IV-P18 埋め込み空間に見る単語分布とモデル地図

山際宏明（京都大学）

- 研究の興味：**埋め込み**（単語や文を表す連続値ベクトル）の分布構造や解釈性を明らかにしたい.
- 紹介内容：埋め込み分布を調べ (①ICA, ②ノルムと分散), ③対数尤度ベクトルでモデル地図を作成した.

研究① 独立成分分析 (ICA) で埋め込みを変換

arXiv 2305.13175 | 2401.06112 | 2406.10984 | 2409.19919

動機 埋め込み空間の形状について，PCA より高い解像度で理解したい.

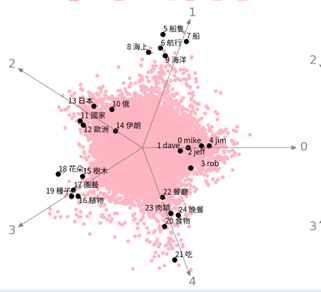
道具 PCA より高次の情報を扱える ICA で埋め込みを変換する.

発見 ICA 変換後の空間は尖っていて，多言語で共通の意味軸が存在する.
➡ その軸を重ねるだけで大体同じ形.

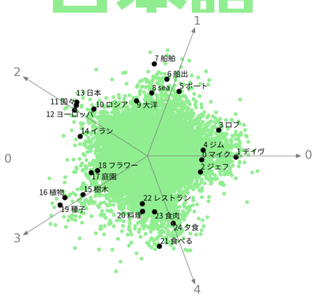
英語



中国語



日本語



事前に計算しておいた n 個の d 次元埋め込みを中心化し $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ とする.

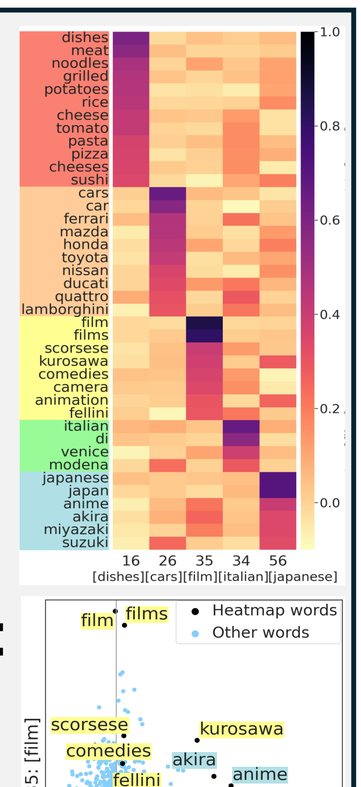
PCA 行列 $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ で白色化 PCA 変換:

$$\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{A} \in \mathbb{R}^{n\times d},$$
$$\boldsymbol{Z}^\top \boldsymbol{Z}/n = \boldsymbol{I}_d.$$

ICA $\boldsymbol{Z}$ に対し，直交行列 $\boldsymbol{R} \in \mathbb{R}^{d\times d}$ を用いて，$\boldsymbol{Z}\boldsymbol{R}$ の各軸をできるだけ独立にしたい．これは 3 次以上の高次モーメントに基づく ICA 変換で実現する:

$$\boldsymbol{S} = \boldsymbol{Z}\boldsymbol{R} \in \mathbb{R}^{n\times d}.$$

- $\boldsymbol{S}$ の軸は成分が大きな単語から解釈可能.
- $\boldsymbol{S}$ はスパースで埋め込み空間は尖った形状.

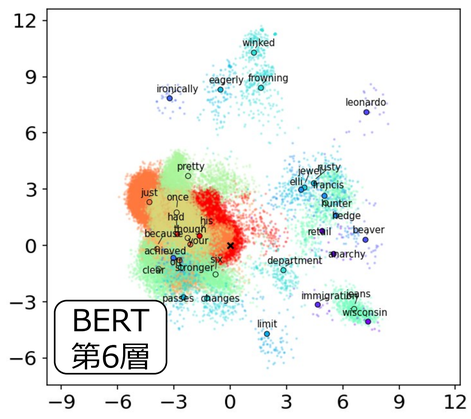


研究② ノルムと分散を用いて文脈化埋め込み集合の形状を調べる

arXiv 2409.11253

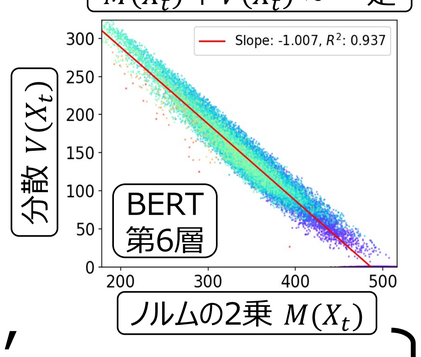
動機

- 文脈ごとにトークンの埋め込みは異なる.
- その埋め込み集合がもつ分布的特徴を調べたい.



道具 埋め込み集合について
- 平均ベクトルのノルム：意味の強さを表す.
- 分散：意味の広がりを表す.

