視覚的に解釈できる埋め込み空間

ACL2026:
Mapping the Circumplex of Affect: Geometric Analysis of Emotion Representations via Hyperspherical Contrastive Learning

2026/11/04 NLPコロキウム

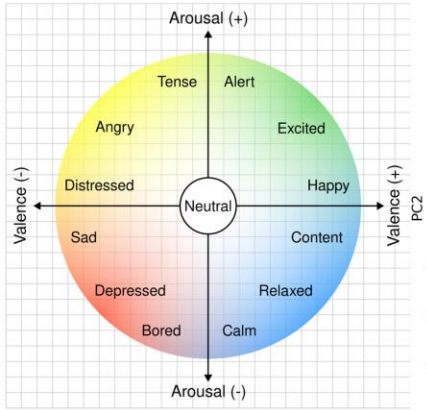
山内 悠輔

自己紹介

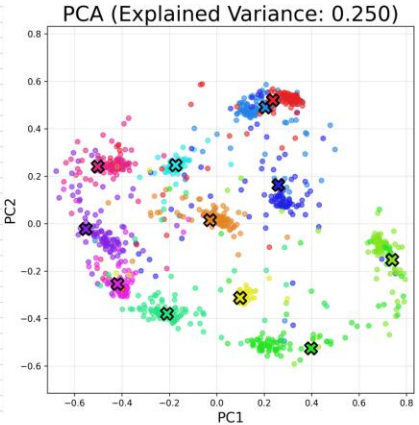
山内悠輔 (Yusuke Yamauchi)

- 所属
  - 東京大学 宮尾研究室 博士1年
- 研究分野
  - 大規模言語モデルの解釈性
    - **モデルの仕組みを埋め込み表現の幾何の観点から理解したいと考えています**

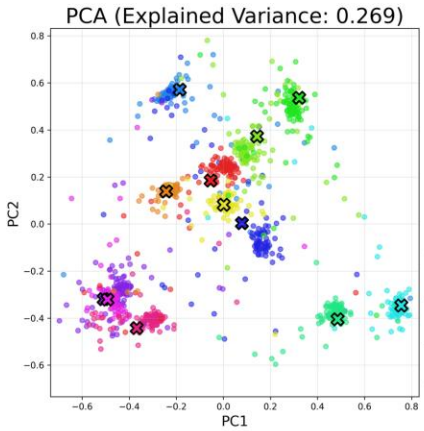
論文の概要：心理学で提唱される感情モデルの再現



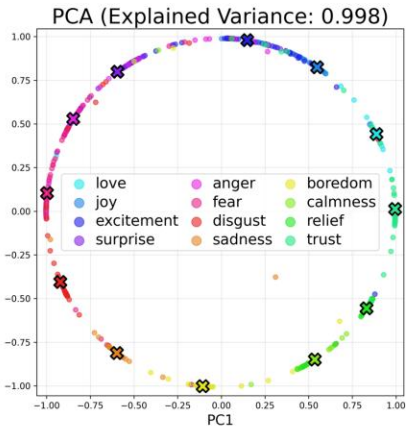
(a) Circumplex model example



(b) SINCERE-nGPT



(c) SoftCSE-nGPT



(d) CircularCSE-nGPT

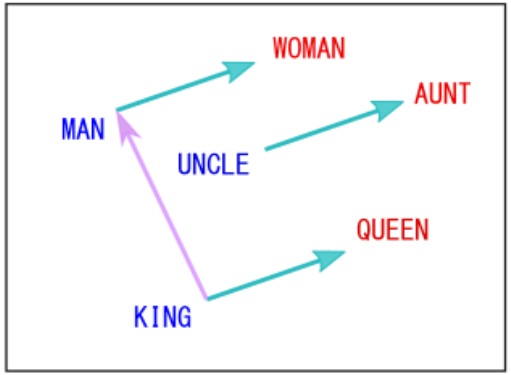
- 言語モデルの**埋め込み空間は解釈(視覚化)が難しい** (左図b, c)
  - 次元削減を行うと情報損失が大きい
  - 人間が想定する感情構造(左図a)と形が全く一致しない

⇒ 心理学で提唱される感情構造を再現するように埋め込みモデルを訓練するフレームワークを提案

- 埋め込み空間を視覚的に解釈できるようになった👍 (左図d)
- 埋め込みの解釈性(可視化性)の向上には構造上識別精度の劣化が避けられない⚖️ことを発見した

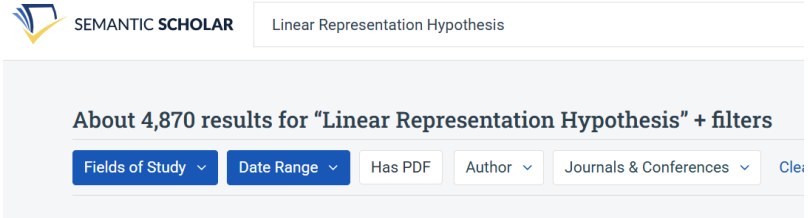
背景：線形表現仮説 (Linear Representation Hypothesis)

