Norm of Mean Contextualized Embeddings Determines their Variance

arXiv

Hiroaki Yamagiwa[1] and Hidetoshi Shimodaira[1,2]
[1]Kyoto University [2]RIKEN AIP

The 31st International Conference on Computational Linguistics
COLING 2025 • Abu Dhabi

Main Contributions

- We propose statistical measures to analyze the **distribution of contextualized embeddings** and provide an **efficient sequential computation** method.
- We show that the norm of mean embeddings and their variance exhibit **a strong trade-off**.
- These values explain **embedding anisotropy across layers** more precisely than prior work.

Token-wise embedding set X_t

Def. For the set of contextualized embeddings $X_t \subset \mathbb{R}^d$ for token t, we define the mean embedding $\boldsymbol{\mu}(X_t) \coloneqq \mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\}$ and three values:

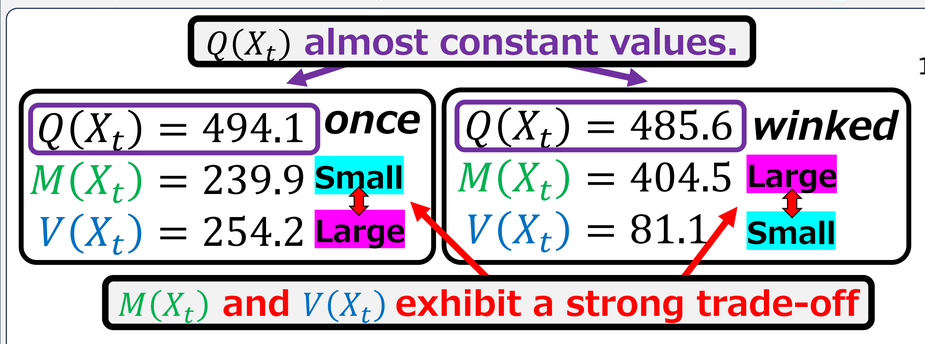
$$Q(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\|\boldsymbol{x}\|^2\},$$
$$M(X_t) = \left\|\mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\}\right\|^2 = \|\boldsymbol{\mu}(X_t)\|^2,$$
$$V(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\|\boldsymbol{x}-\boldsymbol{\mu}(X_t)\|^2\} = \sum_{i=1}^{d}\left\{\left(x_i-\mu_i(X_t)\right)^2\right\}.$$

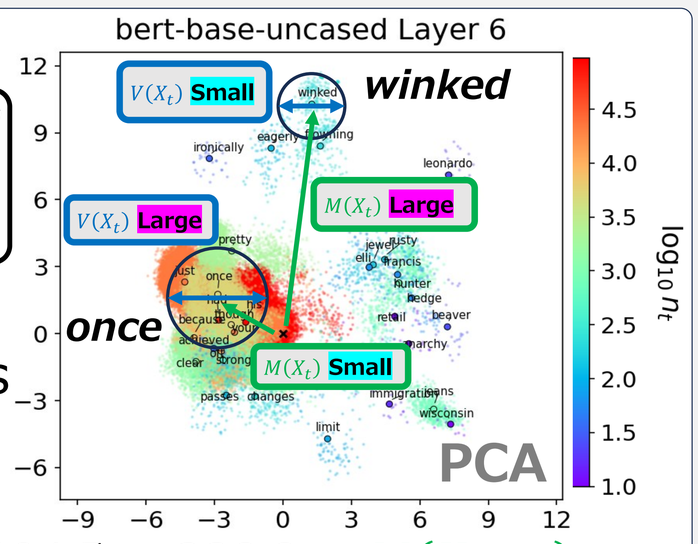
These satisfy $Q(X_t) = M(X_t) + V(X_t)$.

- $M(X_t)$: the **squared norm of the mean embedding** $\boldsymbol{\mu}(X_t)$.
- $V(X_t)$: the **spread of the distribution** based on **variance**.

Computation $\boldsymbol{\mu}(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\boldsymbol{x}\}$ and $Q(X_t) = \mathbb{E}_{\boldsymbol{x}\in X_t}\{\|\boldsymbol{x}\|^2\}$ are computed by **the sequential method**. $M(X_t) = \|\boldsymbol{\mu}(X_t)\|^2$ and $V(X_t) = Q(X_t) - M(X_t)$ can also be **efficiently** computed.

