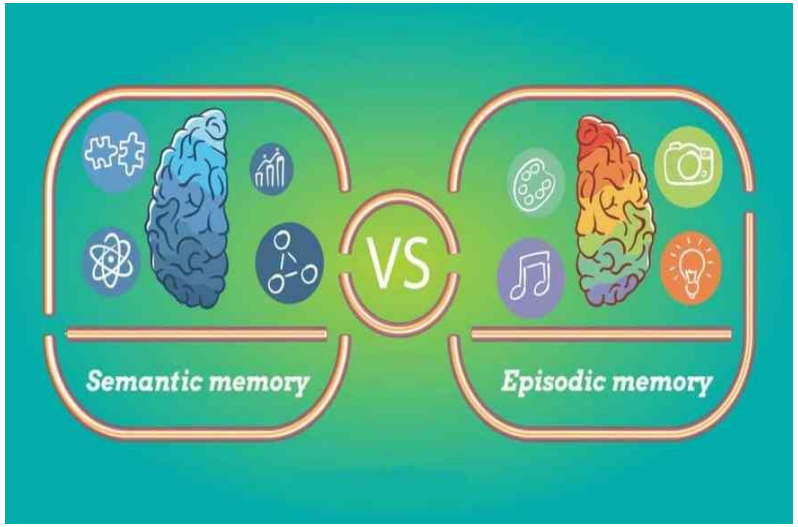


- Recall sensitive to exact query
- Semantics not well understood
- Order seems to matter

How does the brain relate concepts?

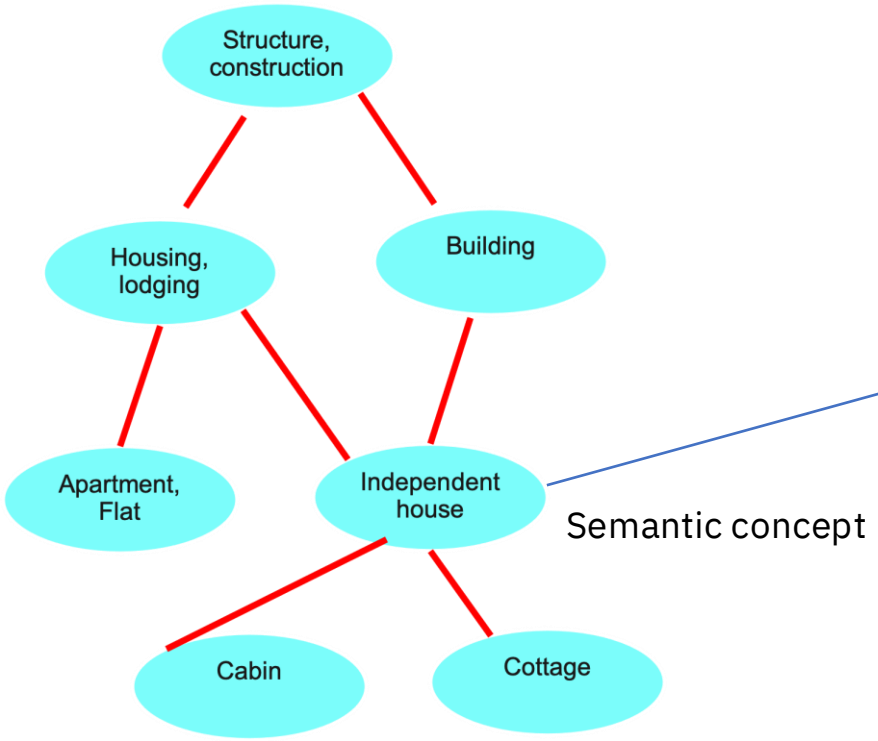
- **Vision language associations are essential to formation of memories**
  - Extensive behavioral, lesion, and functional imaging studies have demonstrated existence and interdependence of semantic and episodic memories
- **Semantic memory system:**
  - Organizes and interprets episodic memories
  - Facilitates the acquisition of new episodic memories
- **Episodic memory system:**
  - Facilitates the addition of new information to the semantic store.
  - Facilitates the retrieval of information from semantic memory
  - Uses semantic memory to form complex and detailed episodic memories