- 言語モデルの埋め込み空間では概念が線形的な方向として表現される
  - あるいは、ある方向に沿って介入を行えば、対象概念に関連するモデルの振る舞いが変化する



Word2vecでおなじみの例．ここでは↗↙方向が性別に対応

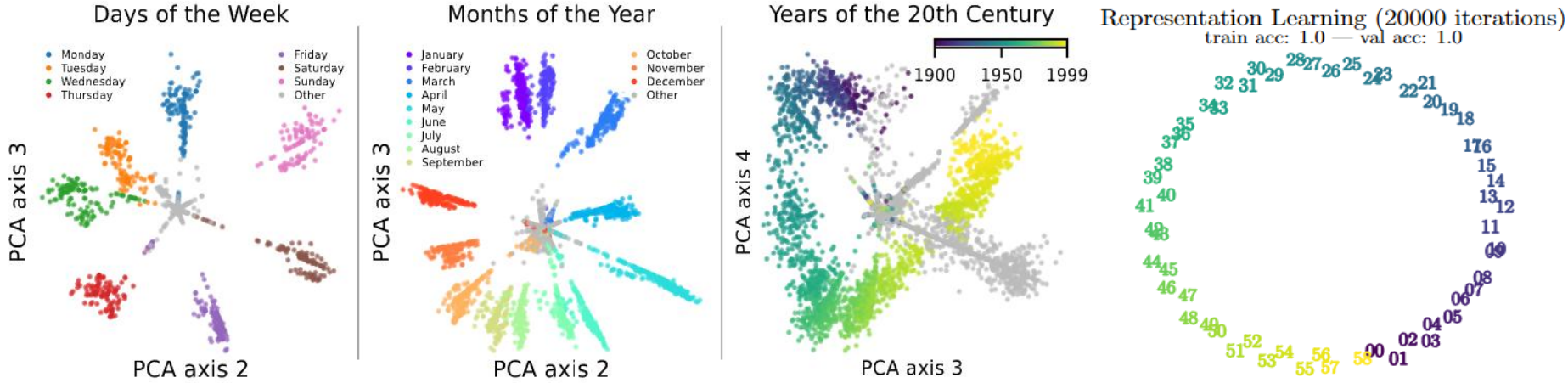
- 直感的に分かりやすく、多くの研究手法の基礎になっている



2026年現在でも主流な考え方

Park et al., “The Linear Representation Hypothesis and the Geometry of Large Language Models”, arXiv (2024)
https://arxiv.org/abs/2311.03658

全ての概念が線形的に表現されるとは限らない



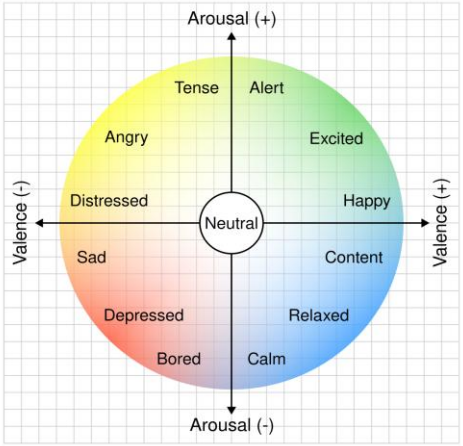
- “週”や”月”のような周期性のある概念や、mod演算を学習したモデルの数値表現は円形に並んでいることが発見された
  - 一方で、これら以外に明確な非線形構造を可視化できる概念は (私が知る限り) 見つかっていない
- 他に非線形に表現される概念は存在しないのだろうか？🤔

Engels et al., “Not All Language Model Features Are One-Dimensionally Linear” arXiv (2025)
https://arxiv.org/abs/2405.14860

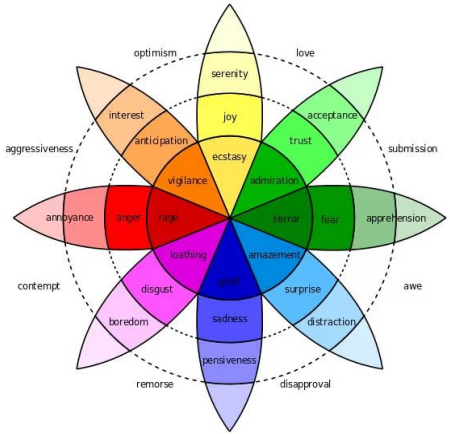
Liu et al., “Towards Understanding Grokking: An Effective Theory of Representation Learning” arXiv (2022)
https://arxiv.org/pdf/2205.10343#page=7.67

感情は非線形の構造をとっている？

ラッセルの円環感情モデル

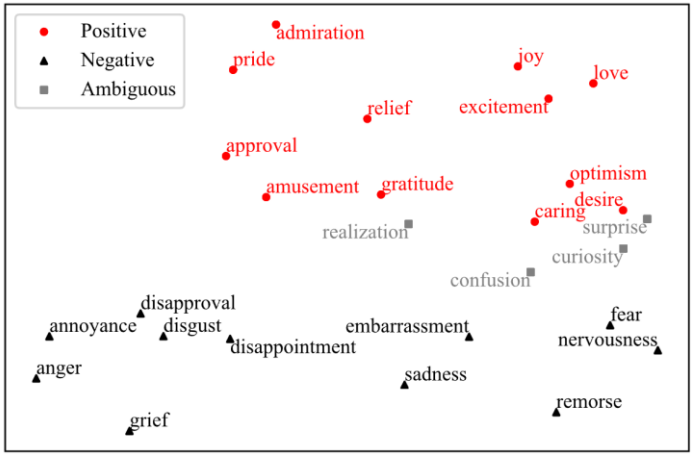


プルチックの感情の輪



gap

言語モデルが表現する感情構造



- どうやら心理学では感情を円形構造としてモデル化することが一般的らしい
  - 言語モデルでは線形的に表現される概念の代表例

- 心理学で提唱される感情構造を再現するように埋め込みを学習すれば
  精度も解釈性も上がるのでは! というのが本研究のResearch Question

Wang et al.,“Learning Emotion Category Representation to Detect Emotion Relations Across Languages”, TPAMI (2025)
https://ieeexplore.ieee.org/document/10902477

関連研究： 感情ラベルの座標の直接的な学習

- 既存研究には座標がアノテーションされている
  データセットを使用して、回帰学習を行うものがある
  - 連続値のラベルがついているものは希少
  - 実際に学習してみると思ったほど円環にならない