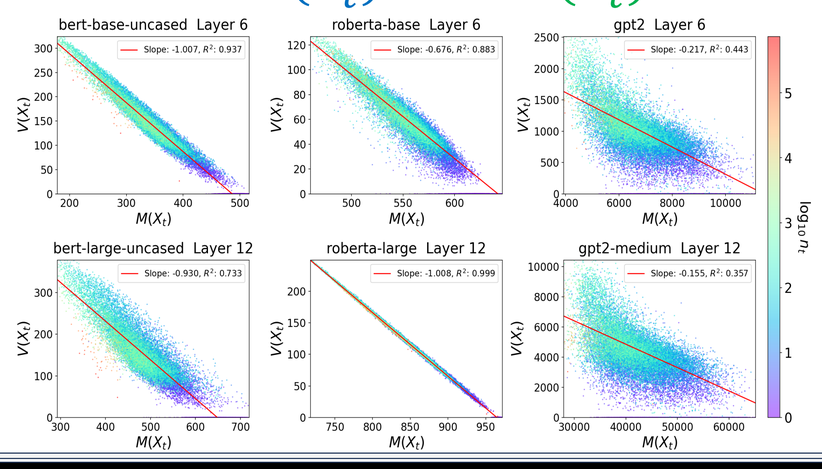
Findings In intermediate layers, **the variation of $Q(X_t)$ is small**; **$M(X_t)$ and $V(X_t)$ exhibit a strong trade-off**.



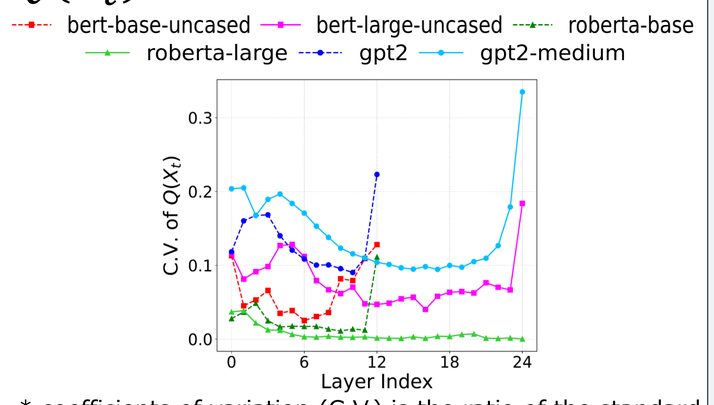


The norm of an embedding reflects meaning strength [Oyama+2023].
➡ $M(X_t)$ represents **the meaning strength** of token t: $M(X_{\text{winked}}) = 404.5 > 239.9 = M(X_{\text{once}})$.

Scatterplots of $V(X_t)$ on $M(X_t)$ for the six models confirm **trade-off between $M(X_t)$ and $V(X_t)$**.



In the intermediate layers of the six models, **C.V.s of $Q(X_t)$ are small**.



* coefficients of variation (C.V.) is the ratio of the standard deviation to the mean, the relative variability in the data.

The entire embedding set X

Def. For $X = \mathrm{U}_{t\in T} X_t \subset \mathbb{R}^d$, $\boldsymbol{\mu}(X)$, $Q(X)$, $M(X)$, and $V(X)$ are also defined.
Additionally, **the within-group variance**

$$V_W(X) = \textstyle\sum_{t\in T} p_t V(X_t)$$

and **the between-group variance**

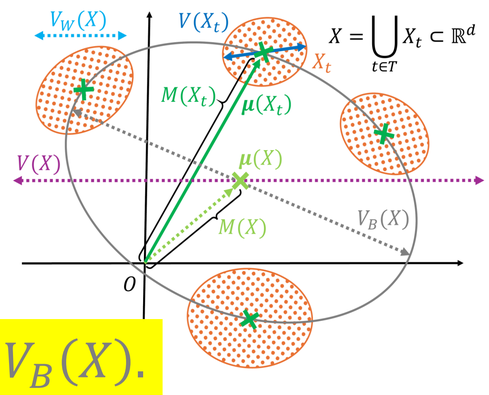
$$V_B(X) = \textstyle\sum_{t\in T} p_t \|\boldsymbol{\mu}(X_t) - \boldsymbol{\mu}(X)\|^2$$