https://human-memory.net/episodic-semantic-memory/

Formation of episodic memory by association



Semantic concept

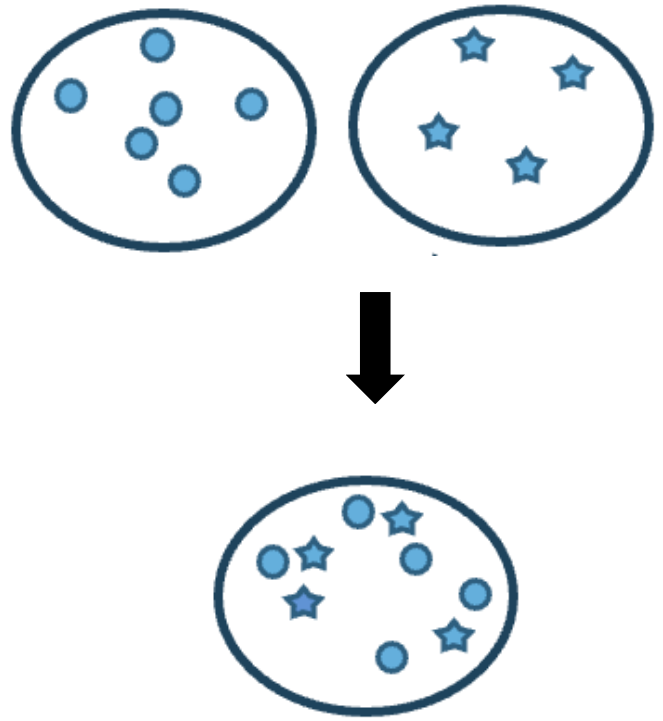
Episodic concept

John's House

A vector representing this memory of John's house

Episodic concepts

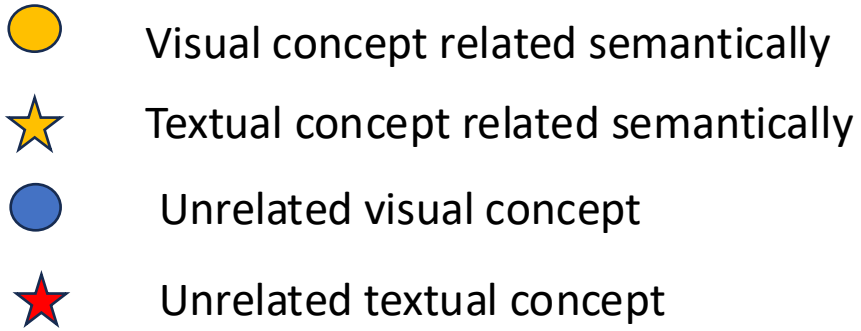
Semantic concepts

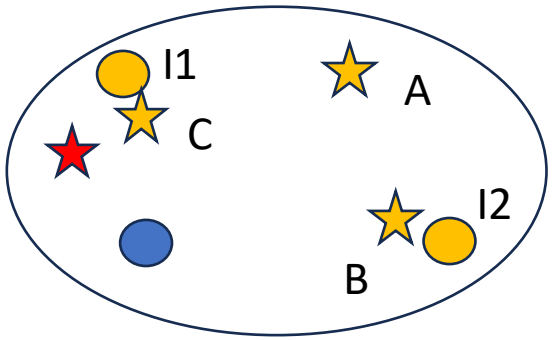


Episodic memory

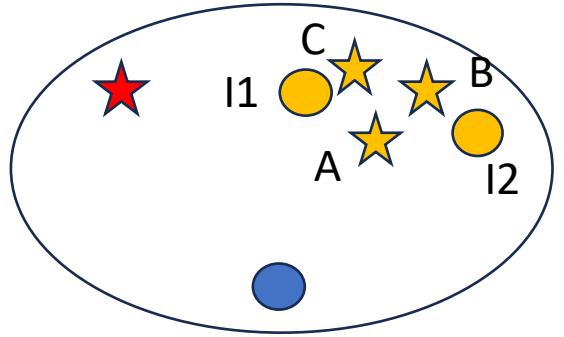
Key Ideas

- Build a semantic memory language embedding
- Semantic textual alignment
- Visual alignment into semantic memory space