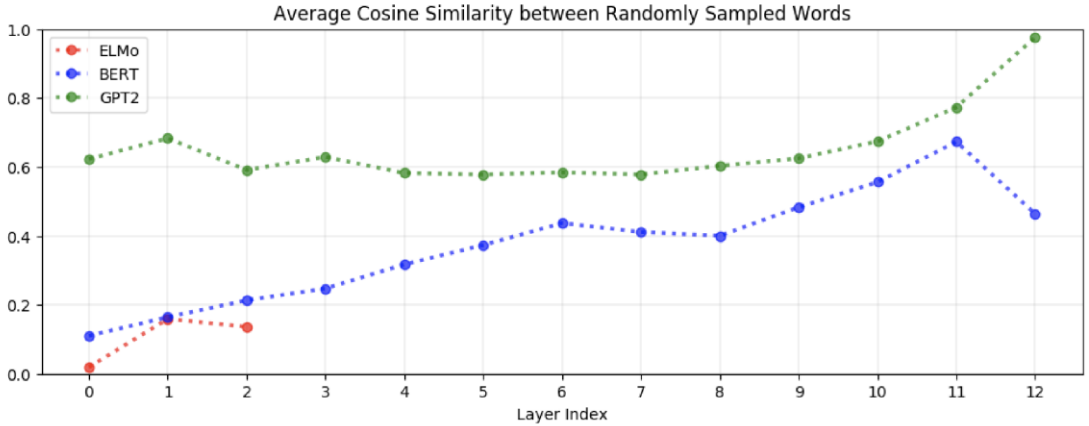
⇒ 円環を視覚的に再現できるようにするためには
幾何構造への強い制約が必要

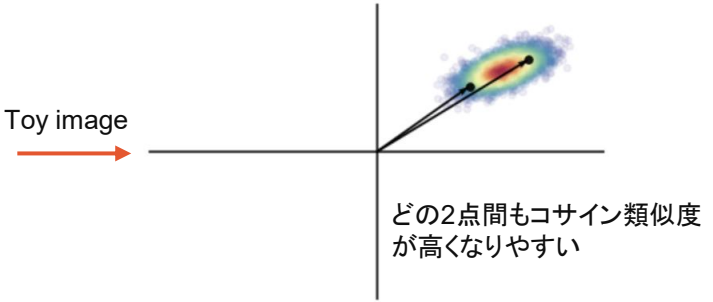
Text	v	a	d
IEMOCAP			
'Excuse me.'	2.5	2.5	2.5
'That's out of control.'	2.5	3.5	3.5
'Did you get the mail? So you saw my letter?'	2.5	2.0	1.5
'Did you get the letter?'	3.5	3.0	2.0
EmoBank			
'Remember what she said in my last letter?'	3	3	3.2
'If I wasn't working here.'	2.8	3.1	2.8
'Goodwill helps people get off of public assistance.'	3.44	3	3.22
'Sherry learned through our Future Works class thatshe could rise out of the mire of the welfare system and support her family.'	3.55	3.27	3.46

Deep Learning-based Categorical and Dimensional Emotion Recognition for Written and Spoken Text, (Atmaja et al., 2019)

言語モデルの埋め込み空間は異方的

- そもそも言語モデルの埋め込み空間は非常に歪んだ形(異方的)をしており、円環とは程遠い





異方的な空間になる主な要因として以下の要素が挙げられる

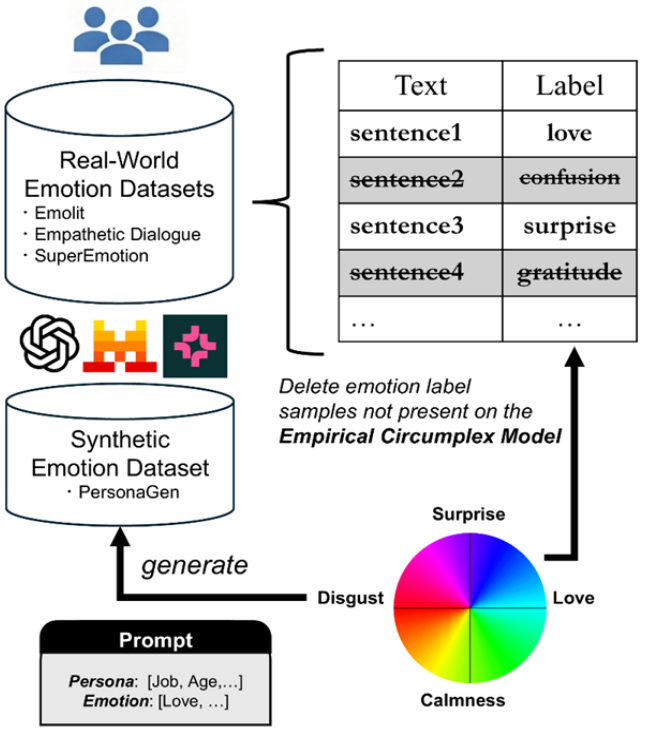
- 言語の頻度分布自体に大きな偏りがある
  - Zipfの法則で知られるように言語の頻度分布には大きな偏りがあるため、それを表現する空間も偏るのは当然？
- ノルムに対する制約がない
  - 概念間の距離を、角度差ではなくノルム差で表現する場合がある

