Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video の要約

arxiv : [[2510.03458] Omni-Embed-Nemotron: A Unified Multimodal Retrieval Model for Text, Image, Audio, and Video](https://arxiv.org/abs/2510.03458)

hugging face: [nvidia/omni-embed-nemotron-3b · Hugging Face](https://huggingface.co/nvidia/omni-embed-nemotron-3b)

ライセンスは、nvidia-open-model-licenseで商用利用不可。

＜概要＞

NVIDIAがマルチモーダル検索埋め込みモデルであるOmni-Embed-Nemotronを提案。

Omni-Embed-Nemotron は、Qwen-Omni / Qwen2.5-Omni-3Bのマルチモーダルモデルの機能から着想を得て、検索をテキストと画像だけでなく、音声と動画のモダリティにも拡張した。

＜Omni-Embed-Nemotronの特徴＞

アーキテクチャ：

クエリとコーパスを独立にエンコードするBi-Encoder構造を採用。類似度計算にはコサイン類似度やドット積を使用。

ダイアグラム

AI 生成コンテンツは誤りを含む可能性があります。

モダリティ対応：

テキスト、画像、音声、動画の単一モーダルおよび複合モーダル（例：音声＋動画から画像+動画を検索）の検索に対応。

これにより、音声による説明とテキストキーワードを用いて動画データベースを検索したり、マルチモーダルプロンプトを用いて関連する音声・映像コンテンツを検索したりするなど、豊富で柔軟な検索ユースケースが可能となる。

エンコーディング戦略：

Qwen-OmniのThinkerバックボーンを使用している。

Qwen-Omniとの違いは、動画に対する処理の方法にある。

Qwen-Omniは音声と画像を交互に処理している(図3(a))が、Omni-Embed-Nemotronでは、音声と動画を非インターリーブ方式で別々にエンコードする(図3(b))。これにより、各モダリティの時間構造を保持することで、検索パフォーマンスが向上。

グラフィカル ユーザー インターフェイス

AI 生成コンテンツは誤りを含む可能性があります。

＜学習＞

音声と映像の両方のエンコーダをフリーズし、LLMにのみLoRAファインチューニングを実施。

学習データセットには、テキストと画像のペアやテキスト同士のペアの検索データセットを使用。

（学習データセットに音声や動画データを使用しなくても、本モデルはこれらのモダリティで優れたパフォーマンスを発揮する。）

また、テキストと動画、およびテキストと音声のペアのデータセットを追加して学習することで、精度の向上が見られる。

＜実験結果＞

動画検索：

以下2つのデータセットに対して、テキストによる質問(クエリ)に対応する動画の検索の精度をNDCG@10/ NDCG@5で評価（1に近いほど精度が良い）。

FineVideo：平均再生時間4.7分の43,751本の動画が含まれており、122の多様なカテゴリが含まれる。

LPM：スライドと講演者の頭が移っている講義動画。

結果：FineVideoベンチマークで従来のテキストベースモデルを上回る性能（NDCG@10で2.6ポイント向上）を示した。

グラフィカル ユーザー インターフェイス, アプリケーション, テーブル

AI 生成コンテンツは誤りを含む可能性があります。

画像検索：

　以下のデータセットに対して、テキストによる質問(クエリ)に対応する画像の検索の精度をnDCG@5で評価(これは100に近いほど精度が良い)。

　　ViDoRe：多様な画像と自然言語クエリの組み合わせで構成されている。

　結果：ViDoReベンチマークで85.7の平均スコアを達成した。特に特定領域（AI、政府報告書、エネルギー）で高精度を示した。

テーブル

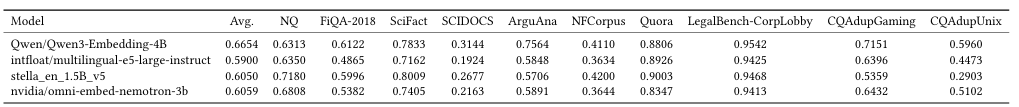
AI 生成コンテンツは誤りを含む可能性があります。

テキスト検索：

以下のデータセットに対して、テキストによる質問(クエリ)に対応するテキストの検索の精度をnDCG@10で評価（1に近いほど精度が良い）。

　MTEB：オープンドメインの質問応答、科学文献検索、議論型テキストマッチングなど、様々な領域をカバーしている。

結果：MTEBベンチマークで平均スコア0.6059を達成し、テキスト特化のモデルに匹敵する性能を示した。



＜結論＞

Omni-Embed-Nemotronは、マルチモーダル検索の新たな基準を打ち立てるモデルであり、テキスト・画像・音声・動画を統一的に扱える点が最大の強みである。特定モダリティに特化したモデルには及ばない場面もあるものの、汎用性と拡張性に優れた検索性能を示している。今後は、ドメイン特化型ファインチューニングやより洗練されたマルチモーダルアライメント戦略によってさらなる性能向上が期待される。

＜感想＞

実験については、今年の5月に公開されたOmni-Embed-Nemotronと同じマルチモーダル埋め込みモデルであるOmniEmbedとの比較は欲しいと思った。

[[2505.02466] Tevatron 2.0: Unified Document Retrieval Toolkit across Scale, Language, and Modality](https://arxiv.org/abs/2505.02466)

[Tevatron/OmniEmbed-v0.1 · Hugging Face](https://huggingface.co/Tevatron/OmniEmbed-v0.1) (こっちはMITライセンス)