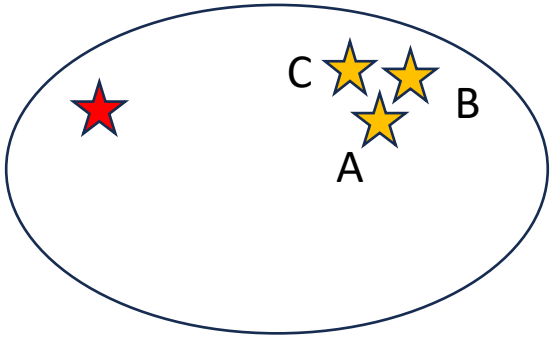


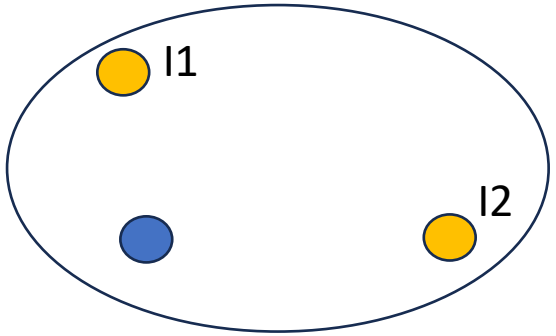


Ideal Episodic Memory transform

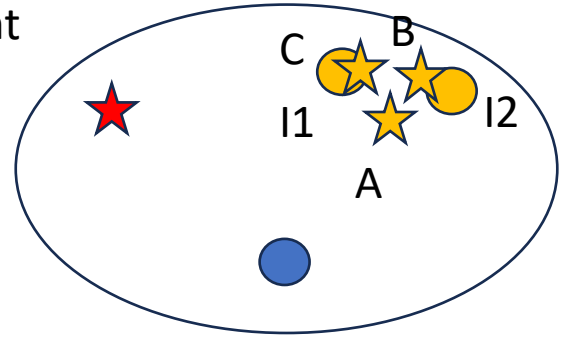


Semantic textual alignment



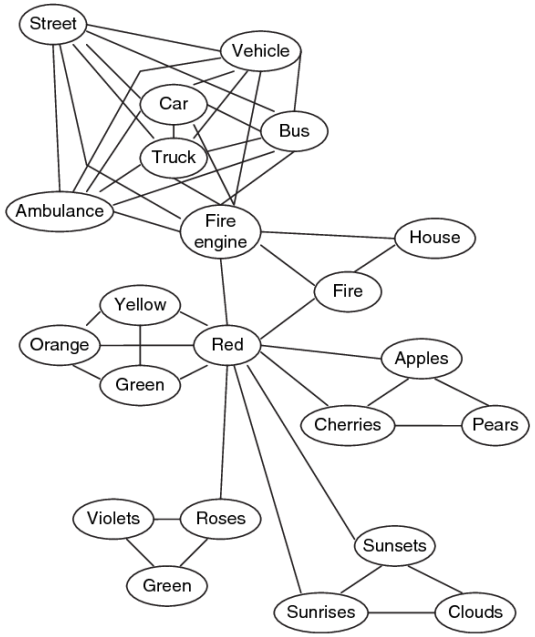


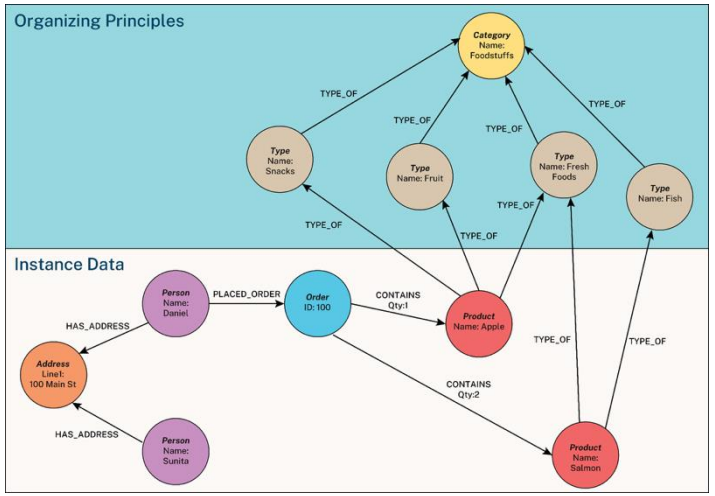
Visual Alignment

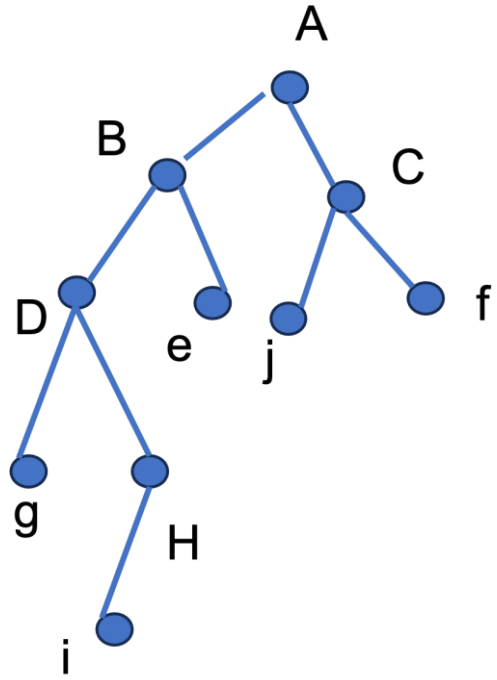


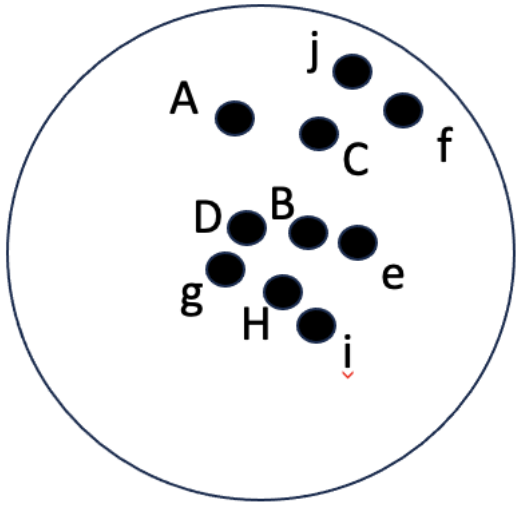
Modeling semantic memory

- Computational models have so far considered :
  - Knowledge graphs
  - Semantic networks
  - Graph embeddings
- Are there any neural models of semantic memory that capture semantic similarity?
  - Only concepts similar in meaning are close to each other