⇒ これらの要因を是正するアプローチでモデルの埋め込みを学習する

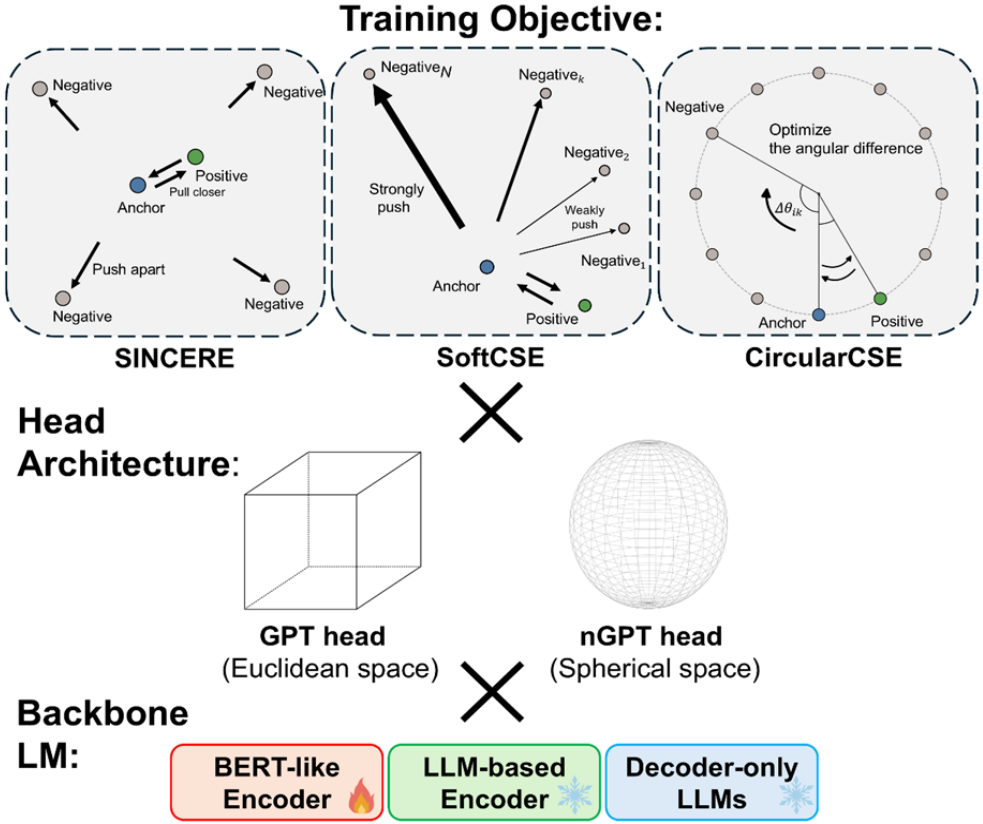
Ethayarajh, “How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings”, EMNLP (2019)
https://aclanthology.org/D19-1006.pdf

提案手法の工夫点：学習データの均一化と幾何制約を意識した訓練

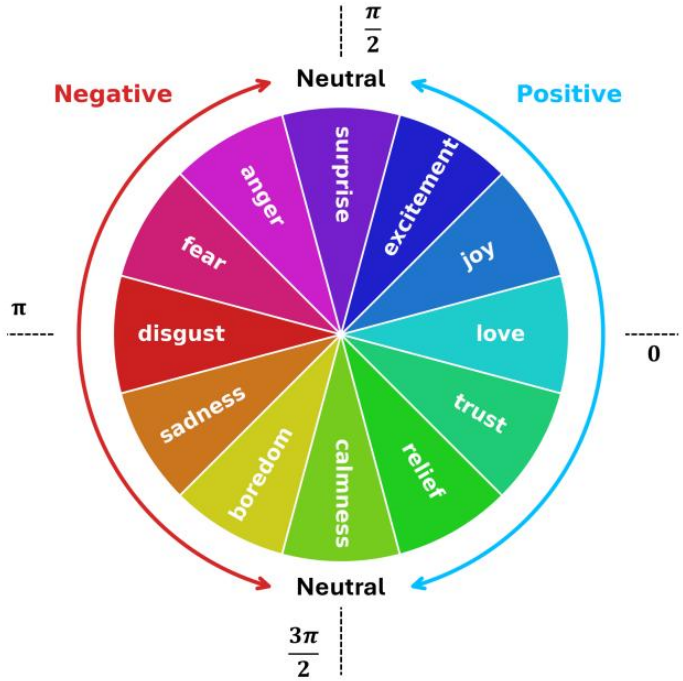
① 学習データの均一化

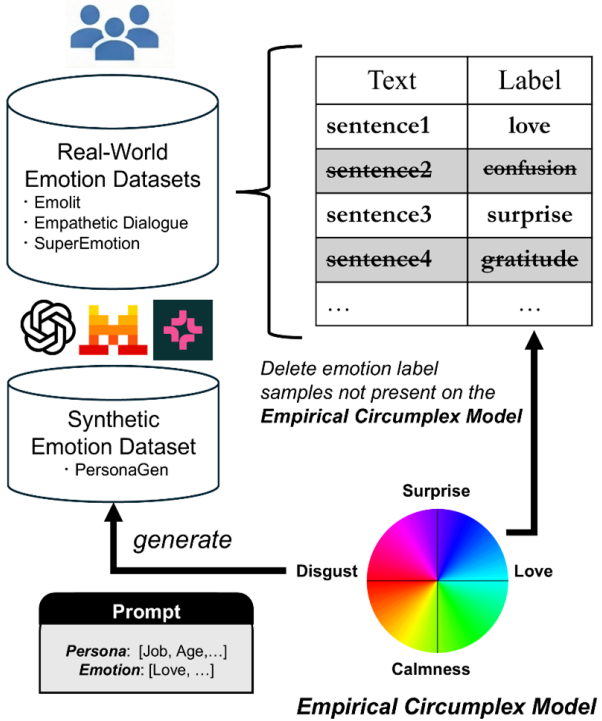


② 幾何制約が異なる損失関数とアーキテクチャ



感情ラベルの配置と学習データの合成





Empirical Circumplex Model (ECM) という独自の円環モデルを定義

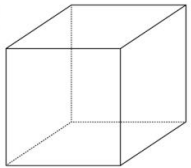
- メジャーかつ位置的に妥当な感情を選定

- ECMに則りデータをフィルタリング
- 足りないものは合成

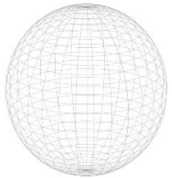
超球面アーキテクチャと対照学習の損失関数



Head Architecture:



GPT head
(Euclidean space)



nGPT head
(Spherical space)

Backbone LM:

BERT-like Encoder | LLM-based Encoder | Decoder-only LLMs

一層のTransformer Blockを事前学習モデルに追加し、その層の出力空間自体が円環になるように学習

- 分類headはつけない
- 各損失関数は負例ペアに対するペナルティの大きさが異なる

Transformer	Normalized Transformer
$\boldsymbol{h}_{\mathrm{A}} \leftarrow \mathrm{ATTN}(\mathrm{RMSNorm}(\boldsymbol{h}))$	$\boldsymbol{h}_{\mathrm{A}} \leftarrow \mathrm{Norm}(\mathrm{ATTN}(\boldsymbol{h}))$
$\boldsymbol{h} \leftarrow \boldsymbol{h} + \boldsymbol{h}_{\mathrm{A}}$	$\boldsymbol{h} \leftarrow \mathrm{Norm}(\boldsymbol{h} + \boldsymbol{\alpha}_{\mathrm{A}}(\boldsymbol{h}_{\mathrm{A}} - \boldsymbol{h}))$
$\boldsymbol{h}_{\mathrm{M}} \leftarrow \mathrm{MLP}(\mathrm{RMSNorm}(\boldsymbol{h}))$	$\boldsymbol{h}_{\mathrm{M}} \leftarrow \mathrm{Norm}(\mathrm{MLP}(\boldsymbol{h}))$
$\boldsymbol{h} \leftarrow \boldsymbol{h} + \boldsymbol{h}_{\mathrm{M}}$	$\boldsymbol{h} \leftarrow \mathrm{Norm}(\boldsymbol{h} + \boldsymbol{\alpha}_{\mathrm{M}}(\boldsymbol{h}_{\mathrm{M}} - \boldsymbol{h}))$
Final: $\boldsymbol{h} \leftarrow \mathrm{RMSNorm}(\boldsymbol{h})$	

nGPT headでは全ての中間表現が超球面上にあり、入力表現は球体の表面上を移動していく

SINCERE × GPTが標準的(無制約)な学習設計
CircularCSE × nGPTが最も円環を誘導しやすい

Loshchilov et al., "**nGPT: Normalized Transformer with Representation Learning on the Hypersphere**", ICLR (2025)
https://openreview.net/forum?id=se4vjm7h4E

実験設定：

- Input:
  - 事前学習済みモデルの文埋め込み {h}
  - Pooling処理済みの768 ~ 4096次元のベクトル
- Output:
  - 追加したTransformer blockの出力 {e}
  - 次元数はbackboneモデルと同じ

- 評価指標:
  - V-Measure : テストセットを教師なしクラスタリングした際のクラスタの識別精度
  - CD-r : ECM上で定義される感情クラス間の距離との一致度

