Accelerating ScaNN via pruning-based vector quantization

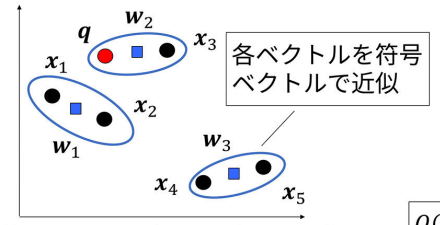
04 枝刈りによるベクトル量子化手法ScaNNの高速化

大量のベクトルから高速に似たものを見つけます

どんな研究	ScaNNはベクトルを符号語に置き換えて量子化し、内積による類似検索を高速・高精度で行う、機械学習における重要な技術です。一方でその置き換え処理自体に時間がかかるという課題がありました。本展示ではScaNNの精度を犠牲にせずにベクトル量子化を高速に行う技術を紹介します。
どこが凄い	従来のScaNNはすべてのベクトルと符号語との誤差を計算して最適な符号語に置き換える必要がありました。本技術は量子化誤差の上限に着目し、それを高速に計算することで最適になりえない候補を枝刈りします。その結果、ベクトル量子化の大幅な高速化を実現しました。
めざす未来	本技術は計算結果の精度を落とさずにベクトル量子化を高速化できます。内積計算は画像検索、深層学習、自然言語処理など幅広い分野で不可欠です。本技術はベクトル量子化を大規模データに対しても適用可能にし、情報処理基盤を支える中核技術となることが期待されます。

背景：最大内積検索

ベクトル$\boldsymbol{q}$との内積が最大になるベクトル$\boldsymbol{x}$を検索
⭕画像検索、深層学習、自然言語処理などに適用可能
❌データベース内の全てのベクトルに対して内積計算が必要なため、計算コストが高くなる
ベクトル量子化: ベクトル$\boldsymbol{x}$を符号ベクトル$\boldsymbol{w}$に近似し検索
⭕高速な検索が可能だが、❌検索精度が低下



符号ベクトル$\boldsymbol{w}$はデータの分布により設定

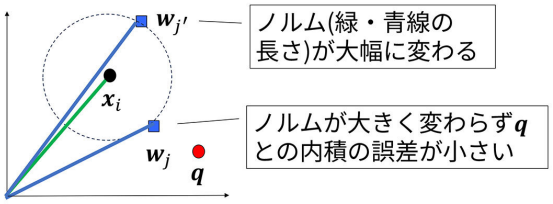
事前に計算した符号ベクトル間の内積

内積	w_1	w_2	w_3
w_1	⋯	⋯	⋯
w_2	⋯	⋯	⋯
w_3	⋯	⋯	⋯

$O(1)$で内積を求め検索

既存のベクトル量子化手法：ScaNN

アイデア: ベクトルの大きさ(ノルム)が変わらないように近似し、内積の誤差を小さくすることで検索精度を向上



手法: 誤差関数$l(\boldsymbol{x}_i, \boldsymbol{w}_j)$が最小の符号ベクトルに割当
⭕誤差関数が小さい ⇔ ノルムが大きく変わらない
❌検索精度は向上するが、誤差関数の計算コストが高い

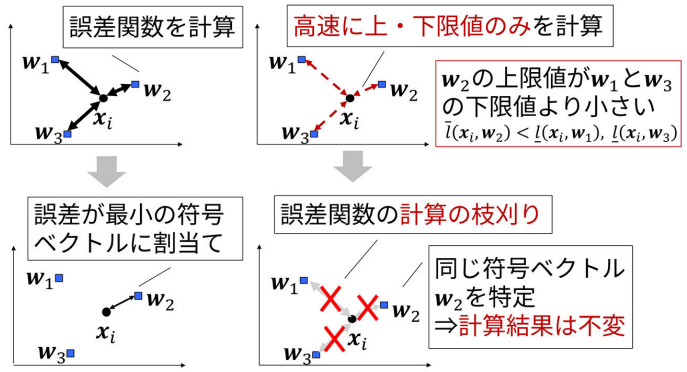
$$l(\boldsymbol{x}_i, \boldsymbol{w}_j) = \frac{h_i - 1}{\|\boldsymbol{x}_i\|^2}\left(p(\boldsymbol{x}_i, \boldsymbol{w}_j)\right)^2 - 2h_i p(\boldsymbol{x}_i, \boldsymbol{w}_j) + h_i\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{w}_j\|^2$$

h_i : 1より大きな値，$p(\boldsymbol{x}_i, \boldsymbol{w}_j)$：ベクトル$\boldsymbol{x}_i$と$\boldsymbol{w}_j$の内積

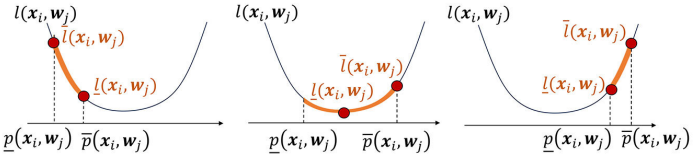
提案手法：枝刈りによる高速化

高速に計算できる誤差関数の上限値$\bar{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$と下限値$\underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$を用いて誤差関数の計算を枝刈り

・既存手法　・提案手法

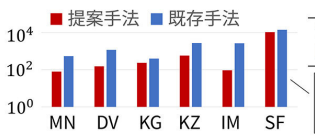


誤差関数$l(\boldsymbol{x}_i, \boldsymbol{w}_j)$は内積$p(\boldsymbol{x}_i, \boldsymbol{w}_j)$の2次関数
⇒ $\bar{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$と$\underline{l}(\boldsymbol{x}_i, \boldsymbol{w}_j)$は、内積の上限値$\bar{p}(\boldsymbol{x}_i, \boldsymbol{w}_j)$と下限値$\underline{p}(\boldsymbol{x}_i, \boldsymbol{w}_j)$から計算



既存手法の検索精度を変えずに大幅な高速化を実現

・処理時間 [秒]



・検索精度[%]

	MN	EM	KZ	IM	TN	SF
提案手法	56.5	35.5	59.7	53.5	6.28	69.8
既存手法	56.5	35.5	59.7	53.5	6.28	69.8

検索精度は変わらない

・実験データ(ベクトルの数×ベクトルの長さ)
MN：MNIST (70,000 × 780)
DV：Devnagari-Script (92,000 ×1,024)
KG：kaggle (200,000 × 202)
KZ：Kuzushiji (270,912 × 784)
IM：imagenet (2,340,373 × 150)
SF：sift10m (10,000,000 × 128)

関連文献

[1] Y. Fujiwara, Á. López, Y. Ida, A. Kumagai, M. Nakano, M. Nakatsuji, A. Kimura, “Fast Vector Quantization Algorithm for ScaNN”, in *Proc KDD*, 2026.

連絡先

藤原 靖宏（Yasuhiro Fujiwara）メディア情報研究部 メディア認識研究グループ