Tversky Neural Networks: Psychologically Plausible Deep Learning with Differentiable Tversky Similarity

Moussa Koulako Bala Doumbouya

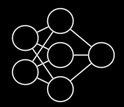
Dan Jurafsky

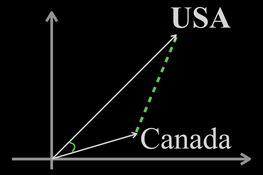
Christopher D. Manning

Similarity: Geometric ≠ Psychological

- Geometric similarity is standard in neural networks
  - However it is incompatible with human cognition
- Tversky similarity better models human judgement
  - Goal: How can it be used as a basic neural network operation ?

Geometric Similarity



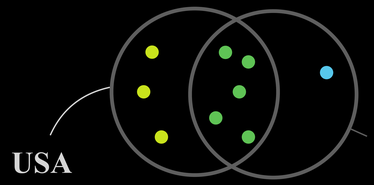


vectors

sim(A, B) = cosine | dot product | euclidian proximity | etc.

$$sim(Canada, USA) = sim(USA, Canada)$$

Tversky Similarity



sets

$$sim(A, B) = \theta f(A \cap B) - \alpha f(A - B) - \beta f(B - A)$$

$$sim(Canada, USA) > sim(USA, Canada)$$

Tversky Neural Networks

Dual Vector - Set Representation

Learnable Feature bank: $\Omega = \{f_k\}_{k=1}^{|\Omega|}$

$$x \in \mathbb{R}^d \quad \equiv \quad X = \{f_k \in \Omega \,|\, x \cdot f_k > 0\}$$

vector *set*

Salience

$$f(A) = \sum_{k=1}^{|\Omega|} a \cdot f_k \cdot \mathbf{1}[a \cdot f_k > 0]$$

Set Operations

$$f(A \cap B) = \sum_{k=1}^{|\Omega|} \Psi(a \cdot f_k, b \cdot f_k) \times \mathbf{1}[a \cdot f_k > 0 \wedge b \cdot f_k > 0]$$

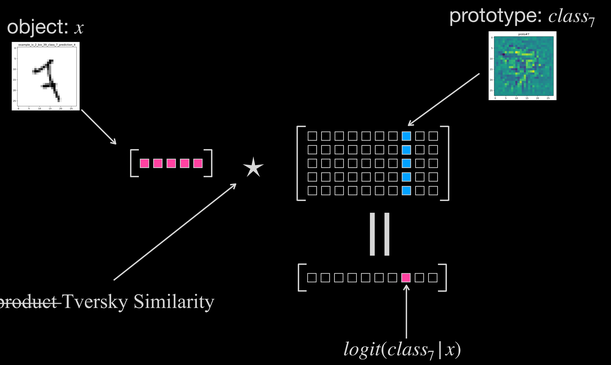
$$f^i(A - B) = \sum_{k=1}^{|\Omega|} (a \cdot f_k) \times \mathbf{1}[a \cdot f_k > 0 \wedge b \cdot f_k \leq 0]$$

Tversky Similarity Layer

$$\mathcal{S}^{\Omega,\alpha,\beta,\theta}(a,b): \begin{cases} \mathbb{R}^d \times \mathbb{R}^d & \longrightarrow \mathbb{R} \\ (a,b) & \longmapsto \begin{bmatrix} \theta \\ -\alpha \\ -\beta \end{bmatrix} \cdot \begin{bmatrix} f(A \cap B) \\ f(A - B) \\ f(B - A) \end{bmatrix} \end{cases}$$

Tversky Projection Layer

$$\mathcal{P}^{\Omega,\alpha,\beta,\theta,\Pi}(a): \begin{cases} \mathbb{R}^d & \longrightarrow \mathbb{R}^p \\ a & \longmapsto \begin{bmatrix} \mathcal{S}^{\Omega,\alpha,\beta,\theta}(a,\Pi_0) \\ \mathcal{S}^{\Omega,\alpha,\beta,\theta}(a,\Pi_1) \\ \vdots \\ \mathcal{S}^{\Omega,\alpha,\beta,\theta}(a,\Pi_{p-1}) \end{bmatrix} \end{cases}$$



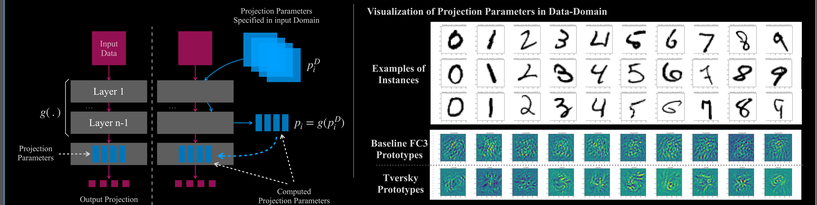
Better Performance

Resnet50	MNIST ACC	NABirds ACC
Baseline	99.54 ± 0.04	62.84 ± 0.45
Tversky	99.56 ± 0.06	65.20 ± 0.26

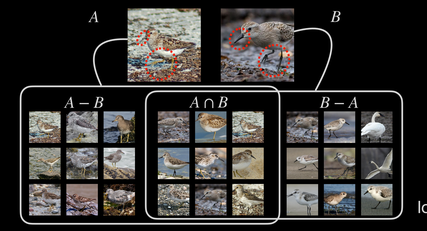
GPT-2	Params	Perplexity
Baseline	124M	113
Tversky	81M (−35%)	104 (−8%)

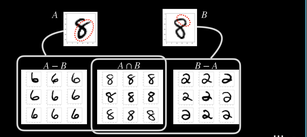
Better Interpretability

Explainable Prototypes



Algebraic Semantic Fields





Concept	Semantic Expression	Top-scoring tokens
Adjective	*bad ∩ good – worse – better – worst – best*	*lousy, evil, crappy, poor, shitty, nice, terrible, horrible, decent, mediocre, valid, great, excellent*
Comparative	*worse ∩ better – bad – good – worst – best*	*sharper, nicer, clearer, smarter, wiser, smoother, safer, hotter, happier, tighter, louder, preferable*
Superlative	*worst ∩ best – worse – better – bad – good*	*happiest, safest, quickest, deadliest, finest, hottest, busiest, toughest, darkest, fastest, brightest, holiest*
Industrial Plant	*plant – aceae – vegetation – vegetable – planting – flower – herb – vine – crop – tree – mushroom – plantation – leaves*	*facility, stall, brate, factory, Laboratories, Shed, implant, Berm, Sew, Manufacturing, reactor, Diesel, refinery, strain, distribut, planting, Industrial, microbial, actory, reactors, Unit, Install, depot, spill, Cutter, Indust, Bot, Nuclear, ineries*

like

like – love
like, alike, likes, akin, unlike, incent, agre, suspic, liked, esque, such, reminiscent, resembling, analogous

love ∩ like
love, like, agre, arrang, comr, incent, trave, reluct, loves, adore, neighb

love – like
love, loves, adore, loving, loved, affection, passion, hate, xual, despise, romance, lover, lovers, reverence, hates, hatred, hating, loving, jealousy, dislike, crave, cherish, adultery, kisses, romantic, lov, passion, fascination, admiration, devotion, kissing, friendship, marriage, lust, passions, romance, longing, desire, obsession

love

ICLR 2026 Rio de Janeiro, Brazil
The Fourteenth International Conference on Learning Representations

Stanford University