  V-Measureを分類器としての精度のよさ、CD-rを空間の解釈のしやすさとみなす

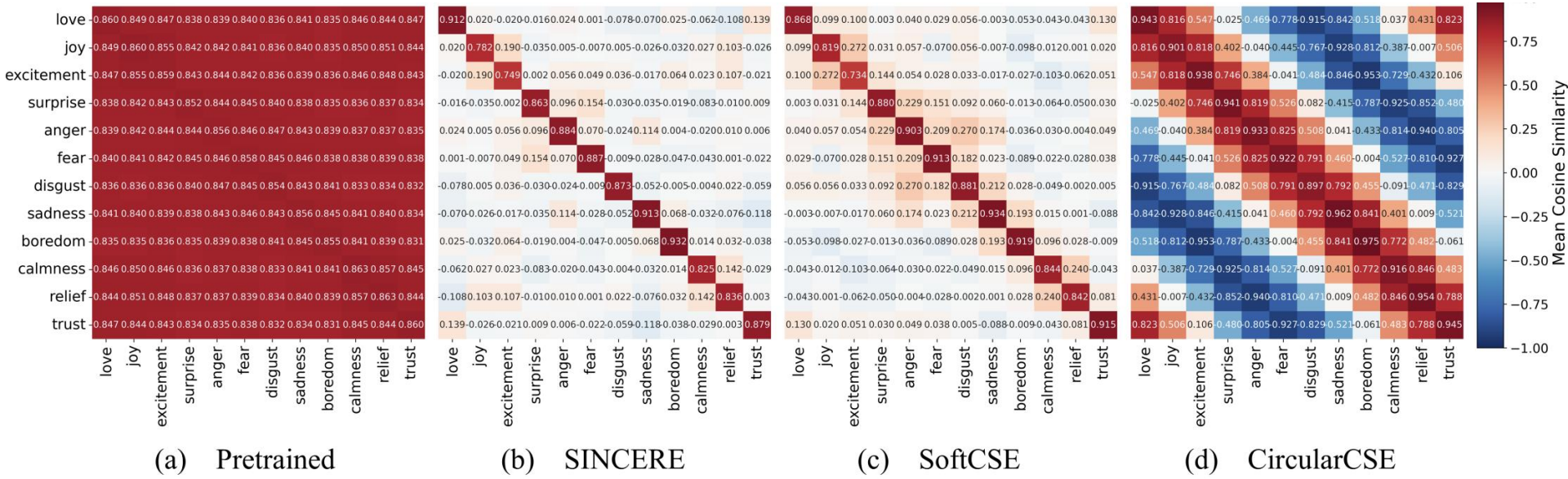
定量的評価：クラスタリング精度と解釈性

Training Objective	Head Arch.	mE5		Qwen3-Embedding-4B		Llama-3.2-3B	
		V_{Measure}	CD-r	V_{Measure}	CD-r	V_{Measure}	CD-r
Pretrained		0.342	0.574	0.495	0.522	0.094	0.217
SINCERE	- GPT	**0.760**	0.317	**0.756**	0.305	**0.725**	0.358
	- nGPT	0.744	0.221	0.739	0.545	0.577	0.425
SoftCSE	- GPT	0.755	0.477	0.751	0.552	0.710	0.548
	- nGPT	0.753	0.499	0.723	0.708	0.516	**0.728**
CircularCSE	- GPT	0.717	0.757	0.643	0.747	0.579	**0.728**
	- nGPT	0.720	**0.764**	0.659	**0.753**	0.382	0.708

- 識別精度においては従来手法 (SINCERE × GPT) が最良
  解釈性においては提案手法 (CircularCSE × nGPT) が最良
- よく見るとV-MeasureとCD-r自体に若干の負の相関の傾向がある
  なぜ？🤔

解釈性を取り入れることのトレードオフ

- 各感情クラスペア間の平均コサイン類似度のヒートマップ



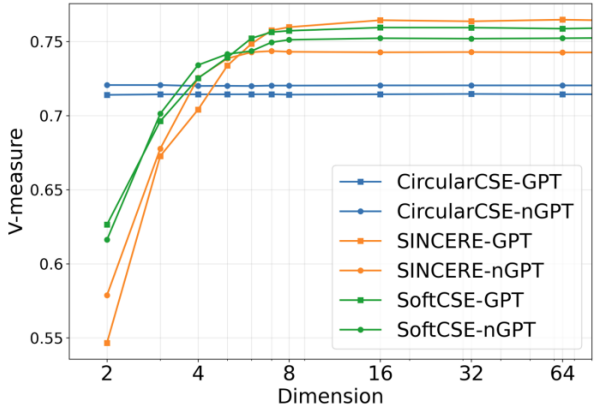
(a) Pretrained (b) SINCERE (c) SoftCSE (d) CircularCSE

- SINCEREは負例ペアをおおよそ直交上に配置する
  CircularCSEは違いの大きさに応じて負例ペアのコサイン類似度に段階がある

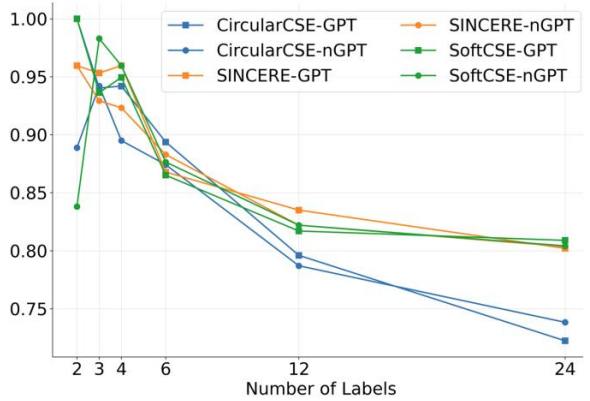
- **解釈性を持たせようとする (≈ 負例ペア間にも差分をつける) とすると**
  どうしても隣接する負例ペア間の距離が近くなり、混同が起きやすくなる

異なる感情ラベル数、次元数での比較

- 次元削減後のクラスタリング精度



- 学習する感情クラス数を変えた際のクラスタリング精度



- 空間の次元数が少ない or 学習するクラス数が少ない場合は提案手法が同等以上の精度を示すが、次元数が増えて感情クラス数も増加すると従来手法に劣るようになる

⇒ **複雑なタスク条件(高次元・多クラス)下では解釈性をとりいれることが、かえって悪影響を及ぼす**ことがあることを示唆している

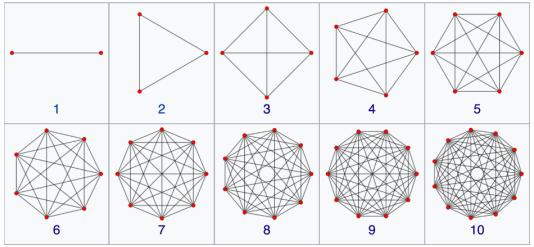
補足：損失関数の下界に関する観点

実は今回適用した損失関数の下界を達成する条件から、各手法がどのような幾何的形状をとるかが推測できる

$$\mathcal{L}_{\text{CircularCSE}} = \frac{1}{B(B-1)} \sum_{i,j \in \mathcal{B}: i \neq j} \ell_{ij},$$

$$\ell_{ij} = \begin{cases} [\max(0, |e_i^T e_j - \cos(\Delta\theta_{ij})| - m)]^2 & \text{if } y_i = y_j \\ (e_i^T e_j - \cos(\Delta\theta_{ij}))^2 & \text{otherwise} \end{cases}$$

最適解の形状



次元が増えても常に2次元の円環が最適解
最も隣接する負例ペア間のコサイン類似度は$cos\left(\frac{2\pi}{E}\right)$
(Eは感情クラス数)

$$\mathcal{L}_{\text{SINCERE}} = \mathbb{E}_i \mathbb{E}_{j \sim \mathcal{P}} \left[-\log \frac{\exp(e_i^T e_j / \tau)}{\exp(e_i^T e_j / \tau) + \sum_{k \in \mathcal{N}} \exp(e_i^T e_k / \tau)} \right]$$