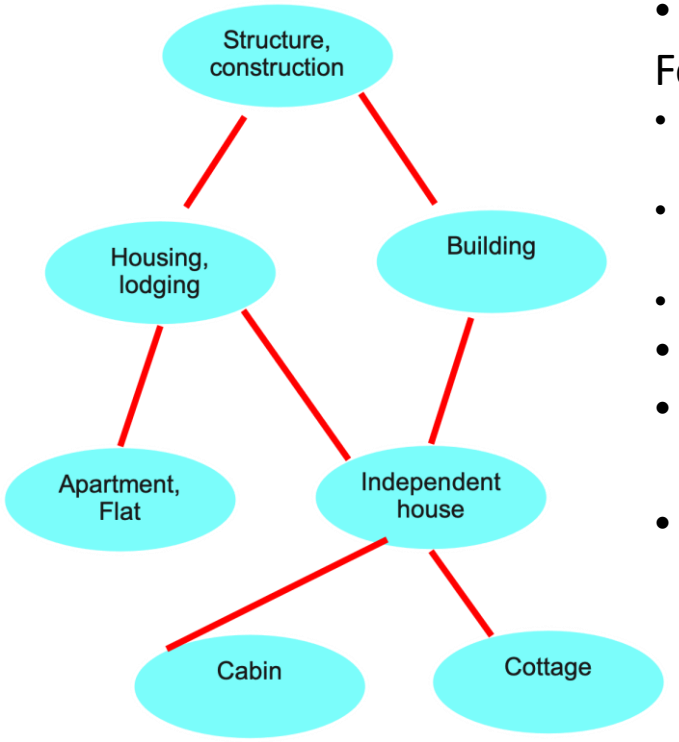


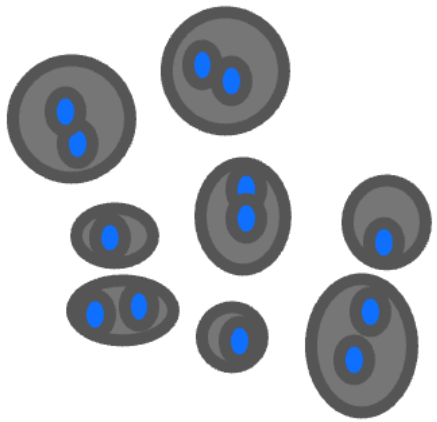




https://neo4j.com/blog/knowledge-graph/what-is-knowledge-graph/

Building semantic memory embedding





- Core vocabulary from Wordnet words:

Form similarity lists per anchor word

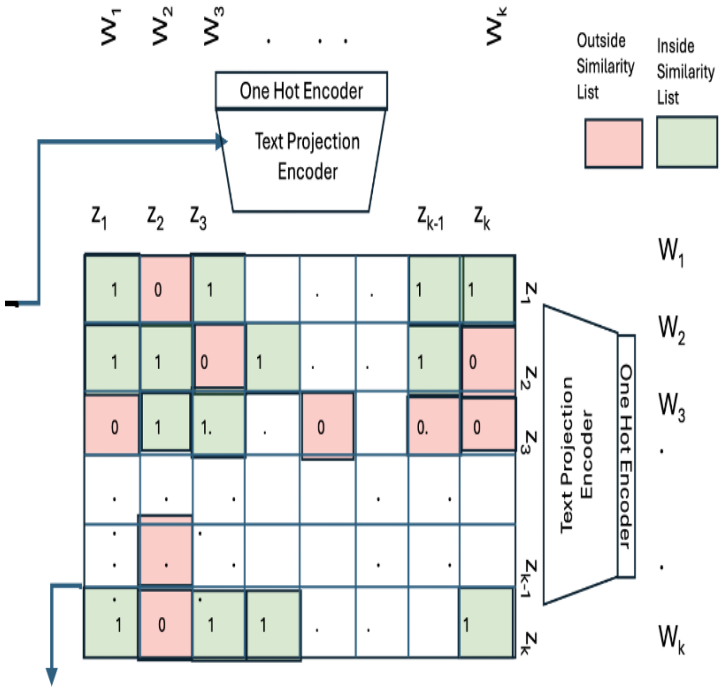
- Cabin ->(cottage, hut, shack, shanty, lodge, chalet, hovel, independent house)
- Independent house -> (detached house, single family home, single-detached dwelling, building, housing, lodging)
- Building -> (edifice, construction, erection, structure)
- All other words become negative words
- Becomes a multi-label supervised contrastive learning problem
- Non-diagonal similarity matrix

Multi-label contrastive loss

$$L_{contrast}(S_i) = \sum_{W_j \in S_i} \log \frac{\exp(z_i \cdot z_j/\tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a/\tau)}$$