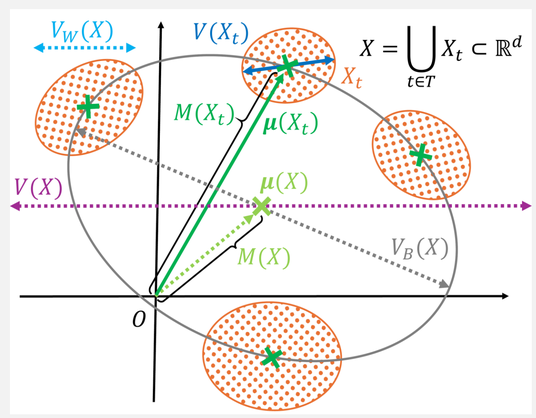
発見1 中間層において，ノルムと分散はトレードオフ.



発見2 層が深くなるにつれ，
- 埋め込み全体は原点から離れ，密集する.
- 埋め込み集合は広がっていく.
- 異なる集合同士は密集していく.

トークン t の埋め込み集合 $X_t \subset \mathbb{R}^d$ の統計量
- X_t の平均ベクトル：$\boldsymbol{\mu}(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\} \in \mathbb{R}^d$
- 平均 $\boldsymbol{\mu}(X_t)$ のノルムの 2 乗：$M(X_t) = \|\boldsymbol{\mu}(X_t)\|^2$
- X_t の分散：$V(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\|\boldsymbol{x} - \boldsymbol{\mu}(X_t)\|^2\}$

埋め込み全体 X の統計量
- $\boldsymbol{\mu}(X), M(X), V(X)$ は同様.
- 群内分散：$V_W(X)$
- 群間分散：$V_B(X)$
- 特に $V(X) = V_W(X) + V_B(X)$.

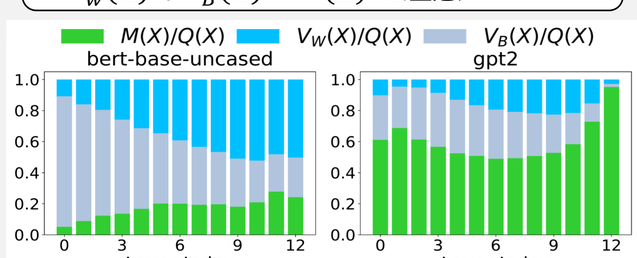


統計量の計算 効率的な逐次計算が可能.

$M(X), V(X)$ の相対比較
$M(X)$ 増加，$V(X)$ 減少.

$V_W(X), V_B(X)$ の相対比較
$V_W(X)$ 増加，$V_B(X)$ 減少.

- $Q(X) = M(X) + V_w(X) + V_B(X)$ で規格化
- $V_w(X) + V_B(X) = V(X)$ に注意



研究③ 対数尤度ベクトルを用いた言語モデルの地図

arXiv 2502.16173 | 2505.15353 | 2505.15428

動機 言語モデルを比較したい. (例) Qwen2-7B と gemma-2b の”距離”は？

道具 対数尤度ベクトルを使用．モデル p_i に対し，テキスト $x_1, \ldots, x_N$ を用いて

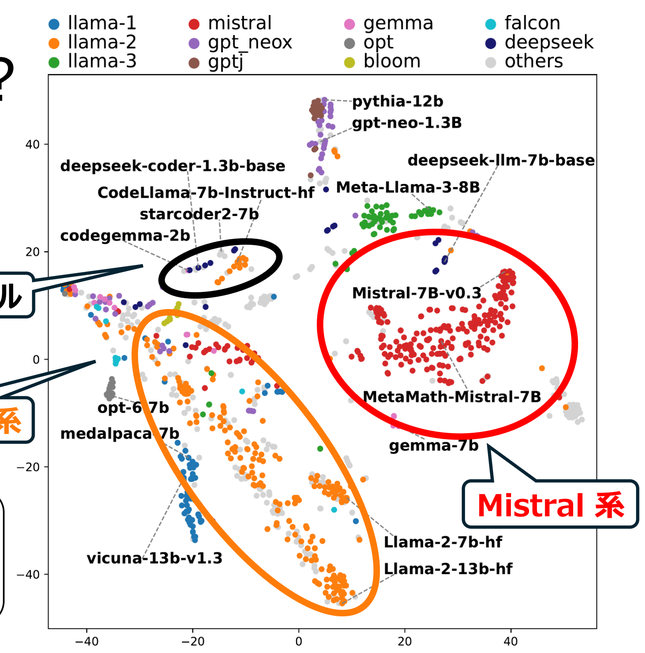
$$\boldsymbol{\ell}_i = (\log p_i(x_1), \ldots, \log p_i(x_N)) \in \mathbb{R}^N.$$

➡ 異なる言語モデルでも N 次元ベクトルとして比較可能！

性質 中心化した対数尤度ベクトルを $\boldsymbol{q}_i$ とすると

$$2\,\mathrm{KL}(p_i, p_j) \approx \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2/N.$$

➡ $\boldsymbol{q}_i, \boldsymbol{q}_j$ 間のユークリッド距離の 2 乗は近似的に KL に比例.



1,018 モデルについて，
10,000 テキストで $\boldsymbol{q}_i$ 計算.
t-SNE でモデル地図を作成.