正例ペア間の類似度

負例ペア間の類似度

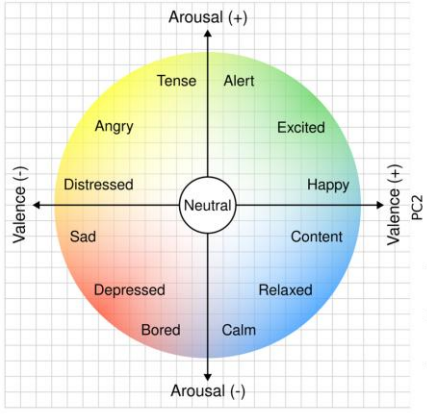
SINCEREでは次元数がE-1以上ある場合は正単体と呼ばれる形状が下界を達成し、負例ペア間のコサイン類似度は

$$e_i^T e_k = -\frac{1}{E-1}$$

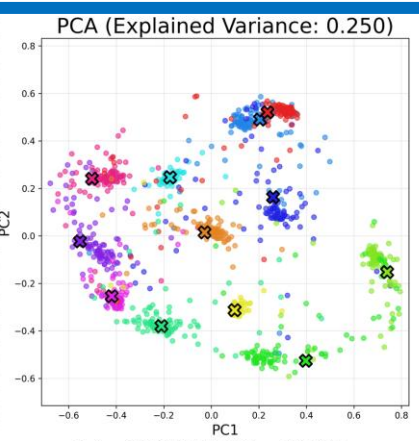
となる

- CircularCSEとSINCEREの最適解の形状が一致するのは次元数が2のときのみ
- 損失関数の下界の観点からも、精度に差が生じることが説明できる

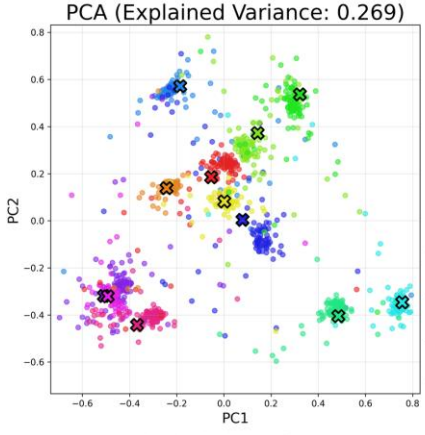
定性的な評価：視覚的に解釈できる埋め込み空間



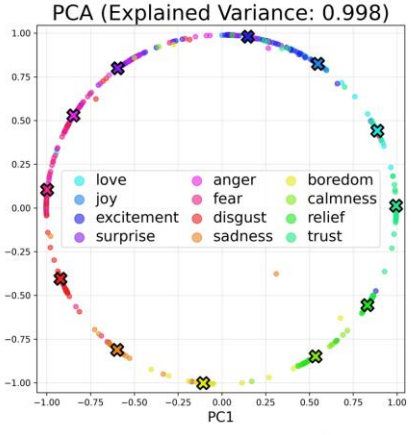
(a) Circumplex model example



(b) SINCERE-nGPT



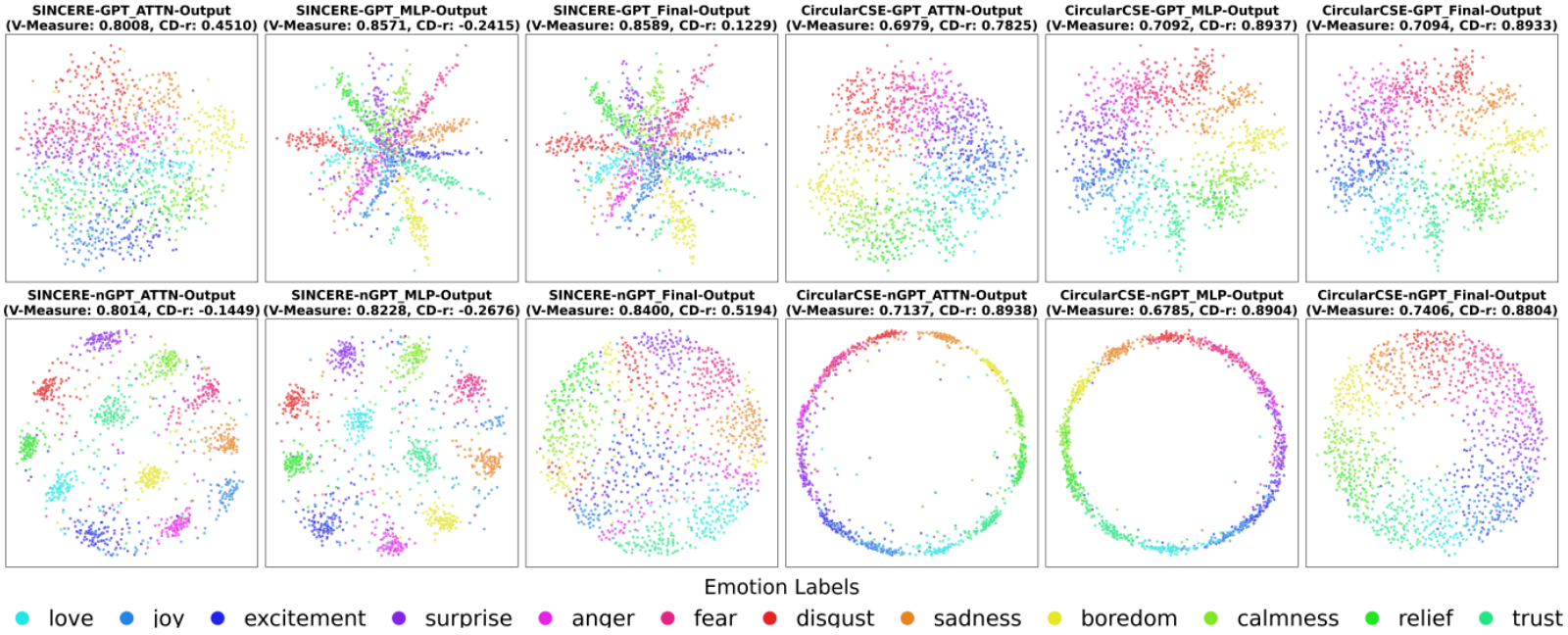
(c) SoftCSE-nGPT



(d) CircularCSE-nGPT

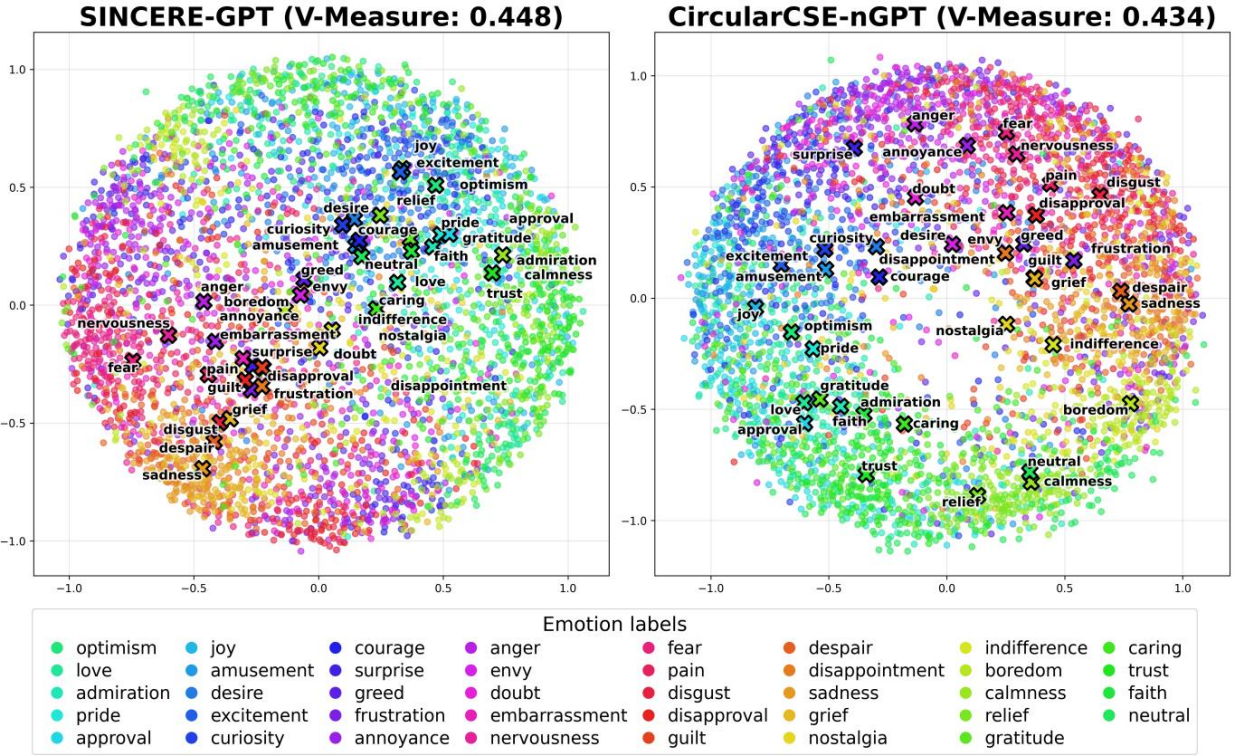
- PCAで2次元に次元削減した際の埋め込み空間
- CircularCSEはほぼ2次元しか使用していないため、次元削減を行い、視覚的に解釈可能!
  エラー分析がしやすい
  ~~賞をとれるぐらい見た目のインパクトが強い~~