$$L_{contrast} = \sum_{j}^{|V|} L_{contrast}(S_j)$$

SemOnt2Vec

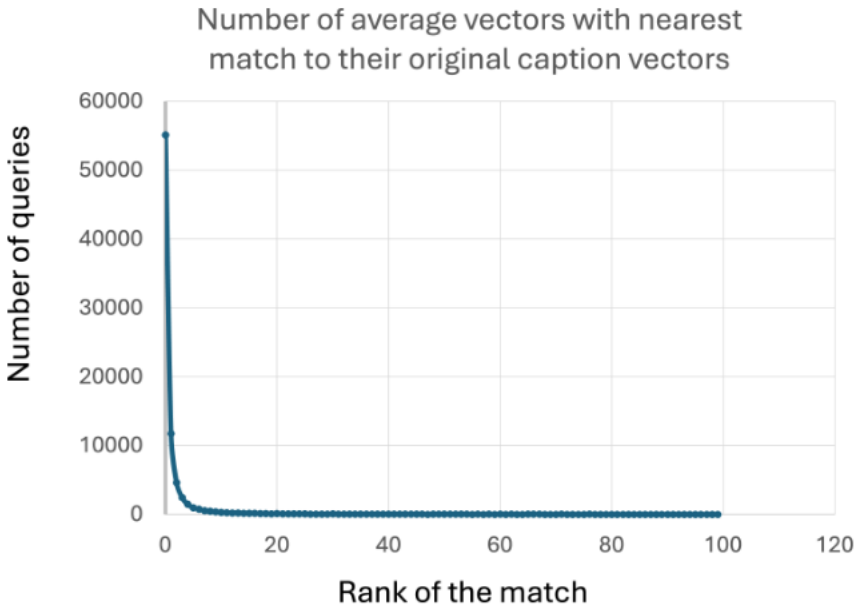


Semantically similar words are inferred from distance in SemOnt2Vec embedding

Handling full captions in semantic memory

- Break a sentence into Wordnet words and average their embeddings
  - E.g. There is a table in the middle of the room
    - Composed nouns: table, room
-
- For long captions, average embeddings of composed words were taken.
- For out-of-vocabulary words, nearest semantic match using SBERT was ta
- Word-sense disambiguation resolved the senses of nouns in the sentence

Corollary-1: Given pairs of synonymous queries Q_1, Q_2 represented by their average vectors C'_{avg1}, C'_{avg2} formed from $C'_{eq11}, ..C'_{eq1k}$ and $C'_{eq21}, ..C'_{eq2k}$, if $|C'_{eq1l} - C'_{eq2l}| < \delta$ for all C'_{eql} then $|C'_{avg1} - C'_{avg2}| < \delta$. This follows directly from vector averaging rules.



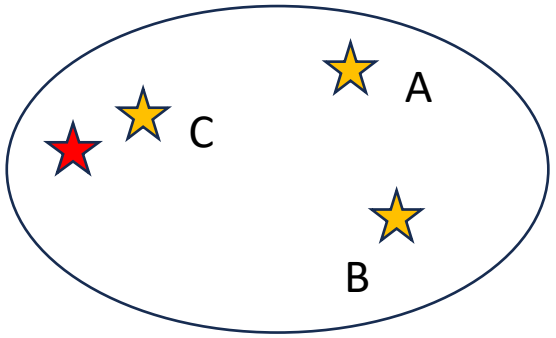
If textual embedding of full captions are close =>
Their embeddings from the average are also close

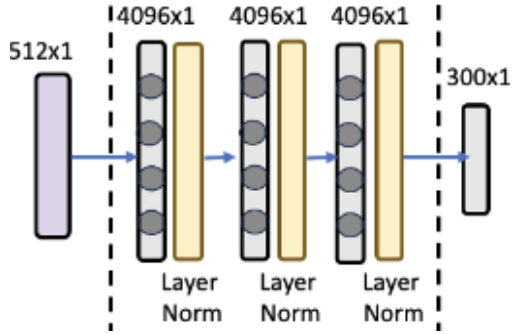
Original caption was the nearest vector for 90% of the average vectors with an
average cosine similarity of 0.958
in Visual Genome dataset.

Any arbitrary caption can be projected into the SemOnt2Vec Embedding

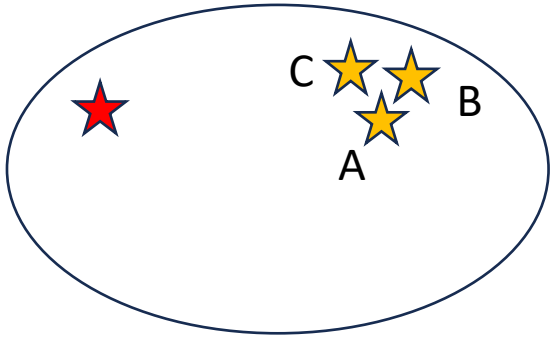
Building the transformation mapping for text

VLM embedding





STE embedding



Source text	Input	Target text	Output
Word2Vec word	VLM embedding	Word2Vec word	STE embedding
Full caption	Average VLM embedding	Full caption	Average STE embedding

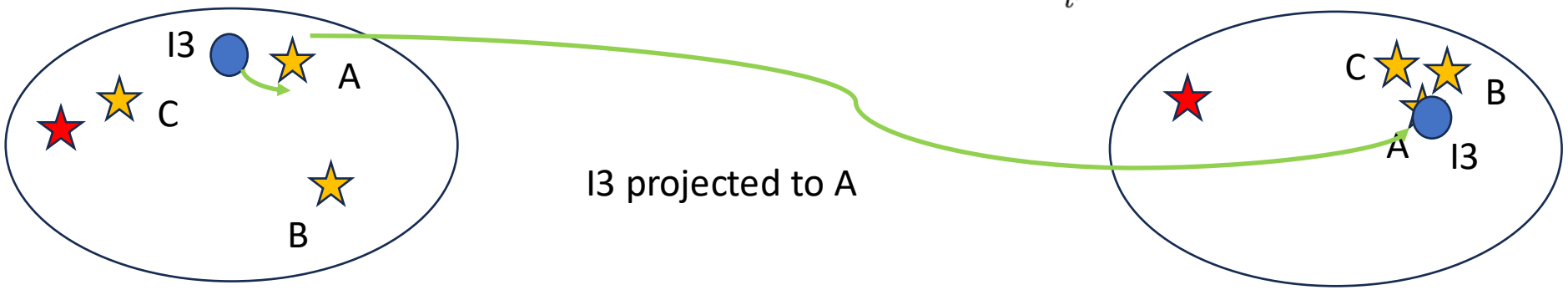
$$\underline{\text{Transform:}}\ \Gamma_t(\cdot) = \mathbf{FC}_3(\mathbf{\Phi}_{\text{relu}}(\mathbf{FC}_2(\mathbf{\Phi}_{\text{relu}}(\mathbf{FC}_1((\cdot)))))\ \ \underline{\text{Loss:}}\ \mathcal{L} = ||\Gamma_t(C_t) - \mathbf{C}_t^{'}||_2^2$$

Trained on 1 M captions + base vocabulary accumulated across WordNet, MS-COCO, Visual Genome etc.

Aligning for images

VLM embedding

$$\Gamma_i(C_i) = \Gamma_t(C_{t_i}) \quad \text{such that} \quad C_{t_i} = \arg\min_{t'} d(C_i, C_{t'})$$



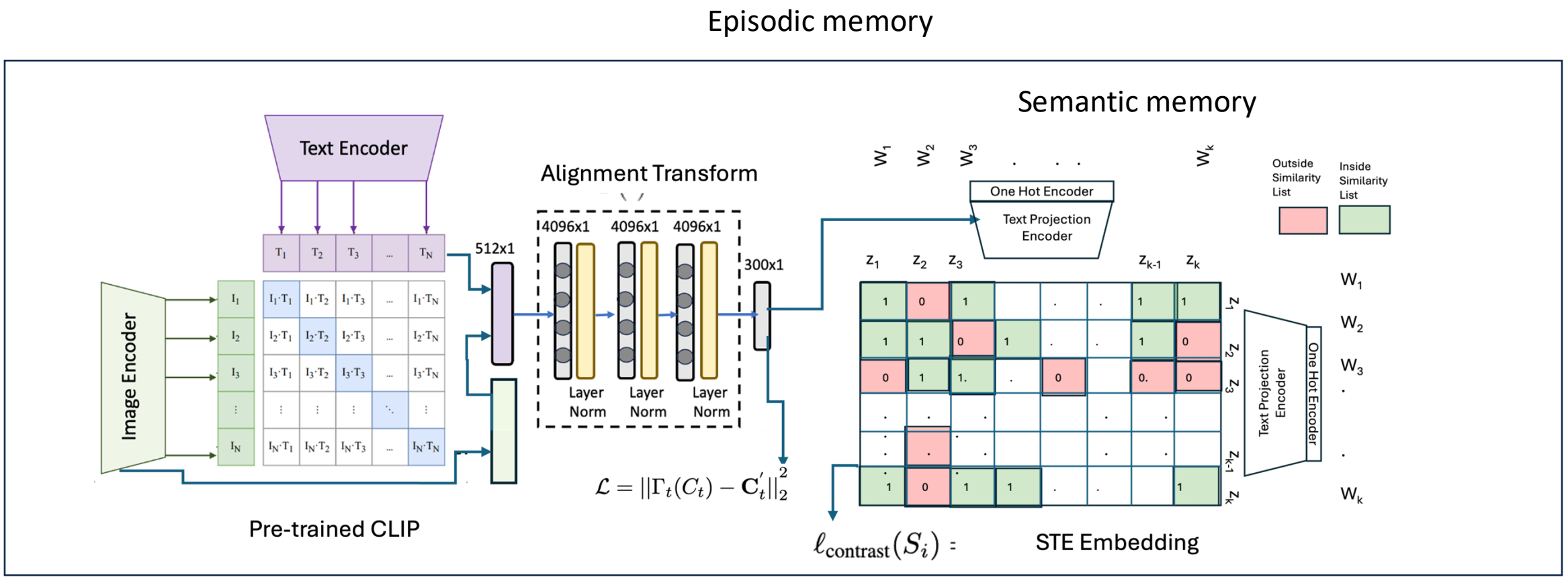
Nearest Tags for I3: A

Plain episodic memory

Nearest Tags for I3: A, B, C

Episodic memory with semantic backing

SemCLIP model



- Alignment transformer needs textual captions only
- Pre-trained STE embedding using large multi-label SupCon
- Pre-trained CLIP-variant VLM (any)
- Trainable Alignment Transform per CLIP variant using MSE loss

- Image-to-text caption
  - Two-step projection
    - Image->VLM->STE->nearest text tags
- Text-to-image retrieval
  - Images already in STE space
  - Text->STE->nearest images

Results

Dataset	Captions	Captions with Noun	Captions with Entity	Caption with Noun Phrase	Caption with Tokens
MS-COCO	568456	568372	96956	568414	568456
Visual Genome	83404	69410	14157	76121	83404
CUB	200	58	46	109	200
SUN	567	239	105	357	567
AWA	50	22	9	39	50
Wordnet		147025			
Total					799702

Semantic memory performance

How is STE embedding different from other word embeddings?

Query	Top 10 Results
van (Word2Vec)	'car', 'parking lot', 'parking meter', 'friend', 'back', 'suv', 'two', 'vehicle', 'street sign', 'front'
tree trunk (Word2Vec)	'tree trunk', 'trunk', 'tree', 'tree branch', 'pole', 'pine tree', 'ski pole', 'christmas tree', 'telephone pole', 'line'
van (STE)	'van.n.01', 'car.n.01', 'sport utility.n.01', 'jeep.n.01', 'cab.n.01', 'minivan.n.01', 'sedan.n.01', 'automotive vehicle.n.01', 'motor vehicle.n.01', 'delivery van.n.01'
tree trunk (STE)	'tree trunk.n.01', 'plant organ.n.01', 'trunk.n.01', 'stalk.n.02', 'bole.n.01', 'stem.n.02', 'wood.n.01', 'pole.n.01', 'structure.n.01'

Is this the same as using WordNet directly?

Query	**Search in Wordnet**	**Search in Semantic Embedding**
'hood	{'hood', vicinity}	{'hood', 'proximity', 'gold_coast', 'locality', 'neighbourhood', 'neck_of_the_woods', 'neighborhood', 'place', 'section', 'vicinity'}
abdominal cavity	{'abdominal_cavity', 'cavity', 'abdomen'}	{'pit_of_the_stomach', 'orbital_cavity', 'glenoid_cavity', 'cavity', 'axillary_fossa', 'abdomen', 'orbit', 'cavum', 'abdominal_cavity', 'bodily_cavity'}
erosion	{'erosion', 'ablation'}	{'erosion', 'deflation', 'wearing_away', 'ablation', 'detrition', 'eroding', 'abrasion', 'attrition', 'eating_away', 'wearing'}
dorm room	{'dorm_room', 'dormitory_room', 'dormitory', 'bedroom'}	{'dormitory_room', 'chamber', 'sleeping_accommodation', 'master_bedroom', 'dormitory', 'dorm_room', 'sleeping_room', 'bedroom', 'bedchamber', 'guestroom'}

Semantic memory embedding performance

- 13 benchmark datasets
- Semantic dissimilarity is respected for datasets that contain antonyms

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman's rank correlation coefficient

d_i = difference between the two ranks of each observation

n = number of observations

Spearman correlation coefficient

Datasets	Original Word #	WordNet Filtered #	Word2Vec	Glove	SBERT	Path2Vec	SME	SME (similar subset)
EM_SIMLEX_SYNS	297	297	0.285	0.240	0.145	0.301	0.265	**0.570**
EN-MC-30	30	30	**0.789**	0.702	0.410	**0.782**	0.650	0.650
EN-MEN-TR-3k	3000	2657	**0.776**	0.743	0.310	0.366	0.257	**0.780**
EN-MTurk-287	287	243	0.767	0.705	0.435	0.317	0.300	**0.810**
EN-MTurk-771	771	771	0.671	0.649	0.335	0.404	0.466	**0.760**
EN-RG-65	65	64	0.761	0.770	0.446	0.723	0.640	**0.820**
EN-RW-STANFORD	2034	910	0.492	0.341	0.226	0.194	0.217	**0.590**
EN-SIMLEX	666	666	0.452	0.397	0.233	0.505	0.398	**0.670**
EN-WS-353-REL	252	248	**0.626**	0.578	0.159	0.136	0	0
EN-WS-353-SIM	203	201	0.774	0.659	0.388	0.599	**0.820**	**0.820**
EN-YP-130	130	43	0.542	0.545	0.326	0.029	0.426	**0.660**
EW-WS-353-Syns	99	98	0.507	0.507	0.366	0.616	**0.655**	**0.655**
EN-WS-353-ALL	352	348	0.694	0.607	0.256	0.406	0.303	**0.720**

Comparing SemCLIP to VLMs

SemCLIP model is more semantically stable

Embedding	# Queries	Synonyms in Top10	%age synonyms covered
CLIP [15]	71895	28070	49.27%
SBERT [17]	71895	37888	52.7%
Ours	71895	67309	87.7%

