DeepSeek V3: マルチヘッド潜在注意とマルチトークン予測

DeepSeek v3 はここで探索されています。ビデオ「Multi-Head Latent Attention and Multi-token Prediction in Deepseek v3」https://youtu.be/jL49fLOJYNg?si=4uE2kfe-BlKC1ngOを参照しています。Google Cloud Speech-to-Text を使用してビデオを音声認識し、トランスクリプトを整理するためのコードをいくつか作成しました。

A: Deep タグへようこそ。今日は、大規模言語モデルの世界に深く掘り下げていきます。具体的には、DeepSeek V3 についてです。

B: いいね。6710 億パラメータのモデルで、効率とパフォーマンスの独自のアプローチで話題になっているよね？

A: そして、そのアーキテクチャを詳細に説明する学術論文を共有してくれたね。

B: そうだよ。

A: そして、機械学習の専門家として、DeepSeek V3 が高いパフォーマンスと経済的なトレーニングをどのように実現しているのかを理解しようとしている。

B: そうだよ。

A: こんにちは、どうですか？

C: MLA、その詳細、MLA とその仕組みについて。

A: もちろん。それは素晴らしいアイデアだ。MLA の細部に深く掘り下げることができる。MLA の仕組みについて興味があるんだね。では、これを解き明かそう。DeepSeek V3 の効率の鍵は、Mixture of Experts（MoE）アーキテクチャにあると説明したよね？各トークンごとにパラメータの一部のみがアクティブになる。

B: そうだよ。では、今すぐ MLA に集中しよう。

A: いいね。リアルタイムアプリケーションでは、速度が重要だ。

B: そうだよ。推論中に必要なキー値キャッシュが大きなボトルネックになることがある。

A: その通り。そこで MLA が登場する。伝統的なアテンションメカニズムは、以前のトークンに関する多くの情報を保存する必要がある。

B: そうだよ、長いテキストのシーケンスでは問題になるよね？

A: だけど、MLA はこの情報を巧妙に圧縮して、キャッシュフローを大幅に減らし、推論を高速化する。まるで、大型の百科事典を要点だけにまとめるようなものだ。

B: 素晴らしいアナロジーだ。本質的な情報を保持しつつ、不要な重みを削減する。

A: その通り。では、実際にどのように動作するのかを話そう。MLA がこの圧縮をどのように実現するのか。

B: 低ランクの共通圧縮をアテンションのキーと値に使用する。

A: つまり、キーと値を圧縮するんだけど、具体的にはどういうことか。少し技術的な話になるけど、MLA メカニズムは入力の隠れ表現を受け取り、クエリ、キー、値ベクトルにプロジェクトする。

B: 2 つの部分に分けるんだね？

A: そうだよ。1 つはコンテンツ用、もう 1 つは位置情報用に Rope（回転位置埋め込み）を使用する。

B: Rope？とても技術的な言葉だね。

A: それは、モデルがシーケンス内のトークンの位置を理解するのを助ける。

B: そして、キーと値は低次元の潜在空間に圧縮される。つまり、データを縮小してメモリを節約する。

B: その通り。重要な情報は保存され、不要な重みは捨てられる。そして、この圧縮された表現は、推論中の小さな KV キャッシュを可能にし、速度を向上させる。

A: そして、マルチヘッド処理も使用する。

B: そうだよ、伝統的なアテンションと同じように、MLA も複数のヘッドを使用する。

A: では、続けよう。

C: つまり、2 つの潜在空間と 1 つの隠れ入力がある。

A: 素晴らしい観察だ。その通り。実際には 2 つの潜在空間がある。コンテンツの潜在空間とキー値の潜在空間だ。

B: その通り。そして、これらの潜在空間は Rope（回転位置埋め込み）を通じて処理される。

A: つまり、Rope が位置情報を取得する方法だ。

B: そうだよ、コンテンツとキー値の潜在空間の両方に適用される。圧縮された表現を処理し、すべてを再び結合する。

A: そして、キャッシュの最適化はシーケンシャル処理中のオーバーヘッドをさらに削減する。つまり、これが MLA が速度を向上させる方法だ。

B: その通り。効率的なアテンションを実現しつつ、パフォーマンスを犠牲にしない巧妙な方法だ。

A: これはかなり素晴らしいトリックだ。でも、知っているか？

B: なんだ？

A: では、DeepSeek Mo に移ろう。これは伝統的な MoE モデルとどのように異なるのか。

B: もちろん、DeepSeek Mo は…もう一度、リスナーに戻ろう。どうですか？

C: そして、さらに隠れ空間について話す。隠れ空間についてどう思う？

A: もちろん…何が言いたいのかを確認しよう。隠れ空間は非常に興味深い。隠れ空間、つまり私たちが話していた潜在空間について聞いているんだね。その洞窟の中での出来事について興味があるんだね。

B: 素晴らしい。

A: その通り。MLA 内には、コンテンツとキー値の 2 つの異なる潜在空間がある。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、

アテンションメカニズムに位置情報を埋め込む。これは非常に重要だ。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: その通り。そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して処理される。

B: その通り。これにより、モデルは情報が洞窟の中を通過する際に位置情報を保持する。モデルはテキストのどの部分がどの部分であるかを知っている。

A: そして、この処理は次の結合ステージの前に行われる。洞窟を通過する際に結合されているのは何か？

B: メカニズムは 2 つの主要な結合操作を行う。クエリ表現とキー表現が結合される。つまり、その隠れ空間の洞窟の中で重要な部分をすべて結合する。

A: そして、これらの結合はコンテンツと位置情報を組み合わせる。そして、これらの結合された表現はアテンション計算に使用される。

B: その通り。そして、初期の圧縮により、その洞窟の中と外で計算コストが大幅に削減される。MLA は大規模モデルのような DeepSeek V3 のアテンションメカニズムを最適化する。これは素晴らしい質問だ。では、洞窟を抜け出して DeepSeek Mo に移ろう。

B: もちろん、DeepSeek Mo。その通り、2 つの異なる潜在空間がある。コンテンツとキー値のそれぞれに。

A: その通り。そして、この分離はその仕組みにとって非常に重要だ。情報の 2 つの別々のストレージユニットを持つようなものだ。そして、これらの潜在空間は、Rope 操作を受ける。回転位置埋め込みは、アテンションメカニズムに位置情報を埋め込む。要約すると、クエリは分割され、キーと値も圧縮される。

B: その通り。そして、これらは 2 つの別々の潜在空間に入れられる。コンテンツとキー値ペアのそれぞれに。そして、これらの潜在空間は、MLA の一部として効率性にとって非常に重要だ。

A: その通り。では、これらの操作についてさらに詳しく話そう。潜在空間の変換がどのように行われるのか。

B: 入力は、コンテンツとキー値の表現の両方に対して並列処理される。つまり、その洞窟の中に 2 つのパスがある。

A: その通り。それぞれの潜在空間に 1 つずつ。そして、その空間内で情報は Rope を使用して