satisfy

$$V(X) = V_W(X) + V_B(X),$$

where $p_t = |X_t|/|X|$.
Then

$$Q(X) = M(X) + V(X) = M(X) + V_W(X) + V_B(X).$$

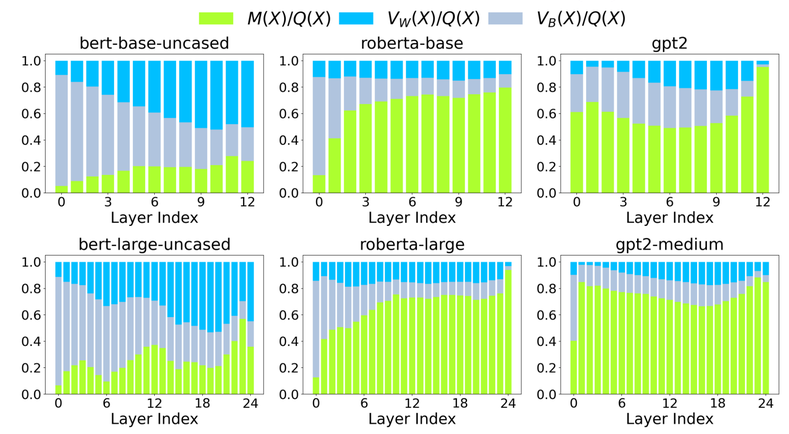


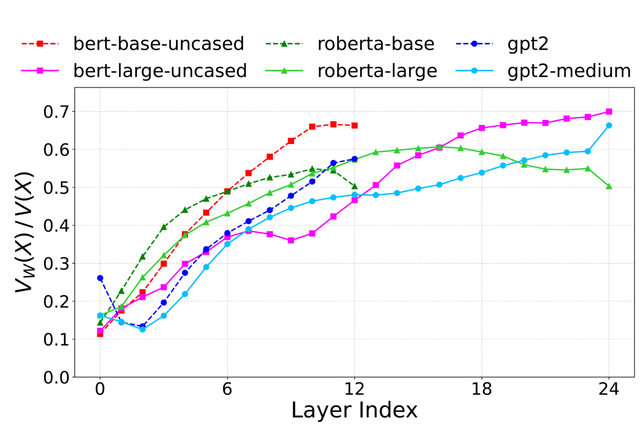
Computation $\boldsymbol{\mu}(X)$ and $Q(X)$ are the frequency-weighted means of $\boldsymbol{\mu}(X_t)$ and $Q(X_t)$:

$$\boldsymbol{\mu}(X) = \sum_{t\in T} p_t \boldsymbol{\mu}(X_t), Q(X) = \sum_{t\in T} p_t Q(X_t).$$

Other values are also **efficiently** computed.

Experiments The left plot shows $\frac{M(X)}{Q(X)}$, $\frac{V_W(X)}{Q(X)}$, and $\frac{V_B(X)}{Q(X)}$, while the right plot shows $\frac{V_W(X)}{V(X)}$, for each layer across models.





As the layers deepen:

- while $M(X)$ increases, $V_W(X)+V_B(X)$ and $V_B(X)$ decrease.
  ➡ Embeddings move farther from the origin, and different X_t become more densely clustered.
- $V_W(X)$ increases.
  ➡ The embeddings in X_t become more contextualized.

This **explains embedding anisotropy across layers more precisely than prior work** that used cosine.

Discussion

Why does the C.V. of $Q(X_t)$ reach its minimum in the intermediate layers?
➡ Input and output layers are distinct and may have different properties from intermediate layers.
Why does GPT-2 behave differently from BERT and RoBERTa for experiments of X_t across layers?
➡ One possible reason is the LayerNorm placement: after FF in BERT/RoBERTa, before FF in GPT-2.