SemCLIP model is more semantically stable

Dataset	Images/Queries	Method	Overlap@1	Overlap@5	Overlap@10	Overlap@50
Visual Genome	7794 / 14513	**SemCLIP**	**0.551**	**0.532**	**0.517**	**0.523**
		CLIP	0.119	0.056	0.038	0.026
		OpenCLIP	0.129	0.055	0.040	0.025
		BLIP	0.119	0.056	0.037	0.024
		NegCLIP	0.109	0.063	0.038	0.024
CUB	11788 / 200	**SemCLIP**	**0.812**	**0.783**	**0.732**	**0.715**
		CLIP	0.118	0.072	0.062	0.053
		OpenCLIP	0.149	0.079	0.062	0.053
		BLIP	0.181	0.084	0.066	0.053
		NegCLIP	0.119	0.078	0.066	0.055
SUN	16657 / 567	**SemCLIP**	**0.554**	**0.531**	**0.523**	**0.511**
		CLIP	0.092	0.048	0.034	0.025
		OpenCLIP	0.070	0.039	0.028	0.021
		BLIP	0.091	0.05	0.035	0.025
		NegCLIP	0.095	0.045	0.032	0.023
AWA2	6985 / 10	**SemCLIP**	**0.751**	**0.723**	**0.702**	**0.715**
		CLIP	0.086	0.051	0.038	0.028
		OpenCLIP	0.139	0.067	0.046	0.032
		BLIP	0.139	0.057	0.039	0.028
		NegCLIP	0.101	0.046	0.034	0.025

SemCLIP model has higher retrieval quality

Dataset	Images / Labels	Model	t2i: NDCG / mAP / Recall (@10)	i2t: NDCG / mAP / Recall (@10)
Visual Genome	7794 / 14513	**SemCLIP**	**0.192 / 0.172**	**0.254 / 0.185**
		CLIP	0.053 / 0.060 / 0.065	0.050 / 0.129 / 0.028
		OpenCLIP	0.066 / 0.075 / 0.081	0.061 / 0.159 / 0.035
		BLIP	0.072 / 0.081 / 0.088	0.069 / 0.167 / 0.039
		NegCLIP	0.063 / 0.072 / 0.077	0.060 / 0.150 / 0.035
		PosCLIP	0.074 / 0.084 / 0.091	0.060 / 0.146 / 0.034
CUB	11788 / 200	**SemCLIP**	**0.721 / 0.845**	**0.891 / 0.812**
		CLIP	0.513 / 0.621 / 0.084	0.619 / 0.554 / 0.826
		OpenCLIP	0.669 / 0.744 / 0.111	0.777 / 0.726 / 0.935
		BLIP	0.204 / 0.341 / 0.033	0.326 / 0.260 / 0.543
		NegCLIP	0.406 / 0.535 / 0.065	0.472 / 0.403 / 0.694
		PosCLIP	0.488 / 0.580 / 0.080	0.553 / 0.479 / 0.788
SUN	16657 / 567	**SemCLIP**	**0.686 / 0.712**	**0.810 / 0.671**
		CLIP	0.414 / 0.562 / 0.191	0.458 / 0.415 / 0.595
		OpenCLIP	0.549 / 0.664 / 0.260	0.514 / 0.476 / 0.634
		BLIP	0.413 / 0.535 / 0.194	0.426 / 0.384 / 0.557
		NegCLIP	0.429 / 0.553 / 0.201	0.425 / 0.380 / 0.567
		PosCLIP	0.463 / 0.602 / 0.214	0.445 / 0.399 / 0.589
AWA2	6985 / 10	**SemCLIP**	0.967 / 0.987	0.995 / 0.989
		CLIP	0.993 / 0.999 / 0.016	0.992 / 0.989 / 1.000
		OpenCLIP	**1.000** / **1.000** / 0.016	**0.994** / **0.991** / 1.000
		BLIP	**1.000** / **1.000** / 0.016	0.991 / 0.988 / 1.000
		NegCLIP	**1.000** / **1.000** / 0.016	0.987 / 0.982 / 1.000
		PosCLIP	**1.000** / **1.000** / 0.016	0.983 / 0.977 / 1.000
COCO	5000 / 80	**SemCLIP**	**0.940** / 0.91	**0.895 / 0.923**
		CLIP	0.810 / 0.869 / 0.081	0.706 / 0.771 / 0.745
		OpenCLIP	0.866 / 0.926 / 0.087	0.756 / 0.788 / 0.799
		BLIP	0.897 / 0.942 / 0.089	0.774 / 0.852 / 0.781
		NegCLIP	0.834 / 0.892 / 0.083	0.766 / 0.797 / 0.808
		PosCLIP	0.919 / 0.946 / 0.088	0.807 / 0.834 / 0.843
Flickr30k	31014 / 158391	**SemCLIP**	**0.571 / 0.523**	**0.580 / 0.572**
		CLIP	0.354 / 0.306 / 0.510	0.314 / 0.430 / 0.320
		OpenCLIP	0.427 / 0.377 / 0.588	0.376 / 0.489 / 0.382
		BLIP	0.562 / 0.510 / 0.725	0.475 / 0.587 / 0.480
		NegCLIP	0.425 / 0.373 / 0.591	0.354 / 0.465 / 0.360
		PosCLIP	0.323 / 0.279 / 0.468	0.273 / 0.367 / 0.283

Conclusions

- An embedding modeling semantic memory is important for capturing semantic relationships between concepts in text
  - Multi-label supervised contrastive learning captures ontological structure
- Episodic memory developed using semantic memory substrate gives more semantic stability
- SemCLIP a new VLM integrating episodic and semantic memory