定性的な評価：各手法が異なる幾何模様を誘導する



- Transformer blockの中間表現を抽出し、可視化した図
  - GPT headではMLP層の出力が最終的な出力に対して支配的
  - nGPT headではAttn層とMLP層が似たような形状

未知データに対する頑健性



- テストセットに未知感情ラベルを含めた際のクラスタリング精度
  - 差がかなり縮まっている
  - 提案手法はクラス重心が円形を保っている

- 解釈性を取り入れたモデルはデータセットの偏りの推定や未知の感情の検出など、**応用面で有用な可能性**

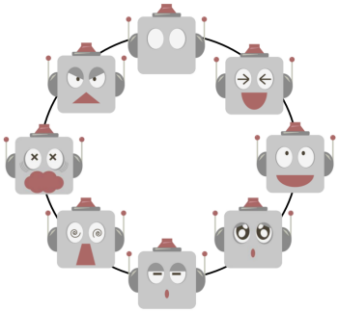
まとめ：心理学で提唱される感情モデルを埋め込み空間に再現

- 学習データの均一化、制約の強さが異なる対照学習損失関数、
  超球面アーキテクチャにより円環構造を誘導
- 解釈性と識別精度の間に構造上避けられないトレードオフを発見
  - “人間にとって分かりやすい”は暗黙的に低次元構造を仮定してしまうことがあり、
    性能劣化につながってしまう
  - 両立はできないので用途に応じて使い分けることが重要

- とにかく見た目にこだわった研究
  - 今は論文数が非常に多いので、ぱっと見で印象に残る要素がないと注目
    を浴びにくいかもしれません

ご清聴ありがとうございました

UTokyo



NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics