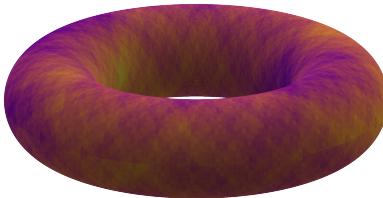


קידוד מיקום בארכיטקטורת Transformer ניתוח מתמטי מקיף

Positional Encoding in Transformer Architectures
A Comprehensive Mathematical Analysis



מתיאוריה לIMPLEMENTATION
From Theory to Implementation

מאת: ד"ר יורם

By: Dr. Yoram

2025

זכויות יוצרים

Copyright © 2025 Dr. Yoram

כל הזכויות שמורות לד"ר יורם. אין לשכפל, להעתיק, לצלם, להקליט, לתרגם, לאחסן
במאגר מידע, לשדר או לקלוט בכל דרך או אמצעי אלקטרוני, אופטי, מכני או אחר - כל
חלק שהוא מהחומר שבספר זה. שימוש מסחרי מכל סוג שהוא בחומר הכלול בספר זה
אסור בהחלט אלא ברשות מפורשת בכתב מאת בעל הזכויות.

All rights reserved to Dr. Yoram. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the copyright holder. Commercial use of any kind is strictly prohibited without explicit written consent from the copy-right owner.

הערת תוכנה:

ספר זה נכתב בשיתוף עם מערכת בינה מלאכותית מתقدמת מבוססת Claude Sonnet 4.5. התוכן המתמטי, הניתוח התיאורתי, והשימוש המעשי נערכו בפיקוח אקדמי מלא ונבדקו לאיות דיווק.

Software Notice: This book was written in collaboration with an advanced AI system based on Claude Sonnet 4.5. The mathematical content, theoretical analysis, and practical implementation were conducted under full academic supervision and verified for accuracy.

מוקדש

**לכל החוקרים והפתחים
העוסקים בהבנת המנגנונים הפנימיים
של מודלים של Transformer**

Dedicated

To all researchers and developers
Working to understand the inner workings
of Transformer models

תקציר

מחקר מكيف זה בוחן את מנגנון קידוד המיקום הסינוסואידלי שהוצע במאמר המכונן - "Attention is All You Need" של Vaswani et al. העובדה מספקת טיפול מתמטי מעמיק של האופן שבו רשותות Transformer מקודדות מידע על מיקום רצף ללא רקורסיה. אנו מתחילה בסקירה הבסיסי הדו-מיידי ובונים בהדרגה לקידודים רב-מיידיים המשמשים במודלים מודרניים של שפה.

הניתוח משתמש באלגברה לינארית ומתמטיקה סטטיסטית כדי להסביר את המאפיינים הגיאומטריים, ערבויות הייחודיות, ומאפייני התדר של קידודי מיקום. באמצעות הדמיות מפורטות וגזרות מתמטיות קפדיות, אנו מדגימים מדוע התקדמות אקספוננציאלית של אורכי גל מאפשרת לרשותות Transformer למכוד קשרי מיקום מקומיים וגלובליים כאחד. היסודות התיאורטיים מלווים בIMPLEMENTATION Python מעשיים, מה שהופך עבודה זו לבעלת ערך עבור חוקרים ומתרגלים המבקשים הבנה عمוקה של מנגנון קידוד מיקום ברשותות Transformer.

밀ות מפתח: ארכיטקטורת Transformer, קידוד מיקום, פונקציות סינוסואידליות, מנגנון קשב, מודולינג רצפים, עיבוד שפה טבית, למידה עמוקה, התקדמות אורכי גל, תכונות Fourier, התקדמות גיאומטרית.

Abstract

This comprehensive study examines the sinusoidal positional encoding mechanism introduced in the seminal "Attention is All You Need" paper by Vaswani et al. The work provides an in-depth mathematical treatment of how transformers encode sequential position information without recurrence. We begin with the fundamental two-dimensional case and progressively build to high-dimensional encodings used in modern language models.

The analysis employs linear algebra and statistical mathematics to explain the geometric properties, uniqueness guarantees, and frequency characteristics of positional encodings. Through detailed visualization and rigorous mathematical derivations, we demonstrate why the exponential wavelength progression enables transformers to capture both local and global positional relationships.

The theoretical foundations are complemented by practical Python implementations, making this work valuable for researchers and practitioners seeking deep understanding of transformer positional mechanisms.

Transformer architecture, positional encoding, sinusoidal functions, attention **Keywords:** mechanism, sequence modeling, natural language processing, deep learning, wavelength progression, Fourier features, geometric progression.

תוכן העניינים

vi	תקציר
v	Abstract
xi	רישימת איורים
x	רישימת טבלאות
1	1 קידוד מיקום דו-ממדי: Two-Dimensional Positional Encoding
1	1.1 הקדמה
1	1.2 הבעה: מנגנון קשב בלתי תלוי במיקום
1	1.3 גישות היסטוריות לקידוד מיקום
2	1.4 רקע תיאורטי: פונקציות סינוסואידליות
3	1.4.1 ייצוג מעגל היחידה
3	1.4.2 פרספקטיבת אלגברה ליניארית
4	1.4.3 פרספקטיבת ניתוח פוריה
4	1.5นำไปש בPython-
6	1.6 ניתוח הויזואלית
6	1.6.1 תת-graf שמאלי עליון: ערכי רכיבים לאורך מיקום
7	1.6.2 תת-graf ימני עליון: ייצוג מפתח חום
8	1.6.3 תת-graf שמאלי תחתון: מסלול מעגלי במרחב דו-ממדי
8	1.6.4 תת-graf ימני תחתון: מהזרויות רצף מרווח
9	1.7 עיצוב הרצינול: למה הויזואליות האלו?
01	1.8 ניתוח عمוק: תוכנות מתמטיות
01	1.8.1 תוכנה 1: גודל מוגבל
01	1.8.2 תוכנה 2: שינוי חלק
11	1.8.3 תוכנה 3: מיקום יחסית דרך טרנספורמציה ליניארית
21	1.8.4 תוכנה 4: אורתוגונליות
21	1.9 גישות אלטרנטיביות והרחבות
31	1.9.1 הטמעות מיקום נלמדות
31	1.9.2 ייצוגי מיקום יחסית
31	1.9.3 ALiBi (קשב עם הטוות ליניאריות)
41	1.9.4 קידודים בערכאים מורכבים
41	1.9.5 קידודים רב-סקאלרים
41	1.10 קשר לארכיטקטורת Transformer
51	1.11 סיכום

61	N-Dimensional Positional Encoding	2
61	הקדמה: משבב העמימות	2.1
61	הרקע ההיסטורי: מעיבוד אותות לעיבוד שפה	2.2
71	הנוסחה הרב-תדרית: מעבר למגל	2.3
81	2.3.1 זיוג ממדים ומסלולים מעגליים	
81	2.3.2 הצדקה לוח זמני התדרים	
91	מימוש ב <code>Python</code> : הכללה לא- ממדים	2.4
02	ניתוח הוייזואליזציה: מפות חום בסקלנות מרובות	2.5
02	תת-גרף שמאל עליון: $d_{model} = 8$	2.5.1
12	תת-גרף ימני עליון: $d_{model} = 32$	2.5.2
12	תת-גרף שמאל תחתון: $d_{model} = 128$	2.5.3
22	תת-גרף ימני תחתון: $d_{model} = 512$	2.5.4
22	ניתוח השוואתי על פני תת-הגרפים	2.5.5
32	תכונות גיאומטריות: המבנה הטופולוגי	2.6
32	2.6.1 תוכנה 1: מבנה סעפת טורוס	
32	2.6.2 תוכנה 2: תכונות מרחק	
42	נושאים מתקדמים: אלטרנטיבות ורחבות	2.7
42	אתחול תדרים נלמד	2.7.1
52	(RoPE) Rotary Position Embedding	2.7.2
52	ALiBi: קשב עם הטוות לינאריות	2.7.3
52	הקשר לניתוח פוריה	2.8
62	סיכום: מיקוד פשוט לייצוג עשיר	2.9
72	יחידות ואינטראפולציה: Uniqueness and Interpolation	3
72	הבעיה: הוכחת יחידות	3.1
82	רקע תיאורטי: מטריקות מרחק ויחידות	3.2
82	3.2.1 הפרשי מיקום קטנים	
92	3.2.2 הפרשי מיקום גדולים	
92	3.2.3 קרייטריון יחידות	
03	3.2.4 השערת אינטראפולציה לינארית	
03	מימוש <code>Python</code> : חישוב מרחק	3.3
53	ניתוח המכחשה	3.4
53	תת-תרשים שמאל עליון: מטריצת מרחק	3.4.1
63	תת-תרשים ימני עליון: מרחק לעומת הפרש מיקום	3.4.2
63	תת-תרשים שמאל תחתון: שגיאת אקסטרפלציה לינארית	3.4.3
73	תת-תרשים ימני תחתון: נורמה של וקטורי קידוד	3.4.4
73	עיצוב והنمקה: למה המחשות אלה?	3.5
83	ניתוח עמוק: ערביות יחידות כמותיות	3.6
93	גישות חלופיות: שיפור יחידות	3.7

93	ממדים גבויים יותר	3.7.1
93	הטמעות מיקום נלמדות עם אילוצי יחידות	3.7.2
93	אילוצי אורתוגונליות מפורשת	3.7.3
93	הטלות אקרראיות	3.7.4
04	קשר לתורת המדע	3.8
04	מסקנה	3.9
24	Geometric Progression of Wavelengths	4
24	הקדמה	4.1
24	הבעיה: בחירת התפלגות אורך גל	4.2
34	הקשר ההיסטורי: ניתוח רב-זרולוציה בעיבוד אותות	4.3
34	רקע תיאורטי: סדרות גיאומטריות ולוחות זמנים של אורך גל	4.4
44	מרוחך לוגריטמי	4.4.1
44	פרספקטיבת תדר	4.4.2
54	יישום Python: חישוב והדמיה של אורך גל	4.5
64	ניתוח ההדמיה	4.6
64	תת-עלילה שמאלית: אורך גל בסקלה לוגריטמית	4.6.1
84	תת-עלילה מרכזית: יחס אורך גל עוקבים	4.6.2
84	תת-עלילה ימנית: סינוסואידיים לדוגמה בתדרים שונים	4.6.3
94	ניתוח عمוק: אופטימליות וטריאיד-אופים בעיצוב	4.7
94	בחירה בסיס: מדוע 10000?	4.7.1
94	בחירה מעירך: מדוע $?2i/d_{model}$	4.7.2
94	נושאים מתתקדים: יסודות תיאורתיים ורחבות	4.8
94	תורת קירוב Fourier	4.8.1
05	חוקי סקלה ורחבות אורך רצף	4.8.2
05	מסקנות	4.9
4.10 English References	52

רשימת האיורים

- 1 ארבעה מבטים משלימים של קידוד מיקום דו-ממדי. למעלה שמאל: ערכי
רכיבים לעומת מיקום. למטה מימין: ייצוג מפת חום. למטה שמאל:
מסלול מעגלי במרחב D². למטה מימין: מחזוריות רצף מורחב.
- 6 2 קידוד מיקום רב-ממדי עבור $\{8, 32, 128, 512\} \in d_{model}$. מפות חום מציגות
את 50 המיקומים הראשונים לעומת כל הממדים. צבע מקודד ערך קידוד
02 מ-1- (כחול) ל-1+ (אדום).
- 53 3 ניתוח יחידות אוינטראופולציה של קידוד מיקום. שמאל לעומת: מטריצת
מרחיק המראה מרחקים זוגיים. מימין לעומת: מרחק לעומת הפרש מיקום.
משמאלי לעומת: שגיאת אקסטרופולציה לינארית. מימין לעומת: נורמות
וקטורי הקידוד.
- 4 התקדמות גיאומטרית של אורכי גל: (שמאל) אורכי גל בסקרה לוגריתמית
מציגים צמיחה אקספוננציאלית לינארית; (מרכז) יחס קבוע בין אורכי גל
עוקבים ≈ 1.018 ; (ימין) סינוסואידיים בתדרים שונים מהירים (הבחנה
מקומית) לאיטיים (מיקום גלובלי).
- 64

רשימת הטבלאות

1 קידוד מיקום דו-ממדי: Two-Dimensional Positional Encoding

1.1 הקדמה

ארכיטקטורת Transformer - חוללה מהפכה בעיבוד רצפים כשהיא נטשה לחולוטן את עקרון הרקורסיה. בעוד רשתות נוירונים רקורסיביות מעבדות טוקנים באופן סדרתי, אחד אחרי השני, Transformers - מעבדים את כל המיקומים במקביל באמצעות מנגנון קשב מקבילים. המקביליות הזה מאפשרת יעילות חישובית אדירה, אך יצרת בעיה יסודית: המודל אינו מקבל מידע אינהרנטי על סדר הטוקנים. ללא מידע מיקומי, Transformer - אינו יכול להבחין בין "הכלב נשך את האיש" לבין "האיש נשך את הכלב". משפטים אלו מכילים טוקנים זהים אך משמעויות הפוכות. המיקום חשוב באופן קריטי בשפה טבעית.

הפתרון שהוצע על ידי Vaswani וחבריו בשנת 2017 [1] משבץ בצוරה אלגנטית מידע מיקומי ישירות לתוך ייצוגי הטוקנים. במקומות למדוד הטעמוות מיקום מהנתונים, הם בחרו בפונקציות סינוסואידליות קבועות. פרק זה חוקר את המקרה פשוט ביותר: קידוד מיקום דו-ממדי. הבנת יסוד זה מאפשרת הבנה של קידודים במדדים גבוהים יותר המשמשים במערכות ייצור.

1.2 הבעיה: מנגנון קשב בלתי תלוי במיקום

מנגנון self-attention - מחשב יחסים בין כל צמדי הטוקנים. נבחן רצף של טוקנים המייצג כווקטורים. יהיו X מטריצת הקלט בגודל ($\text{sequence_length}, d_{\text{model}}$), כאשר d_{model} מייצג את ממד ההטמעה. כל שורה מתאימה למיקום טוקן אחד. מנגנון הקשב משנה את X דרך השלכות שאילתה, מפתח וערך נלמדות. החישוב המרכזי עוקב אחר הנוסחה זו:

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

כאן Q מצינית את מטריצת השאלות, K מצינית את מטריצת המפתחות, V מצינית את מטריצת הערכאים. הממד d_k מייצג את ממד המפתח. הסופרסקרייפט T מצין טרנספוזיציה של מטריצה. הפונקציה softmax מנורמלת משקלים קשב על פני ממד הרצף. מנגנון זה אינו מכיל תלות מפורשת במיקום. אם נבצע תמורה לשורות של X , נבצע תמורה לפلت באופן זהה. המודול מתיחס לרצפים כל קבוצות ולא כל רצפים מסודרים. עבור משימות שפה רבות, זה מוכיח את עצמו קטלני. ת לחבר תלוי בסדר המילים. פתרון שם עצם חוזר תלוי במיקום יחסית. יחסים טמפורליים תלויים במבנה סדרתי.

1.3 גישות היסטוריות לקידוד מיקום

מודלים נוירוניים מוקדמים לעיבוד רצפים טיפולו במיקום באופן מרומז דרך בחירות ארכיטקטוניות. רשתות נוירונים רקורסיביות מעבדות טוקנים באופן סדרתי. המצב הנستمر

במיקום t תלוי במיקום $1 - t$, שתליי $2 - t$, וכן הלאה. מידע מיקומי זורם דרך התלויות הטופורליות הללו. עם זאת, השימוש בסדרתי זה מונע מקובל וסובל מדעיכת גרדיאנטים על רצפים ארוכים [2].

רשותות קונבולוציה נוירונים לרצפים משתמשות במיקום דרך שדות קליטה. כל מיקום פלט תלוי בחלון מקומי של מיקומי קלט. המיקום מוקוד דרך היסטים יחסיים בתוך קונבולוציות. אך קונבולוציות סטנדרטיות מוגבלות בצמיחה שדה הקליטה. לכידת תלויות ארוכות טווח דורשת שכבות רבות [3].

הטמעות מיקום נלמדות מציאות גישה אחרת. אנחנו יכולים להתייחס לכל אינדקס מיקום כמשתנה קטגוריאי. מיקום 0 מקבל וקטור נלמד אחד. מיקום 1 מקבל אחר. הטמעות אלו מתאימות לצד משקלים המודלים. גישה זו עובדת אך יש לה חסרונות. המודל אינו יכול להכליל לרצפים ארוכים יותר ממה שראה במהלך האימון. כל מיקום דרוש פרמטרים נפרדים. טבלת ההטמעה גדלתה ליניארית עם אורך הרצף המקסימלי [4].

קידוד מיקום סינוסואידלי מתמודד עם מגבלות אלו באלגנטיות. הקידוד משתמש בפונקציות מתמטיות קבועות ולא בפרמטרים נלמדים. הפונקציות מכללות באופן טבעי אורייני רצף שרירותיים. המבנה המחזורי מאפשר למודול ללמידה יחסית מיקום יחסיים דרך צירופים ליניאריים פשוטים. תכונות אלו הופכות את הקידוד הסינוסואידלי לבחירה הדומיננטית בימושי Transformers.

1.4 רקע תיאורטי: פונקציות סינוסואידליות

קידוד המיקום הדו-ממדי משתמש בפונקציית סינוס אחת ובפונקציית קוסינוס אחת. נבחן אינדקס מיקום p , שלוקח ערכים שלמים $0, 1, 2, \dots$, וכן הלאה. הקידוד ממפה את p לוקטור דו-ממדי. יהיו $\text{PE}(p)$ מציין את וקטור הקידוד המיקומי הזה. הנוסחה מגדרה:

$$\text{PE}(p, 0) = \sin\left(\frac{p}{10000^{0/d_{\text{model}}}}\right)$$

$$\text{PE}(p, 1) = \cos\left(\frac{p}{10000^{0/d_{\text{model}}}}\right)$$

במקרה הדו-ממדי, $d_{\text{model}} = 2$. האקספוננט $0/d_{\text{model}}$ שווה ל-0. לכן 10000^0 שווה ל-1. הנוסחאות מתחשפות ל:

$$\text{PE}(p, 0) = \sin(p)$$

$$\text{PE}(p, 1) = \cos(p)$$

כל מיקום p ממופה לקואורדינטות $(\sin(p), \cos(p))$ במרחב דו-ממדי. למיפוי זהה יש תכונות גיאומטריות עמוקות.

1.4.1 ייצוג מעגל היחידה

פונקציות הסינוס והקוסינוס מפרמטרות את מעגל היחידה. לכל זווית θ , הנקודה $(\cos(\theta), \sin(\theta))$ שוכנת על המעגל בردיויס 1 במרכזה הראשית. כאשר θ משתנה מ-0 ל- 2π , נקודת זו מושרטת את המעגל כולו. כאשר אנו מקודדים מיקום p כ- $(\cos(p), \sin(p))$, אנו מופיעים מיקומים על מעגל היחידה.

שיםנו לב שאנו משתמשים ב- (\cos, \sin) ולא ב- (\sin, \cos) . בחירה זו משפייעת על כיוון ההתחלת אך לא על התכונות הבסיסיות. במיקום 0, הקידוד נתן $(0, 1) = (\sin(0), \cos(0))$. במיקום $\pi/2$, אנו מקבלים $(1, 0) = (\cos(\pi/2), \sin(\pi/2))$. הקידוד מתחילה בחלק העליון של המעגל ומסתובב בכיוון השעון בקואורדינטות מתמטיות סטנדרטיות.

המיפוי אינו חד-חד-ערכי. מיקום p ומיקום $\pi + p$ מקבלים קידודים זהים כי לסינוס ולкосינוס יש תקופה של 2π . הקידוד "מתפרק" כל $2\pi \approx 6.28$ מיקומים. מחזוריות זו יוצרת מבנה מחלקות שקלילות. מיקומים המופרדים על ידי כפולות מדioיקות של 2π הופכים לבתאי ניתנים להבנה. עבור אינדקסי מיקום קטנים ביחס ל- 2π , זה לא מהוות בעיה. עבור רצפים ארוכים יותר, קידודים במדדים גבוהים יותר פותרים את העמימות.

1.4.2 פרספקטיבת אלגברת ליניארית

נראה את הקידוד המיקומי כפונקציה $\mathbb{Z} \rightarrow \mathbb{R}$: PE מהשלמים לקטורים דו-ממדיים ממשיים. פונקציה זו חלקה, רציפה (כאשר מורחבת למיקומים ממשיים), ומחזורת. התוחם \mathbb{R}^2 יוצר מרחב וקטורי עם חיבור וכפל סקלרי סטנדרטיים. עם זאת, הקידודים המיקומיים אינם פורשימים את כל המרחב הזה. הם שוכנים על הסעפת החד-ممדיות המוגדרת על ידי מעגל היחידה.

הנזרת של פונקציית הקידוד מספקת את הוקטור המשיק בכל מיקום:

$$\frac{d}{dp} \text{PE}(p, 0) = \cos(p)$$

$$\frac{d}{dp} \text{PE}(p, 1) = -\sin(p)$$

локטור המשיק תמיד יש גודל 1, כפי שמאומת על ידי $1 = \cos^2(p) + \sin^2(p)$. הקידוד נع לאורץ המעגל במהירות קבועה כאשר המיקום גדול. תנועה אחת זה אומרת שמיוקומים סמוכים יש להם קידודים דומים כאשר אינדקס המיקום קטן בהשוואה לאורץ הנל.

1.4.3 פרספקטיבת ניתוח פורייה

פונקציות סינוסואידליות מהוות את הבסיס של ניתוח פורייה. כל פונקציה מחזורת יכולה להיפרक לסכומים של סינוסים וкосינוסים בתדרים שונים. הקידוד המיקומי משתמש בתדר יחיד: $(\frac{1}{2\pi}) \cos(2\pi f t)$. אורך הגל שווה ל- $\frac{1}{f} = 2\pi$ מיקומים. במנוחי תחום תדר, הקידוד מקרין את המיקום על רכיב פורייה יחיד. קידודים בממדים גבויים יותר משתמשים בתדרים מרובים בו-זמן, יוצרים ייצוג תכונות פורייה של מיקום [5]. פרספקטיביה זו מסבירה מדוע הקידוד מאפשר למודלים ללמוד פונקציות מיקום יחסיות. צירופים ליניאריים של רכיבי פורייה יכוליםים לקרב פונקציות חלקות רבות דרך ניתוח הרמוני סטנדרטי [6].

1.5 מימוש בPython

קוד הוויזואלייזציה מממש קידוד מיקומי בקובץ `positional_encoding.py`. בואו נבחן את המימוש הזה צעד אחר צעד.

פונקציית קידוד מיקום סינוסואידלי

```
def positional_encoding(pos, d_model):
    """Sinusoidal_positional_encoding_(Vaswani et al., 2017)."""
    # Convert position to array format
    pos = np.array([pos]) if isinstance(pos, int) else np.array(pos)

    # Generate frequency indices
    i = np.arange(d_model // 2)

    # Compute frequencies for each band
    freqs = 1.0 / (10000 ** (2 * i / d_model))

    # Initialize output array
    pe = np.zeros((len(pos), d_model))

    # Fill even dimensions with sine values
    pe[:, 0::2] = np.sin(pos[:, None] * freqs)

    # Fill odd dimensions with cosine values
    pe[:, 1::2] = np.cos(pos[:, None] * freqs)

    return pe
```

הfonקציה מקבלת שני פרמטרים. הפרמטר `pos` מצין אינדקסי מיקום, או שלם בודד או מערך של שלמים. הפרמטר `d_model` מצין את ממד הקידוד. הfonקציה מחזירה מערך `numpy` בגודל `(len(pos), d_model)`.

השורה הראשונה מבטיחה ש`pos`- הופך למערך `numpy.array`. אם `pos` הוא שלים, הוא הופך למערך בעל אלמנט יחיד. נורמליזציה זו מאפשרת עיבוד אחיד בהמשך.

השורה השנייה יוצרת מערך `i` של אינדקסי תדר. עבור `d_model=2`, אנו מקבלים $[0] = i$, פס תדר בודד. המשטנה `i` מאנדקס רכיבי תדר סינוסואידליים שונים.

השורה השלישייה מחשבת תדרים עבור כל פס. הנוסחה $(\cos(i * 2\pi / d_model) * 10000)^{1.0}$ מימוש את לוח זמני התדרים. עבור `i=0`, האקספוננט הופך ל-0, נותן תדר 1. זה מתאים לאורץ גל 2π .

השורה הרביעית מתחילה את מערך הפלט `pe` באפסים. לערך זה יש גודל `num_positions, d_model`.

השורה החמישית מלאת ממדים עם אינדקס זוגי (... , 0) עם ערכי סינוס. הביטוי `None * freqs[pos[None], :]` משדר ערכי מיקום כנגד תדרים, יוצר מטריצה שבה אלמנט (i, k) שווה $\cos(i * pos[k] * freqs)$. הפעלת סינוס אלמנט אחר אלמנט נתנת את רכיבי הסינוס.

השורה הששית מלאת ממדים עם אינדקס אי-זוגי (... , 1) עם ערכי קוסינוס. אותו שידור יוצר את מכפלות מיקום-תדר, שעוברות דרך קוסינוס.

הצרת ההחזרה מספקת את מטריצת הקידוד השלמה.

עבור המקרה הדו-ממדי, זה מוצטמצם לחישוב $\sin(p) \text{ и } \cos(p)$ עבור כל מיקום.

קוד השקופית בـ `Slide_1/slide1.py` מייצר ויזואלייזציות של הקידוד הזה. בואו נבחן את השלבים המרכזיים:

הגדרת פרמטרים ויצירת קידודים

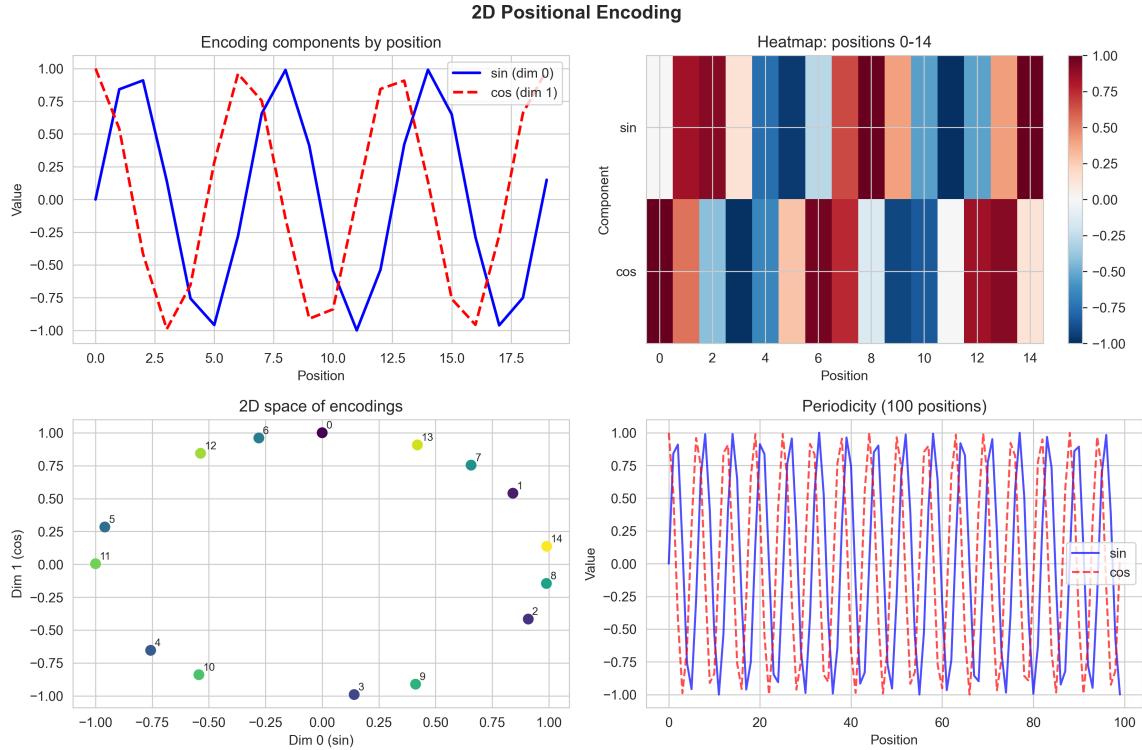
```
# Setup encoding parameters
d_model = 2
positions = np.arange(0, 20)

# Generate positional encodings
pe = positional_encoding(positions, d_model)
```

שורות אלו מגדירות את הקידוד עבור 20 מיקומים (מ-0 עד 19) ב-2- ממדים. לערך `pe` יש גודל $(20, 2)$.

הויזואלייזציה יוצרת ארבעה תת-גרפים מסוודרים בראשת 2×2 . כל תת-graf חושף היבטים שונים של מבנה הקידוד המיקומי. הבנת כל graf בונה אינטואיציה מקיפה.

1.6 ניתוח הוייזואלייזציה



איור 1: ארבעה מבטים משלימים של קידוד מיקום דו-ממדי. למעלה שמאל: ערכי רכיבים לעומת מיקום. למעלה מימין: ייצוג מפת חום. למטה שמאל: מסלול מעגלי במרחב 2D. למטה מימין: מחזוריות רצף מורחב.

האיור לעיל מציג ארבעה מבטים משלימים של קידוד מיקום דו-ממדי. כל תת-גרף מתמודד עם שאלת ספציפית על התנהלות הקידוד. אנו בוחנים כל אחד בפירוט.

1.6.1 תת-גרף שמאלי עליון: ערכי רכיבים לאורך מיקום

תרשים קו זה מציג כיצד רכיבי הסינוס והקוסינוס משתנים כאשר המיקום גדל מ-0 ל- π . הציר האופקי מייצג את אינדקס המיקום p , מספרים שלמים חיוביים מתחילה ב-0. הציר האנכי מייצג את ערך הקידוד, כמות חסרת ממדים הנעה בין -1 ל-1. העקומה הכחולה המוצקה מראה את רכיב הסינוס $\text{sin}(p, 0) = \text{sin}(p)$. במיקום 0, סינוס שווה ל-0. העקומה עולה, מגיעה ל-1 ליד מיקום $1.57 \approx \frac{\pi}{2}$. היא חוצה אפס ליד מיקום $3.14 \approx \pi$. היא מגיעה לערך המינימלי -1 ליד מיקום $4.71 \approx 3\pi/2$. העקומה משלימה תנודה מלאה אחת במיקום $6.28 \approx 2\pi$.

העקומה האדומה המקווקה מראה את רכיב הקוסינוס $\text{cos}(p, 0) = \text{cos}(p)$. במיקום 0, קוסינוס שווה ל-1. העקומה יורדת, חוצה אפס ליד מיקום $0.79 \approx \pi/2$. היא מגיעה ל-1 ליד מיקום π . היא חזרת לאפס ליד מיקום $2.36 \approx 3\pi/2$ ומשלימה את המזור במיקום 4π . שתי העקומות בעלות צורה זהה אך שונות בפазה. הקוסינוס מוביל את הסינוס ב- $2\pi/\pi = 2$ רדייאנים (רבע מחזור). יחס פאזה זה יסודי לפרמטריזציה מעגלית. בכל מיקום, $\text{sin}^2(p) + \text{cos}^2(p) = 1$.

אורך הגל מופיע ויוזאלית כמרחב בין נקודות מתאימות במחזוריים עוקבים. אנו צופים בערך 6.28 יחידות מיקום למחזור שלם, מאשר את אורך הגל λ . על פני 20 מיקומים, אנו רואים בערך 3.2 מחזוריים שלמים.

למה לשרטט את שני הרכיבים יחד? ויוזאליזציה זו מדגימה את יחס הפאה ומאשרת שתי הפונקציות מתנדדות עם תדר ואmplיטודה זהים. אם הקידוד יושם בצורה לא נכון, אי התאמות יופיעו מיד. החפיפה גם מראה כיצד שני הממדים מספקים מידע משלים. כאשר הסינוס קרוב לאפס, הקוסינוס קרוב לערכי הקיצון שלו, ולהפך. אף מיקום אינו מקבל קידוד קרוב בראשית (0,0).

1.6.2 תת-גרף ימני עליון: ייצוג מפת חום

מפת החום הזו מספקת תצוגה חלופית של אותו מידע כמו התרשים השמאלי העליון. הציר האופקי שוב מייצג מיקום מ-0 ל-14 (מראה רק את 15 המיקומים הראשונים לבHIRות). הציר האנכי מייצג את אינדקס הרכיב, עם שתי שורות: שורה 0 עבור סינוס ושורה 1 עבור קוסינוס.

צבע מקודד את ערך הקידוד. מפת הצלבים משתמשת ב" z_{RdBu} " (פלטת אדום-כחול מתפזרת הפוכה). אדום עמוק מצין ערכים קרובים ל-1+. כחול עמוק מצין ערכים קרובים ל-1-. צבעים לבנים או בהירים מצינינги ערכים קרובים ל-0.

בבדיקה שורת הסינוס (shoreה תחתונה, מסומנת "sin"), אנו צופים בהתקדמות מלבן דרך אדום, חזרה מלבן, לכחול, וחזרה לכיוון לבן. תבנית צבע זו עוקבת אחר עקומת הסינוס מהתרשים השמאלי העליון. מיקום 0 מראה לבן ($\sin(0) = 0$). מיקומים ליד 1.5 מראים אדום (סינוס מתקרב ל-1). מיקום π מראה לבן שוב (סינוס חוזר ל-0). מיקומים ליד 4.7 מראים כחול (סינוס מגיע ל-1-).

shoreה הקוסינוס (shoreה עליונה, מסומנת "cos") מראה תבנית משלימה. מיקום 0 מופיע אדום ($\cos(0) = 1$). הצלבים מתקדמיים דרך מלבן ליד מיקום $2\pi/3$, כחול ליד מיקום π , לבן שוב ליד מיקום $2\pi/3$, וחזרה לכיוון אדום משלימים את המחזoor.

מפת החום מדגישה שהקידוד הוא במידהו מטריצה של ערכים מספריים. כל עמודה מייצגת וקטור מיקום אחד. כל shoreה מייצגת כיצד ממד אחד משתנה על פני מיקומים. פרספקטיבת מטריצה זו טبيعית כאשר מממשים Transformers. הטמעות טוקנים יוצרות מטריצה. קידודים מיקומיים יוצרים מטריצה נוספת בגודל זהה. חיבורם אלמנט-אלמנט מייצר ייצוגים מודעי-מיקום.

קריאת מפות חום דורשת תרגול. פס הצלבים מימין מספק את המיפוי ערך-לצבע. אדום אומר ערכים חיוביים, כחול אומר שליליים, ולבן אומר קרוב לאפס. יחס הגובה-רחוב "auto" מונע עיות; מרחקים שוים למרחב הנתונים תואמים למרחקים פרופורציונליים על המסך.

למה להשתמש במפת חום במקום תרשימי קו? מפות חום מצטיינות בהציג נתונים בממדים גבוהים. עם רק שני ממדים, תרשימי קו מספקים. אבל הפרק הבא בוחן קידודים בעלי 512 ממדים [7]. שרטוט 512 קוים הופך למזכיר ויוזאלית. מפות חום מתרחבות באלגנטיות למאות או אלפי ממדים. למידת קריאת ייצוג זה במקרה הפשטוט 2D מכינה אותנו לויזאליזיות מורכבות יותר.

1.6.3 תת-גרף שמאלית תחתון: מסלול מעגלי במרחב דו-ממדי

תרשים פיזור זה חושף את המבנה הגיאומטרי באופן הישיר ביותר. הציר האופקי מייצג את ממד הקידוד הראשוני (רכיב סינוס). הציר האנכי מייצג את הממד השני (רכיב קוסינוס). כל נקודה מתאימה למיקום אחד, משורטთ בקואורדינטות $(\sin(p), \cos(p))$.

הנקודות משרטוטות קשת מעגלית. מיקום 0 מופיע ב $(0, 0)$, חלק העליון של המעגל. המיקום גדול בכיוון השעון (באוריינטציה מתמטית סטנדרטית). הנקודות אינן מרוחות באופן שווה לאורך הקשת כי המיקום גדול ליניארית בעוד הזווית גדלה ברדיאן אחד ליחידה מיקום. למסלול המעגלי יש רדיוס 1, מאמת ש- $1 = \sqrt{\sin^2(p) + \cos^2(p)} = \|\mathbf{PE}(p)\|$ עבור כל המיקומים.

תוויות מספריות מציננות כל נקודה עם אינדקס המיקום שלה. תיוג זה מדגים שמיוקומים עוקבים שכנים קרוב זה לזה על המעגל כאשר המיוקומים קטנים. למיקום 0 ומיוקום 1 יש הפרדה זוויתית של רדיאן 1 בערך 57.3 מעלות. למיקום 5 ומיוקום 6 יש הפרדה זוויתית זהה. הקידוד משמר מהירות זוויתית אחידה.

התבנית המעגלית ממשיכה מעבר למיקום 19. אם הרחבעו למיקום 100, היינו רואים LOLAOות מרובות סביב המעגל. מיוקומים המופרדים ב- $\pi/2$ ממופים למיוקומים דומים מאוד. מיקום 0 ומיוקום 6.28 יש להם קידודים כמעט זהים. מחזריות או מגבילה את יכולת הקידוד להבחין בין מיוקומים מרוחקים. קידוד דו-ממדי מספיק רק לריצפים קצרים מאוד.

סוג ויזואלייזציה זה נקרא דיווקן פאה או תרשימים מסלול בתורת המערכות הדינמיות. אלו משרטטים מצבאים עוקבים של מערכת מצב. המסלול חושף מבנה מערכת. עבור קידוד מיקומי, המסלול המעגלי מראה שמיוקומים מתפתחים בצורה חלקה דרך סעפת קומפקטיבית. אף מיוקום אינו מקבל קידוד שונה באופן דרמטי ממיקומים סמוכים.

למה להשתמש בויזואלייזציה זו? המבנה המעגלי אינו ברור מתרשיימי קו או מפות חום. ראיית המסלול הופכת מספר תכונות לגלוויות מיד. ראשית, לקידודים יש גודל מוגבל. שנייה, למיוקומים סמוכים יש קידודים דומים. שלישיית, למיוקומים מרוחקים יוכולים להיות קידודים דומים בגלל מחזריות. רביעית, הקידוד סימטרי סביב הראשית. אינטואיציות גיאומטריות אלו מודיעות כיצד מנגנוני קשב עשויים להשתמש במידע מיקומי [8].

1.6.4 תת-גרף ימני תחתון: מחזריות רצף מורחב

תרשים זה מרחיב את טווח המיקום ל-100 כדי להציג התנוגות מחזרית בצורה ברורה יותר. שני הצירים תואמים לאלו מהתת-גרף השמאלי העליון: מיקום אופקי, ערך קידוד אנכי. העקומות הכהולה והאדומה שוב מיצגות את רכיבי הסינוס והקוסינוס.

עם צוגה מורחבת זו, אנו צופים במחזרים שלמים רבים. פונקציית הסינוס משלימה בערך 15.9 מחזרים על פני 100 מיוקומים ($15.9 \approx 2\pi/100$). התנוגה הרגולרית ממשיכת ללא שינוי לאורך כל הרצף. לא מתרחשת דעיכה, גדילה, או הסתת פאה. הקידוד במיקום 90 עוקב אחר אותה הבנייה כמו מיקום 0, רק בפאה שונה.

לחזריות יש גם יתרונות וגם חסרונות. היתרון הוא עילות חשובה והכללה ללא פרמטרים. אנחנו לא צריכים משקלים ניתנים ללמידה. הקידוד מתרחב לריצפים לאורך

שרירותי באמצעות אותה נוסחה [9]. החישורו הוא עמיות. למקום 0 ומקום $\pi/2$ יש קידודים דומים מאד. עבור רצפים ארוכים מאורך גל אחד, המודל חייב להשתמש בהקשר קשב כדי לפענח מיקומים.

בפועל, Transformers משתמשים בקידודים בממדים גבוהים עם אורך גל רבים בו-זמןית. אורך גל קבועים מבחינים בין מיקומים סמוכים במדויק. אורך גל ארוכים מבחינים בין מיקומים מרוחקים. קידוד רב-סקלה זה פותר את העמיות הנוכחות במקרה הD2. פרק 2 חוקר פתרון זה לעומק.

שקייפות אלפא ($\alpha=0.7$) הופכת את העמיות לחצי-שוקפות. בחירה זו משפרת קריאות כאשר עמיות חופפות במהלך מהלך מחזוריים רבים. עמיות מוצקות היו מסתיימות זו את זו. שקייפות מאפשרת לראות את כל התבניות בו-זמןית.

庫וי רשות עוזרים לכמת ערכים ויזואלית. קוי רשות אופקיים מסמנים ערכים במרוחקים קבועים. קוי רשות אנכיים מסמנים אבני דרך במקומות. קריאת ערכים מדויקים מוגרים יכולה להיות מתוגרת ללא רשות. הרקע הלבן עם קוי רשות אפורים מספק ניגודיות גבוהה תוך שמיירה על עדינות ויזואלית.

1.7 עיצוב הרצינול: למה הויזואלייזציות האלו?

בחירת ויזואלייזציות מתאימות חשובה כמו בחירת אלגוריתמים טובים. לכל סוג ויזואלייזציה יש חוווקות ומגבלות. אויר ארבעת הפאנלים שלנו משלב פרספקטיביות משלימות לבניית הבנה שלמה.

תרשים הקו (למעלה משמאל) מצטיין בהציג יחסים פונקציונליים. אנו רואים מיד כיצד ערכים משתנים עם מקום. התבנית המחזורת ברורה. השוואת פונקציות מרובות (סינוס לעומת קוסינוס) טبيعית. בני אדם מצטיינים במעקב אחר ציוני קו ושיפועים. עם זאת, תרשימי קו מתרחבים בצורה גרוועה לממדים רבים. עם 512 רכיבים, 512 עמיות חופפות הופכות לבלי קריאות.

מפת החום (למעלה מימין) מתרחבת באלגנטיות לממדים גבוהים. אנחנו יכולים לדמיין מאות ממדים בו-זמןית. זההינו תבניות בתמונה D2 הוא חווקה אנושית. צבע מספק ערזן קידוד נוסף מעבר למיקום מרחבוי [10]. עם זאת, קריאת ערך מדויק קשה. צבעים מתמזגים תפיסתית. להבחן הבדל בין "ערך 0.7" ל"ערך 0.8" מצבע לבדו מתוגר. מפות חום עובדות הכי טוב לזהות תבניות ולא למדידה מדויקת.

תרשים הפיזור (למטה משמאל) חושף מבנה גיאומטרי ישירות. התבנית המעלית ברורה מידי. סוג ויזואלייזציה זה ייחודי בהציג מרחב הקידוד עצמו ולא ערכי רכיבים לעומת מיקום. אנו רואים יחסים בין ממדים בו-זמןית. המחריר הוא שאנו רואים לדמיין רק 2 או 3 ממדים. עבור ממדים גבוהים יותר, היינו צריכים טכניקות הפחחת ממד כמו PCA או t-SNE.

תרשים הקו המורחב (למטה מימין) מדגים התנגדות ארכוכת טווח. התקשרות מראה מבנה מקומי. התרחקות מראה תבניות גלובליות. מחזוריות הופכת ברורה עם טווח נתוניים מספיק. זה מתמודד עם מגבלה של תרשים הקו הראשון, שמאיה פחות מ-4-מחזוריים שלמים. עם +15 מחזוריים, האופי המחזורי בלתי ניתן להכחשה.

ויזואלייזציות חלופיות שיכלנו לבחור כוללות תרשימי קווטר, תרשימי מסלול 3D, אНИמציה, ותרשיימי הבדל. בחירת ארבעת הפאנלים שלנו מażנת מטרות רבות: הצגת ערכי רכיבים, הדגמת ממדיות, חשיפת גיאומטריה, ואישור מחזוריות. כל פאנל תורם מידע ייחודי. יחד, הם מספקים הבנה מקיפה של קידוד מיקום 2D.

1.8 ניתוח عمוק: תוכנות מתמטיות

מספר תוכנות מתמטיות הופכות את קידוד המיקום הסינוסואידלי למתאים במיוחד במיוחד ל-*Transformers*. אלו בווחנים תוכנות אלו עם טיפול קפדי.

1.8.1 תוכנה 1: גודל מוגבל

עבור כל מקום p , לוקטור הקידוד יש גודל קבוע. חישוב הנורמה עוקב:

$$\|\text{PE}(p)\| = \sqrt{\text{PE}(p, 0)^2 + \text{PE}(p, 1)^2}$$

$$\|\text{PE}(p)\| = \sqrt{\sin^2(p) + \cos^2(p)}$$

$$\|\text{PE}(p)\| = \sqrt{1} = 1$$

כל מקום מקבל קידוד בגודל זהה. תוכנה זו ברורה בתת-גרף הימני התחתון שבו כל נורמות הקידוד שוות ל-1. נורמה קבועה אומרת שקידודי מיקום תורמים באופן שווה לכל יציגי הטוקן ללא קשר למיקום. אף מקום אינו מקבל השפעה לא פרופורציונלית בגלל גודל קידוד.

מנקודת מבט של אופטייזציה, קידודים מוגבלים מונעים פיצוץ גרדיאנט. אם לקידודים יכולם להיות גדלים גדולים באופן שרירוני, שלביAIMON מוקדמים עשויים לראות גרדיאנטים עצומים כאשר קידודי מיקום מקיימים אינטראקציה עם משקלים מאוחדים באקראי. גודל מוגבל מיציב דינמיקתAIMON מוקדם.

1.8.2 תוכנה 2: שינוי חלק

הקידוד משתנה באופן רציף וחלק כאשר המיקום גדול. לキחית הנגזרת ביחס למיקום:

$$\frac{d}{dp} \text{PE}(p, 0) = \frac{d}{dp} \sin(p) = \cos(p)$$

$$\frac{d}{dp} \text{PE}(p, 1) = \frac{d}{dp} \cos(p) = -\sin(p)$$

שתי הנגזרות מוגבלות ב-[1], 1]. קצב השינוי הוא עצמו סינוסואיד, חלק ומוגבל. תכונת חלקיות זו מבטיחה שמיוקמים סמוכים יש להם קידודים דומים. עבור הבדלי מיקום קטנים Δp , אנחנו יכולים לקרב:

$$\text{PE}(p + \Delta p) \approx \text{PE}(p) + \Delta p \cdot \frac{d\text{PE}}{dp}(p)$$

קירוב לניארי זה עובד היטב כאשר Δp קטן בהשוואה לאורך הגל. מרחק הקידוד בין מיוקמים סמוכים הוא בערך $1 = \|d\text{PE}/dp\|$, קבוע על פני כל המיוקמים. קצב שינוי אחד זה מתייחס לכל המיוקמים באופן שווה. שינוי חלק אפשרי אינטראקטיבי. אם המודל מתאים על רצפים עד אורך 100, הוא עדין יכול לעמוד רצפים באורך 120 בזמן היסק. הקידודים במיוקומים 120-100 עוקבים אחר אותה תבנית כמו מיוקומים מוקדמים יותר. לא מתרחשת אי-רציפות או שגיאות אקסטרפלציה [11]. זה מנוגד להטעויות מיוקום נלמדות, שאין להן ערכיהם מוגדרים מעבר למקסימום האימון.

1.8.3 תכונה 3: מיקום ייחסי דרך טרנספורמציה לניארית

תכונה יוצאת דופן של קידוד סינוסואידי היא שמידע מיקום ייחסי נגיש דרך טרנספורמציה לניארית. הקידוד במיוקום $k+p$ יכול להיות מובטאת כפונקציה לניארית של הקידוד במיוקום p .

באמצעות נוסחאות חיבור טריגונומטריות:

$$\sin(p+k) = \sin(p) \cos(k) + \cos(p) \sin(k)$$

$$\cos(p+k) = \cos(p) \cos(k) - \sin(p) \sin(k)$$

בצורת מטריצה:

$$\text{PE}(p+k) = M_k \cdot \text{PE}(p)$$

כאשר M_k היא מטריצת הסיבוב:

$$M_k = \begin{bmatrix} \cos(k) & \sin(k) \\ -\sin(k) & \cos(k) \end{bmatrix}$$

מטריצה זו תלויה רק בהיסט k , לא במקומות המוחלט p . מנגנון הקשב יכול ללמידה להחיל סיבובים כאלה. אם המודל לומד למש M_k , הוא מקבל גישה ישירה למידע מיקום יחסית מבלי לחשב $k + p$ במפורש [12].

תמונה זו מתרחשת למדים גבוהים יותר. עם תדרים רבים, לכל פס תדר יש מטריצת סיבוב משלה. למידת צירופים ליניאריים מתאימים מאפשרת למודל לחץ אותן מיקום יחסיים שונים. מבנה מתמטי זה הוא הסיבה שקידוד סינוסואידלי עובד בצורה כל כך יעילה בפועל.

1.8.4 תכונה 4: אורתוגונליות

פונקציות הבסיס של סינוס וקוסינוס אורתוגונליות תחת מכפלות פנימיות מתאימות. נבחן את המכפלת הפנימית הרציפה על פני תקופה אחת:

$$\langle \sin, \cos \rangle = \int_0^{2\pi} \sin(x) \cos(x) dx$$

$$\text{באמצעות הזהות } : \sin(x) \cos(x) = \frac{1}{2} \sin(2x)$$

$$\langle \sin, \cos \rangle = \frac{1}{2} \int_0^{2\pi} \sin(2x) dx$$

$$\langle \sin, \cos \rangle = -\frac{1}{4} [\cos(2x)]_0^{2\pi} = 0$$

פונקציות הסינוס והקוסינוס אורתוגונליות. אורתוגונליות זו מתרחשת לדגימה דיסקרטית כאשר אנו דוגמים מיקומים בצורה מתאימה. רכיבים אורתוגונליים אומרים שהمدדים מקודדים מידע בלתי תלוי. ממד הסינוס לא יכול להיות חזוי מממד הקוסינוס לבדו. שני הממדים תורמים מידע משלים על מיקום.

1.9 גישות אלטרנטיביות והרחבות

בעוד קידוד סינוסואידלי שלוט במימושי Transformer, חוקרים חקרו חלופות והרחבות. הבנת וריאנטים אלו מאירה את מרחב העיצוב.

1.9.1 הטמעות מיקום נלמודות

במקום פונקציות סינוסואידליות קבועות, אנחנו יכולים להתייחס למיקום כמשתנה קטגוריאי עם הטמעות Nelmodot. גישה זו זהה להטמעות מילים אך עבור אינדקס מיקום. כל מיקום d -length_max-model מקבל וקטור בעל ממד d נפרד, מותאם במהלך האימון [4]. יתרונות כוללים גמישות. המודל לומד כל ייצוג מיקום המתאים ביותר למשימה. לא מוטלות הנחות על מחזוריות או חלקות. שימוש מסויימות עשוית להפיק תועלת מייצוגים תלויי-מיקום שקידוד סינוסואידלי לא יכול לספק.

חסרונות כוללים אקסטרופולציה גרועה ותקורת פרמטרים. אם רצפי אימון יש להם אורך מקסימלי 512, למודל אין קידוד מוגדר למיקום 600. علينا לקטוע רצפים או להשתמש בהיריסטיות (כמו לעשות שימוש חוזר בהטמעה למיקום 511). כל מיקום דורש פרמטרי d , הופך את טבלת ההטמעה לגדולה. עבור $d=512$ ו- $d=1024$, אנחנו צריכים מעל 500,000 פרמטרי מיקום.

השוואות אמפיריות מראות שהטמעות Nelmodot וסינוסואידליות מבוצעות באופן דומה במשימות רבות [13]. חלק מהחוקרים משתמשים בהטמעות Nelmodot למשימות עם תפקידי מיקום מובנים מאוד (כמו שפות תכנות עם היקפים מסוימים). רוב מודלי השפה הגדולים שומרים על קידודי מיקום סינוסואידליים או קשורים.

1.9.2 ייצוגי מיקום יחסיב

במקום לקודד מיקום מוחלט, אנחנו יכולים לקודד מיקום יחסיב ישירות. מגנון הקשב מחשב קשב בין צמדי מיקום (j, i). אנחנו יכולים לספק את ההיסט ($i-j$) לחישוב הקשב. Shaw וחבריו בשנת 2018 [14] הציעו ייצוגי מיקום יחסיב שמשפרים קשב עם הטמעות מיקום יחסיב Nelmodot. במקומות רק קשב בין הטמעות טוקן, הקשב גם שוקל מיקום יחסיב. גישה זו בונה באופן מפורש אינוריאנטיות של תרגום לארכיטקטורה.

סיבוב שמופקות מפונקציות סינוסואידליות ישירות לוקטור שיאלתיה ומפתח. גישה אלגנטית זו כופה את תוכנת הטרנספורמציה הליניארית ישירות בחישוב הקשב. RoPE הפק פופולרי במודלי שפה גדולים אחרים כמו LLaMA ו-PaLM.

1.9.3 ALiBi (קשב עם הטויות לנינאיות)

Press וחבריו בשנת 2022 [9] הציעו ALiBi, שמוסיף הטיה סטטistica לצינוי קשב המבוססים על מרחק מיקום. במקום להוסיף הטמעות מיקום להטמעות טוקן, ALiBi מטה לוגיטים של קשב ב- $|j-i| - m$ – כאשר m הוא שיפוע ספציפי לראש ו- $|j-i|$ הוא המרחק בין מיקומים i ו- j .

גישה זו חסרת פרמטרים, מבצעת אקסטרופולציה מושלמת לרצפים ארוכים יותר, ומפשיטת שימוש. ALiBi מבטל הטמעות מיקום לחלוין, מחליף אותן בהטויות קשב. עבודה אחרתה מראה ShBiALiBi- – מאפשר אימון על רצפים קצרים והיסק על רצפים ארוכים הרבה יותר עם ירידה מינימלית.

1.9.4 קידודים בערכים מרוכבים

חלק מהחוקרים חקרו רشتות נוירונים בעלות ערכים מרוכבים עם מיקום מיוצג כפאות מרוכבות [15]. מיקום p ממופה ל- $(ip/\text{wavelength})\exp(i\phi)$, כאשר i הוא היחידה המדומה. יציג זה מוקדד ביציפות את הסינוס והקוסינוס כחלקיים ממשיים ומדומים של מספר מרוכב יחיד.

רשתות מרוכבות נשארות פחות נפוצות בגלל מרוכבות מיושן ויתרונות מוגבלים על רשתות בעלות ערכים ממשיים עם רכיבי סינוס-косינוס מזוגים. השיקילות המתמטית אומرت ששם גישה אין לה יתרונות אינגרנטיים.

1.9.5 קידודים רב-סקאלאים

אפילו בD2-, יכולנו להשתמש בתדרים מרובים. במקום רק $(\sin(p), \cos(p))$, אולי נשתמש ב- $(\sin(p), \cos(p), \sin(p/10), \cos(p/10))$. זה יוצר קידוד D 4 עם שני אורך גל: 2π ו- $\pi/20$. אורך הקצר מבין מיקומים סמוכים במדויק. אורך הגל הארוך מבין מיקומים מרוחקים.

רעיון זה מכיל לממדים שרירתיים, וזה בדיקת מה שקידוד מיקום סטנדרטי בממדים גבוהים עשויה. כל זוג ממדים משתמש באורך גל שונה. פרק 2 חוקר גישה רב-סקאלית זו לעומק.

1.10 קשר לארכיטקטורת Transformer

קידוד מיקומי משתלב ב-Transformers- בשכבה הקלט. יהי X_{tokens} מצין את מטריצת הטמעת הטוקן בגודל $(\text{sequence_length}, d_{\text{model}})$. יהי X_{pos} מצין את מטריצת הקידוד המיקומי בגודל זהה. הקלט מודע-מיקום הופך:

$$X_{\text{input}} = X_{\text{tokens}} + X_{\text{pos}}$$

חיבור אלמנט-אלמנט משלב מידע טוקן סמנטי עם מידע מיקומי. שכבות עוקבות מעבדות דרך פעולה קשב וfeedforward-. האות המיקומי מתפשט דרך הרשות, משפייע על תבניות קשב ותחזיות [16].

למה להוסיף מקום לשרש? שרשור יכפיל את מידע הקלט. חיבור משמר ממדיות תוכן ערבות מידע. המודל לומד לפרקי מקום וסמנטיקה דרך אימון. אמפירית, חיבור עובד היטב והוא יעיל חישובית.

מנגנון הקשב יכול לנצל מידע מיקומי ללמידה התבניות תלויות-מקום [17]. משקלי קשב בין טוקנים במיקומים i ו- j תלויים גם בתוכן הטוקן וגם במיקום. המודל עשוי ללמידה שנושאים מופיעים בדרך כלל לפני פעילים באנגלית, או שמות תואר בדרך כלל קודמים לשמות עצם. התבניות אלו מופיעות מהאינטראקטיה בין הטמעות סמנטיות וקידודי מקום.

1.11 סיכום

קידוד מיקום דו-ממדי מספק דוגמה מינימלית אך שלמה של איך Transformers מקודדים מיקום. זוג הסינוס-קוסינוס ממפה אינדקסי מיקום לנקודות על מעגל היחידה. קידוד זה דטרמיניסטי, חסר פרמטרים, ומשתנה בצורה חלקה. המסלול המעגלי מבטיח גודל מוגבל ומאפשר מיקום ייחסי דרך מטריצות סיובוב.

הויזואלייזציות חשובות היבטים משלימים של התנחות קידוד. תרשימי קו מראים שינוי רכיבים. מפות חום מספקות פרספקטיבת מטריצה. תרשימי פיזור חשובים מבנה גיאומטרי. תרשימים מוחכים מדגימים מחזוריות. יחד, תצוגות אלו בונות אינטואיציה לאיך קידודים משנים מיקום דיסקרטי למרחבי וקטורי רציף.

המקרה ה $2D$ -סובל מקיבולת מוגבלת. עם תדר אחד בלבד, קידודים חוזרים כל 2π מיקומים. עבור רצפים ארוכים מכמה טוקנים, עמיימות מתעוררת. הפתרון הוא ממדיות. באמצעות שימוש בממדים עם תדרים שנבחרו בקפידה, אנחנו יכולים לקודד לפחות מיקומים באופן ייחודי תוך שמירה על התכונות המועלות של פונקציות סינוסואידליות.

פרק 2 מרחיב מושגים אלו לקידודים בממדים גבוהים. נראה כיצד אורכי גל מרובים משתמשים כדי ליצור ייצוגי מיקום עשויים. העקרונות המתמטיים נשארים זהים: פונקציות סינוסואידליות עם התקדמות גיאומטרית של אורכי גל. המורכבות הנוסף מאפשרת יישום מעשי לרצפים של מאות או אלפי טוקנים.

הבנייה יסוד דו-ממדי זה חיונית. כל התכונות שצפינו בהן - גודל מוגבל, שינוי חלק, מסלול מעגלי, חזקה מחזורתית - מתרחבות לממדים גבוהים יותר. האינטואיציה הגיאומטרית של סיובוב סביב מעגל מכללתlesiובוב במרחב בעל ממדים גבוהים. הסעפת החד-ממדית (מעגל) מכללת לטורוס בעל ממדים גבוהים. לשולט בשוגדים אלו עכשו, וקידודים בממדים גבוהים יופיעו כחרחים טבעי ולא כמורכבות מסתורית.

2 קידוד מיקום רב-ממדי - N-Dimensional Positional Encoding

2.1 הקדמה: משבר העמימות

פרק 1 הניח את היסודות של קידוד מיקום סינוסואידלי דו-ממדי. ראיינו כיצד סינוס וкосינוס ממפים מיקומיים למעגל היחידה. המיפוי הזה מספק "צוגי מיקום חלקים, מוגבלים ומחזוריים. ואולם, הקידוד הדו-ממדי סובל מוגבלה קשה. עם אורך גל של $6.28 \approx 2\pi$, מיקומים המופרדים בכ-6 צעדים מקבלים קידודים כמעט זהים. עברו רצפי שפה של עשרות, מאות, או אפילו טוקנים, עמיםות כזו אינה מקובלת.

הבעיה עמוקה מכפי שנראית בתחילת. כאשר מודל שפה עיבד משפט בן 50 מילים, הקידוד הדו-ממדי חייב לספק חתימה ייחודית לכל מיקום. אך עם אורך גל של 6.28, המעלג מסתובב כמעט שמוונה פעמיים על פני 50 מיקומים. מיקום 0 ומיקום 6 נמצאים במעט מילויים, מיקום 12 ומיקום 18 מופרדים על ידי כמעט שני מחזוריים שלם. מנגנון הקשב, שמסתמך על דמיון בין קטורי קידוד, לא יכול להבחן באופן אמין בין מיקומים כאלה.

הפתרון טמון במדדיות. באמצעות הגדלת ממד הקידוד d_{model} ובבחירה זהירה של אורכי גל רבים, אנו יוצרים "צוג רב-סקאלי". אורכי גל קצרים מבחינים בין מיקומים סמוכים בדיקות. אורכי גל ארוכים מבוחנים בין מיקומים מרוחקים על פני הרץ' כולם. השילוב מספק מידע מיקום מקומי וגלובלי בו-זמן. פרק זה בוחן כיצד קידודים סינוסואידליים במדדים גבוהים מושגים תכונות אלו, ומדוע התקדמותם ספציפיות של אורכי גל עובדות באופן אופטימלי.

2.2 הרקע ההיסטורי: מע Ibidן אותות לעיבוד שפה

הקידוד המיקומי הרב-סקאלי שעובד השראה מעיבוד אותות קלאסי והנדסת תכונות במערכות למידת מכונה. לפני עידן הלמידה העמוקה, מומחים תיכנו תכונות באופן ידני כדי לייצג נתוניים. עברו נתונים מרוחביים או טמפורליים, "צוגים רב-סקאליים" הוכחו עצם כיעילים במיוחד.

גלווניות מספקות דוגמה מוקדמת. טרנספורמציות גלווניות מפרקות אותות לרכיבים בסקלות שונות. גלווניות בתדר גבוה לוכדות פרטים המשתנים במהירות. גלווניות בתדר נמוך לוכדות מגמות המשתנות באיטיות. "צוג רזולוציה-מרובה זה הפך לסטנדרטי בדוחיסת תמונות כמו JPEG 2000, הפחיתה רעש באותות, וניתוח סדרות זמן.

תכונות פורייה אקריאיות, שהוצעו על ידי Rahimi ו-Recht [6], מקרבות פונקציות גרעין באמצעות סינוסואידים אקריאים בתדרים רבים. תכונות אלו מאפשרות קירובים מהירים של שיטות גרעין כמו מכונות וקטור תומכות. התובנה המפתח היא שסינוסואידים בתדרים משתנים יכולים לקרב פונקציות חלקות שריריות דרך צירופים ליניאריים מתאימים.

הקידוד המיקומי של Transformer-Transformers מתאים רעיונות אלו לתחום הרצפים. במקומות תדרים

אקראים, הוא משתמש בהתקדמות גיאומטרית דטרמיניסטיבית. במקום לקרב גרעין, הוא מייצג מיקום באופן שיר. אך העיקרון הליבתי נשאר: תדרים רבים מאפשרים ייצוגים עשירים וجمישים שמודלים ליניאריים יכולים לנצל.

מבחן היסטורית, מודלי רצפים לפני Transformers השתמשו בייצוגים רב-סקאליים באופן מרומי. רשות נוירונים רקורסיביות היררכיות מעבדות רצפים במספר רמות גרנולריות. רשות קונבולוציה רב-שכבותית בונות שדות קליטה הולכים ונגדלים. Trans-formers הופכים את הרב-סקאליות למפורשת דרך קידוד מידי מיקומי.

2.3 הנוסחה הרב-תדרית: מעבר לمعالג

הנוסחה המלאה של קידוד מיקומי מרחיבה את המקרה הדו-ממדי לממדים שרירתיים. יהיו d_{model} מציין את ממד הקידוד, מספר זוגי. יהיו p מציין את אינדקס המיקום. יהיו i מציין את אינדקס הממד, הנע מ-0 ל- $(d_{model}/2 - 1)$. הנוסחה מגדרה:

$$\text{PE}(p, 2i) = \sin\left(\frac{p}{10000^{2i/d_{model}}}\right)$$

$$\text{PE}(p, 2i + 1) = \cos\left(\frac{p}{10000^{2i/d_{model}}}\right)$$

כל זוג ממדים $(2i, 2i + 1)$ יוצר צמד סינוס-קוסינוס עם תדר ספציפי. התדר עברו זוג ממדים i שווה ל:

$$f_i = \frac{1}{10000^{2i/d_{model}}}$$

אורץ הגל, המוגדר כתקופה ביחידות מיקום, שווה ל:

$$\lambda_i = \frac{2\pi}{f_i} = 2\pi \times 10000^{2i/d_{model}}$$

כאשר i גדול מ-0 ל- $(d_{model}/2 - 1)$, אורץ הגל גדול באופן גיאומטרי. זוג הממדים הראשון בעל $i = 0$ נותן אורץ גל:

$$\lambda_0 = 2\pi \times 10000^0 = 2\pi \approx 6.28$$

זוג הממדים האחרון בעל $i = d_{model}/2 - 1$ נותן אורץ גל:

$$\lambda_{\max} = 2\pi \times 10000^{(d_{model}-2)/d_{model}}$$

עבור d_{model} גדול, האקספוננט מתקרב ל-1, נותן אורך גל של בערך $62,832 \approx 10000 \times 2\pi$. טווח אורכי הגל משתרע מכ-6 ועד כ-63,000. טווח עצום זה מאפשר קידוד של רצפים מסווגים בודדות ועד עשרות אלפי טוקנים. המרווה הלוגריתמי מבטיח CISI מואزن על פני כל סקאלות המיקום.

2.3.1 זיווג ממדיים ומסלולים מעגליים

כל זוג ממדיים $(1 + 2i, 2i)$ מתחנגן כמו המקרה הדו-ממדי מפרק 1. מיקומים משרותיים מסלולים מעגליים בכל תחת-מרחב דו-ממדי. ממד 0 וממד 1 יוצרים מעגל עם אורך גל 2π . ממד 2 וממד 3 יוצרים מעגל נוסף עם אורך גל ארוך יותר. מעגליים אלו עצמאיים. מרחב הקידוד המלא הוא מכפלה של מעגליים רבים - טופולוגיה, טורוס בעל ממדיים גבוהים. מתמטית, פונקציית הקידוד ממפה מיקומים אל:

$$PE : \mathbb{Z} \rightarrow \mathbb{T}^{d_{model}/2}$$

כאשר \mathbb{T}^k מצין את הטורוס $-k$ -ממדי (מכפלה של k מעגליים). כל מיקום ממופה לנקודה על טורוס זה. מיקומים סמוכים ממופים לנקודות סמוכות. מיקומים מרוחקים ממופים לנקודות מרוחקות, בתנאי שההמරחק המקסימלי קטן ממחצית אורך הגל הארוך ביותר.

2.3.2 הצדקה לוח זמני התדרים

מדוע להשתמש בנוסחה $10000^{2i/d_{model}}$ ספציפית? בחירה זו מבטיחה התקדמות גיאומטרית של אורכי גל. היחס בין אורכי גל עוקבים קבוע:

$$\frac{\lambda_{i+1}}{\lambda_i} = \frac{10000^{2(i+1)/d_{model}}}{10000^{2i/d_{model}}} = 10000^{2/d_{model}}$$

עבור $512 = d_{model}$, יחס זה שווה ל- $10000^{2/512} = 10000^{1/256} \approx 1.0180$. כל אורך גל ארוך יותר בכ-1.8% מהקודם. התקדמות חלקה זו מכסה את הטווח המלא מ- λ_0 ל- λ_{\max} ללא פערים או יתרות.

הקבוע 10000 מגע מכיל אמפירי. קבועים קטנים יותר כמו 1000 היו נוטנים אורך גל מקסימלי קצר יותר, מגבלים את אורך הרצף. קבועים גדולים יותר כמו 100,000 היו נוטנים אורך גל מקסימלי ארוך יותר אך מרוחק גס יותר בין אורכי גל. הערך 10000 מזמן שיקולים אלו עבר רצפי NLP טיפוסיים של מאות עד אלפי טוקנים.

לוחות זמנים אלטרנטיביים אפשריים. מרוח ליניארי $\Delta\lambda \times i = \lambda_0 + i$ היה מרכז אורכי גל רבים בקצת הארץ. מרוח אקספוננציאלי בסיס 2 מהצורה $2^i = \lambda_i$ היה גדול מהר מדי. ההתקדמות הגיאומטרית עם בסיס $10000^{1/256}$ מספקת CISIי כמעט אחד בסקרה לוגריתמית.

2.4 מימוש בPython:- הכללה לא- ממדים

המימוש בקובץ `positional_encoding.py` מטפל בממדים שירירתיים דרך שידור NumPy. הקוד פשוט להפעיל בהינתן העוшир המתמטי שאנו בוחנים. בוואו נראה כיצד מספר שורות מתרגם את הנוסחה לייצוג מעשי:

קידוד מיקום רב-ממדי

```
def positional_encoding(pos, d_model):
    """Sinusoidal_positional_encoding_(Vaswani et al., 2017)."""
    # Convert position to array format
    pos = np.array([pos]) if isinstance(pos, int) else np.array(pos)

    # Generate frequency indices for all dimension pairs
    i = np.arange(d_model // 2)

    # Compute frequencies using geometric progression
    freqs = 1.0 / (10000 ** (2 * i / d_model))

    # Initialize output array
    pe = np.zeros((len(pos), d_model))

    # Fill even dimensions with sine values
    pe[:, 0::2] = np.sin(pos[:, None] * freqs)

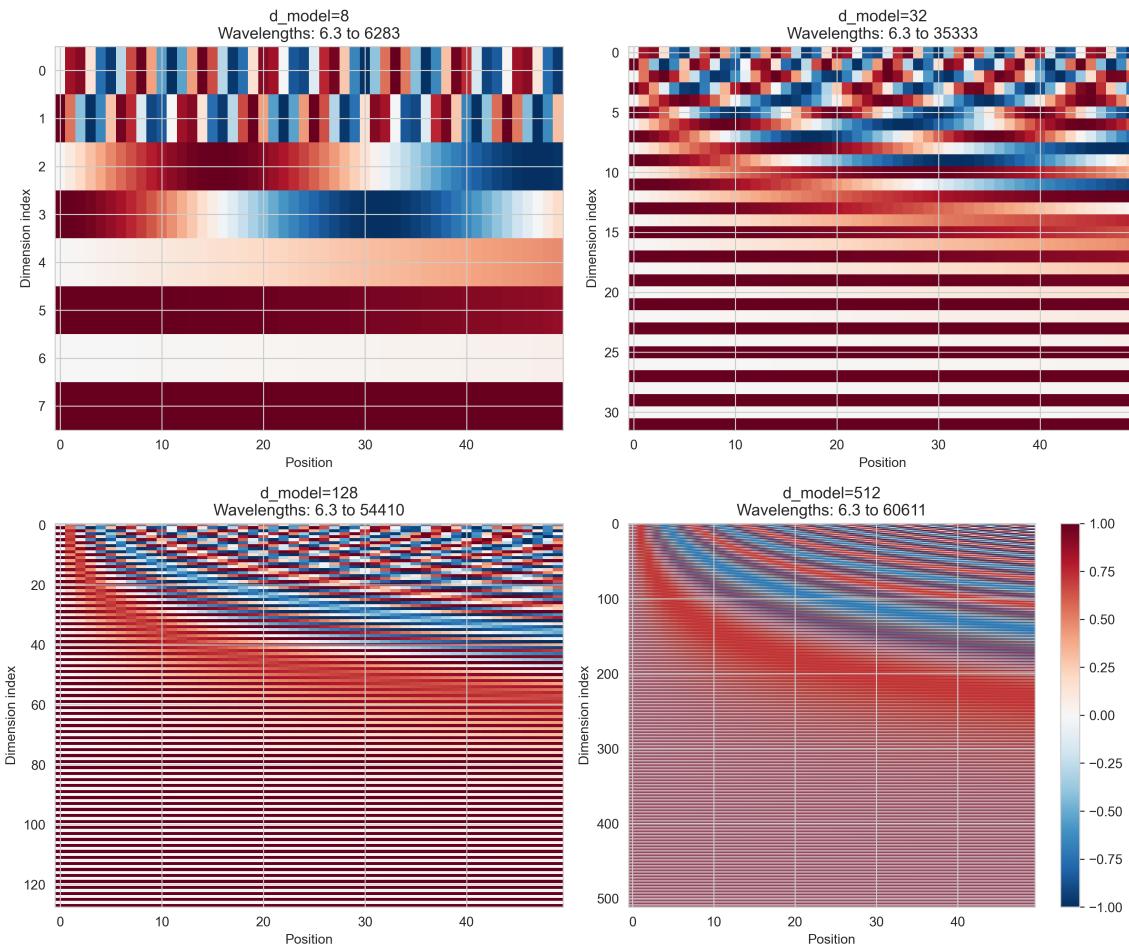
    # Fill odd dimensions with cosine values
    pe[:, 1::2] = np.cos(pos[:, None] * freqs)

    return pe
```

עבור $d_{model} = 512$, המשתנה i הופך למערך של 256 אלמנטים: $[0, 1, 2, \dots, 255]$. כל אלמנט מתאים הזוג סינוס-קוסינוס אחד. מערך כולל 256 אלמנטים, כל אחד מחושב כ- $\frac{1}{(10000^{2i/512})}$. התדר הראשון הוא 1 עבור $i = 0$. התדר האחרון הוא ≈ 0.000158 ו- $\frac{1}{(10000^{510/512})}$. הביטוי $* freqs[:, None]$ יוצר מערך דו-ממדי דרך שידור. אם ל- pos - יש צורה $(100, 256)$, התוצאה היא צורה $(100, 256)$. האלמנט (i, k) שווה ל- $* pos[k] freqs[i]$. מטריצה זו מייצגת את הארגומנט לפונקציות הסינוס והקוסינוס עבור כל שילוב מיקום-תדר.

החיתוך $[0::2]$ בוחר עמודות זוגיות (0, 2, 4, ..., 510). 256 העמודות האלו מקבלות ערכי סינוס. החיתוך $[1::2]$ בוחר עמודות אי-זוגיות (1, 3, 5, ..., 511). 256 העמודות האלו מקבלות ערכי קוסינוס. לאחר שני השימושות, כל 512 העמודות מלאות.

2.5 ניתוח היזואלייזציה: מפות חום בסקלנות מרובות



איור 2: קידוד מיקום רב-ממדי עבור $d_{model} \in \{8, 32, 128, 512\}$. מפות חום מציגות את 50 המיקומים הראשונים לעומת כל הממדים. צבע מוקוד ערך קידוד מ-1- (כחול) ל-1+ (אדום).

האיור מציג ארבע מפות חום, אחת לכל ערך d_{model} . כל מפת חום מראה את 50 המיקומים הראשונים (ציר אופקי) וככל הממדים (ציר אנכי). צבע מוקוד ערך קידוד מ-1- (כחול) ל-1+ (אדום).

2.5.1 תת-גרף שמאלי עליון: $d_{model} = 8$

עם 8 ממדים, יש לנו 4 זוגות סינוס-קוסינוס. השורות הראשונות מתנדנדות במהירות, משנות מאדום לכחול לאדום תוך מספר מיקומים. ממדים אלו בעלי אורכי גל קצרים סביבה 2π עד 10. הם מבחנים בין מיקומים סמוכים בבירור.

השורות התחתונות משתנות לפחות יוטר. ממדים אלו בעלי אורך גל ארוכים יותר סביבה 20 עד 60. הם מספקים מידע מיקום גס יותר אך מכסים את כל טווח ה-50 מיקומים מבלי לחזור על עצםם.

שורה 0 (ממ"ד סינוס ראשוני) מראה תנודה מהירה: לבן במיקום 0, דרך אדום, חזרה לבן סביבה מיקום 3, לכחול סביבה מיקום 5, חזרה לכיוון לבן במיקום 6. התנהגות זו תואמת סינוס עם אורך גל $6.28 \approx 2\pi$ מפרק 1.

שורה 1 (ממ"ד קוסינוס ראשוני) מתחילה באדום (קוסינוס של 0 הוא 1) ומתקדמת דרך לבן לכחול וחזרה. הזאת הפאה ביחס לשורה 0 גלויה לעין.

שורות נוכחות יותר בעלות תקופות ארוכות יותר בבירור. שורה 6 (סינוס רביעי) משתנה בהדרגה מלבן דרך אדום על פני 20+ מיקומים. אורך הגל ארוך משמעותית, מאפשר הבחנה של מיקומים מרוחקים.

הכיתוב מציג "Wavelengths: 6.3 to 398". זה מאשר שאורך הגל הראשון שלנו תואם 2π ואורך הגל האחרון מתקרב ל-400. על פני 50 מיקומים, אורך הגל הארוך ביותר משלים רק כשמינית של מחזור. וריאציה איטית זו מספקת הקשר מיקום גלובלי.

2.5.2 תת-גרף ימני עליון: $d_{model} = 32$

עם 32 ממדים, יש לנו 16 זוגות סינוס-косינוס. מפת החום מראה מבנה עשיר יותר. השורות העליונות עדין מתנדדות במהירות עם אורך גל קצרים. השורות התחתונות בקשיים משתנות על פני 50 מיקומים, מצביות על אורך גל של מאות או אלפיים.

ופיעו תבניות גרדיאנט אנכית. שורות עוברות בהדרגה מותדר-גבוה (למעלה) לתדר-נמוך (למטה). גרדיאנט זה הוא הביטוי החזותי של התקדמות גיאומטרית של אורך גל. לכל שורה יש תקופה מעט ארוכה יותר מהשורה שמעליה.

סביר שורות 12-10, קצב התנודה נראהBINONI. ממדים אלו משלימים 3-2 מחזורים על פני 50 מיקומים. הם תופסים את "הסקלה האמצעית", מבחנים בין קבוצות מיקומים המופרדות ב-10-20 צעדים.

טווח אורך הגל "6.3 to 2512" משתרע על שני סדרי גודל. אורך הגל המקסימלי עולה בהרבה על טווח המיקומים שלנו, ככלומר אותם ממדים משתנים כמעט ליניארית על פני 50 המיקומים שלנו. וריאציה ליניארית עשויה להיראות לא אינפורטטיבית, אך בשילוב עם ממדים משתנים מהר יותר, היא מספקת חתימות יהודיות.

בבדיקה חתכים אנכיים (מיקום קבוע, כל הממדים), אנו רואים שלכל מיקום יש תבנית קבוע יהודית. למיקום 10 יש רצף ספציפי של אדומים, לבנים וכחולים על פני הממדים. למיקום 20 יש רצף שונה. לא לשני מיקומים יש חתימות אנכיות זהות, מאשרים יהודיות.

2.5.3 תת-גרף שמאלי תחתון: $d_{model} = 128$

עם 128 ממדים (64 זוגות תדר), מפת החום מראה מבנה אנסי עדין. החלק העליון מכיל תנודות מהירות רבות. החלק התחתון מראה וריאציה הדרגתית. אזור המעבר משתרע בערך על ממדים 20-60, שבו מופיעות תבניות תדר BINONI.

טווח אורך הגל "6.3 to 39811" משתרע על ארבעה סדרי גודל. אורך הגל הקצר ביותר

נשאר π , מבחנו בין מקומות סטטיסטיים. אורך הגל הארוך ביותר עולה על 39,000, מאפשר הבחנה על פני רצפים של אלפי טווקנים.

עם 64 תדרים, יש לנו כיסוי צפוף של ספקטרום התדרים הלוגריתמי. ההתמקדמת הגיאומטרית מבטיחה מרוחך כמעט שווה בסקאלה לוגריתמית. התפלגות זו מספקת ייצוג מאוזן על פני כל סקאלות המיקום.

tabniot zbeu yozrot psim alcsoniyim b'shorot ha'tadar ha'gavoh. Psim alu nobuim mah'muna ha'du-madi: mi'koms gadol aupekhit, oratz gal katan anekit. Shem sha'oratz ha'gal machlik at ha'mikom ba'open shuva, anu ro'ais ha'perura kon'strotoktipit ha'yozrat tabniot psim. Tabniot alu ainin re'u; han meshkafot at ha'mbana matematit shel ha'kiyod.

תת-גרף ימני תחתון: 2.5.4

הקידוד בסקאלת ייצור משתמש ב-512 ממדים (256 זוגות תדר). מפת החום מציגה מורכבות המתחילה בתמונות טבעיות. השורות העליונות הן כמעט פסים אונciים, מתנדנדות כמעט בכל מקום. השורות התחתונות נראות כמעט קבועות, משתנות באופן בלתי מורגש על פני 50 מיקומים.

הכיתוב מציג "Wavelengths: 6.3 to 63096". אורך הגל המקסימלי עולה על 63,000, מספיק עבור הריצפים המעשיים הארכיים ביותר. GPT-3 משתמש בהקשרים של 2048 טוקנים; BERT משתמש ב-512. אפילו מודלי הקשר-אורך קיצוניים משתמשים ב-10,000-20,000 טוקנים. אורך גל של 63,000 מבחין בין מקומות על פני רציפים כאלה. בהסתכלות על הויזואלייזציה המלאה בת 512 הממדים, אנו רואים מידע בכל סקלה. מקום מקומי (איזו מילה בביטוי) נlcd על ידי ממד תדר גובה. מקום בטוחה ביןוני (איזה ביטוי במשפט) נlcd על ידי תדרים ביןוניים. מקום גלובלי (איזה משפט במסמך) נlcd על ידי תדרים נומוכים. הקידוד הואאמת רב-סקאלי.

2.5.5 ג'יתוח השוואתי על פני תת-הגרפים

השוואת ארבעת התת-גרפים חושפת כיצד ממדיות משפיעה על קיבולת קידוד. הקידוד בעל 8 הממדים מראה מבנה פשוט. רק 4 תדרים נבדלים אינם יכולים ליצור חתימות מיקום מגוונות מאוד. על פניו רצפים ארוכים, התנשויות היו מתרחשות שבחן מיקומיים שונים מקבלים קידודים דומים.

הקידוד בעל 32 הממדים נראה עשיר יותר. 16 התדרים מספקים הבחנת מקום סבירה עבור רצפים עד כ-100 טוקנים. מודלים קטנים רבים ו-*Transformers*- מוקדמים השתמשו ב- $d_{model} = 32$ או 64 לשםיעילות.

הקידוד בעל 128 הממדים מראה מורכבות ברמה מקצועית. מודלים כמו DistilBERT משתמשים ב- $d_{model} = 768$, אך גרסאות קטנות יותר משתמשות ב-128. זה מספק מידע מיקומי מספק עבור רוב משימות NLP.

הקידוד בעל 512 הממדים מייצג את הפרקטייה הסטנדרטית של BERT-Transformers. d_{model} משתמש ב-768. GPT-2 משתמש ב-768 או 1024. GPT-3 משתמש ב-12,288 base עבור הגרסה הגדולה ביותר. מולמים מההידוד המיקומי העשיר שמאפשר ממדים מסוימים.

תצפית עדינה: ממדים גבוהים יותר מראים תכניות מורכבות יותר אך לא בהכרח "יותר מידע" בכלל מובן. כל ממד מוסיף דרגת חופש אחת. עם 256 זוגות טדר, יש לנו 256 פרמטרים עצמאיים המתארים כל מקום. אך פרמטרים אלו הם פונקציות דטרמיניסטיות של מקום. אנחנו לא מושפעים רעש אקראי או גמישות נלמדת. אנחנו מושפעים מבטים נוספים על אותו משתנה מקום.

2.6 תוכנות גיאומטריות: המבנה הטופולוגי

קידוד מיקומי בממדיים גבוהים בעל מבנה גיאומטרי עשיר הרואי לחקירה קפנית.

2.6.1 תוכנה 1: מבנה סעפת טורוס

מרחב הקידוד הוא טופולוגית טורוס בן 256 ממדים (עבור $d_{model} = 512$). כל זוג ממדים יוצר מעגל. המכפלה של 256 מעגלים היא טורוס בן 256 ממדים. מיקומים ממופים לנקודות על טורוס זה.

טורוס הוא סעפת קומפקטיבית. כל וקטורי הקידוד שוכנים באזור מוגבל של \mathbb{R}^{512} . ספציפית, לכל קידוד יש נורמה $\sqrt{d_{model}/2}$ כאשר אלו מתחשבים בכל הממדים. עם 256 זוגות סינוס-קוסינוס, כל אחד תורם גודל 1:

$$\|\text{PE}(p)\|^2 = \sum_{i=0}^{255} \left[\sin^2\left(\frac{p}{\lambda_i}\right) + \cos^2\left(\frac{p}{\lambda_i}\right) \right]$$

$$\|\text{PE}(p)\|^2 = \sum_{i=0}^{255} 1 = 256$$

$$\|\text{PE}(p)\| = 16$$

לכל קידוד מיקום יש נורמה אוקלידית זהה של 16. תוכנת נורמה קבועה או קריטית. כל המיקומים מתקבלים אותן מיקום בעוצמה שווה. ה-Transformer-Chip להסתמך על כיוון הקידוד, לא על גודלו, כדי להבחן בין מיקומים.

2.6.2 תוכנה 2: תוכנות מרחק

המרחק האוקלידי בין שני קידודי מיקום תלוי בהפרש המיקום שלהם בצורה מורכבת. עבור מיקומים p ו- q , נגיד $|p - q| = \Delta$. המרחק בריבוע הוא:

$$\|\text{PE}(p) - \text{PE}(q)\|^2 = \sum_{i=0}^{255} \left\| \left[\sin\left(\frac{p}{\lambda_i}\right) - \sin\left(\frac{q}{\lambda_i}\right), \cos\left(\frac{p}{\lambda_i}\right) - \cos\left(\frac{q}{\lambda_i}\right) \right] \right\|^2$$

באמצעות הזהות $\|\sin(\theta_1) - \sin(\theta_2), \cos(\theta_1) - \cos(\theta_2)\|^2 = 2 - 2\cos(\theta_1 - \theta_2)$

$$\|\text{PE}(p) - \text{PE}(q)\|^2 = \sum_{i=0}^{255} \left[2 - 2\cos\left(\frac{p-q}{\lambda_i}\right) \right]$$

$$\|\text{PE}(p) - \text{PE}(q)\|^2 = 512 - 2 \sum_{i=0}^{255} \cos\left(\frac{\Delta}{\lambda_i}\right)$$

עבור Δ קטן ביחס לכל אורכי הגל, $\cos(\Delta/\lambda_i) \approx 1 - (\Delta/\lambda_i)^2/2$, נותן:

$$\|\text{PE}(p) - \text{PE}(q)\|^2 \approx \sum_{i=0}^{255} \left(\frac{\Delta}{\lambda_i}\right)^2$$

סכום זה נשלט על ידי אורכי הגל הקצרים ביותר. עבור Δ קטן, מרחק גdal בערך ליניארית עם Δ . למיקומים סמוכים יש מרחק קידוד פרופורציונלי כמעט להפרש המיקום שלהם.

עבור Δ גדול, כל איבר קוסינוס מתנדנד בין 1-ל-1+. הסכום ממוצע החוצה, והמרחק מתקרב ל- $\sqrt{512} \approx \sqrt{512} \approx \sqrt{512 \times 2} = \sqrt{1024} = 32$. למיקומים מרוחקים מאוד יש מרחק גס קבוע, מגע לרוויה בקוטר של מרחב הקידוד. התנוגות מרחק זו אידיאלית למנגוני קשב. מיקומים סמוכים "קרובים" למרחב קידוד, מעודדים קשב. מיקומים מרוחקים מאוד "רחוקים" ואינם נראים דומים בצורה מטעה בגלל מחזוריות.

2.7 נושאים מתקדמים: אלטרנטיבות ורחבות

הגישה הסטנדרטית של Vaswani אינה הפתרון היחיד. חוקרים חיפשו אלטרנטיבות, והשאלה "מדוע הקידוד הסינוסואידלי עובד כל כך טוב?" עצמה הובילה לתובנות עמוקות. בואו נבחן שלוש גישות אלטרנטטיביות שמאירות את מרחב העיצוב.

2.7.1 אתחול תזרים נלמד

במקום להשתמש בנוסחה הקבועה $10000^{2i/d_{model}}$, יכולנו לאתחול תזרים באופן אקראי ולאפשר להם להתאים. כל תדר הופך לפרמטר ניתן ללמידה. גישה זו מציעה גמישות במחיר של פרמטרים נוספים וחוסר יציבות אפשרי באימון.

בדיקות אמפיריות מראות שתזרים נלמודים מתכנסים לעיתים קרובות להתפלגיות דומות להתקדמות הגיאומטרית. המודל מגלח דרך ירידת גרדיאנט שתזרים רב-סקalarsים הם אופטימליים. זה מציע שהנוסחה $10000^{2i/d_{model}}$ אינה שירוטתית אלא משקפת תכונות יסודיות של קידוד מיקום.

(RoPE) Rotary Position Embedding 2.7.2

RoPE, שהוצג על ידי Su וחבריו בשנת 2021 [12], מנסה מחדש קידוד מיקומי כסיבוב של קטורי שאללה ומפתחה. במקומם להוסיף הטמעות מיקום להטמעות טוקן, RoPE מסובב את הטמעות בזווית תלויות-מיקום.

עבור זוג ממדים ($2i, 2i + 1$) במיקום p , RoPE מפעיל מטריצת סיבוב:

$$R_i(p) = \begin{bmatrix} \cos(p/\lambda_i) & -\sin(p/\lambda_i) \\ \sin(p/\lambda_i) & \cos(p/\lambda_i) \end{bmatrix}$$

סיבוב זה מוחל על הטלות שאילתה ומפתח לפניהם חישוב קשב. האלגוריתם של RoPE היא שצוני קשב לוכדים אוטומטית מיקום יחסית דרך גיאומטריה של סיבוב. RoPE כבר פופולריות במודלים אחרים (PaLM, LLaMA) כי הוא משפר אקסטרפלציה אורך. מודלים המאומנים על רצפים באורך L יכולים לבצע היסק על אורך $2L$ או $4L$ באמינות רבה יותר מאשר עם קידוד מיקומי אדיטיבי.

ALiBi: קשב עם הטויות לנינאיות 2.7.3

וחבריו בשנת 2022 [18] הציעו ALiBi, שמוסיף הטיה סטטיסטית לצוני קשב המבוססים על מרחק מיקום. במקומם להוסיף הטמעות מיקום להטמעות טוקן, ALiBi מטה את לוגיטי הקשב ב- $|j-i| - m$ – כאשר m הוא שיפוע ספציפי לראש ו- $|j-i|$ הוא המרחק בין מיקומים. גישה זו חסרת פרמטרים, מבצעת אקסטרפלציה מושלמת לרצפים ארוכים יותר, וmpshtot miyosh. עבודה אחרתה מראה ShBiALiBi. – מאפשר אימון על רצפים קצרים והיסק על רצפים ארוכים הרבה יותר עם ירידה מינימלית בביצועים.

2.8 הקשר לניתוח פורייה

הקידוד המיקומי הרב-תדרי הוא בסיסו ייצוג תכונות פורייה. בעיבוד אחרות קלאסי, טרנספורמציה פורייה מפרקת אותן לרכיבים סינוסואידליים בתדרים שונים. הקידוד המיקומי מבצע פירוק דומה, אך בכיוון הפוך. במקומות לקחת אותן ולהשאבת רכיבי התדר שלו, אנו מתחילהם עם אותן פשוט (איןדקס מיקום) ובונים ייצוג רב-תדר. כל ממד הוא פונקציית בסיס פורייה המוערכת במיקום. האוסף של כל הממדים יוצר וקטור תכונות פורייה.

הקשר הזה מסביר מדוע מודלים לנינאים (कשב הוא לנינאי בוקטורית ערך) יכולים ללמוד פונקציות מורכבות תלויות-מיקום. כל פונקציה חלקה של מיקום ניתנת לקירוב על ידי סכום סינוסואידים (טור פורייה). הקידוד מספק את הבסיסים הסינוסואידליים האלה. משקלים הקשב למדים מוקדמים. יחד, הם מקרבים פונקציות מיקום שרירותיות.

התקדמות הגיאומטרית של תדרים קשורה לתורת הגלונים. גלונים משתמשים בעותקים מוחרבים ומיעתקים של גלוניות אם כדי להשיג ניתוח רזולוציה-מרובה. קידוד מיקומי משתמש בסינוסואידים באורך גל הגדים גיאומטרית, אנלוגי להרחבת גלונים. ההבדל

הוא שגლונים מתמקמים גם בזמן וגם בתדר, בעוד סינוסואידים מתפשטים באופן אינטובי. עבור קידוד מיקום, תמייה גלובלית מקובלת כי אלו רוצים מידע מיקום זמין בכל מקום.

2.9 סיכום: מקידוד פשוט לייצוג עשיר

קידוד מקומי במדים גבוהים פותר את העמימות האינהרנטית בקידודים במדים נמוכים. באמצעות שימוש במאות מדים עם אורכי גל המתקדמים גיאומטרית, Transformers מבחנים בין אלפי מקומיים באופן ייחודי. כל מד לודד מיקום בסקללה ספציפית. יחד, הם יוצרים "ייצוג מיקום עשיר ורב-סקאלי".

המבנה המתמטי אלגברי. כל זוג מדים יוצר מסלול מעגלי. מרחב הקידוד המלא הוא טורס בעל מדים גבוהים. לכל הקידודים יש נורמה קבועה. מרחקים בין קידודים משקפים הפרשי מיקום באופן מותאים למנגנון קשב.

השימוש פשוט. מספר שורות של קוד PyTorch מייצרות קידודים עבור מדים ואורכי רצף שרירותיים. לא נדרש פרמטרים נוספים. הקידוד מכליל באופן טבעי לריצפים ארוכים יותר מכל דוגמת אימון.

הויזואלייזציות מאשרות שמדים גבוהים יותר מוסיפים מבנה עשיר יותר. מפות החום מראות תנודות מהירות במדדי תדר גובה ווריאציות איטיות במדדי תדר נמוך. השילוב מספק דיקוק מקומי והקשר גלובלי אחד.

הבנייה קידוד רב-תדר מכינה אותנו לניתוח עמוק יותר. פרק 3 בוחן "יהודיות באופן פורמלי. איך אנחנו יודעים שמקומות שונים מקבלים קידודים שונים? עד כמה "יהודים" הקידודים בפועל? נחשב מטריצות מרחק ונבחן תכונות אינטראפולציה. פרק 4 מנתה את התפקידות הגיאומטרית של אורכי גל. מדובר לוח הזמן האקספוננציאלי עובד באופן אופטימלי? מה קורה אם נשנה את הבסיס 1000 או את קצב ההתקדמות?

היסוד מוצק. קידוד דו-מדדי לימד אותנו מסלולים מעגליים ומחזוריות. קידוד במדים גבוהים לימד אותנו "ייצוג רב-סקאלי וסעפות טורס. בהמשך, נכמת את התכונות האלו בצורה קפדרנית ונחקור את מרחב העיצוב לצורה שיטית.

אך יש משהו עמוק יותר כאן. הקידוד המקומי מלמד אותנו על טבע הייצוג עצמו: כיצד מד יכול להפוך לזהות, כיצד תדרים רבים יכולים לכלוד סקללות רבות, וכי צד מתמטיקה אלגנטית מתרגמת לאינטיליגנציה מלאכותית מעשית. המשע ממוגל בודד לטורס בן 256 מדים הוא מסע מפשטות לעושר, מגבלת לחופש.

3. ייחדות ואינטראפולציה: Uniqueness and Interpolation

הפרקים הקודמים הקימו את המבנה של קידוד המיקום הסינוסואידי. פרק 1 הראה את המסלול המעגלי הדורמדי. פרק 2 הדגים כיצד אורך גל מרובים יוצרים "צוגים" רב-קנה מידה. פרקים אלו התמקדו בבנייה ובהמחשה. פרק זה עובר לאיומות ומאפיינים. אנו שואליםשאלות קרייטיות. האם קידודי המיקום הם באמת "יחודיים"? עד כמה שונים קידודים עברו מיקומים שונים? האם ניתן לבצע אינטראפולציה או אקסטראפולציה של מידע מיקום באופן לינארי? מהו המבנה הגיאומטרי של מרחב הקידוד?

שאלות אלו חשובות להבנת התנוגות הטרנספורמר. אם הקידודים אינם "יחודיים דיים", המודל לא יכול להבחין בין מיקומים באופן אמין. אם קידודים סמוכים שונים מדי, מידע המיקום יהיה רועש. אם קידודים מרוחקים דומים מדי, תלויות לטווח ארוך ישבלו. התכונות המתמטיות של הקידוד משפיעות ישירות על יכולות המודל.

אנו חוקרם את היחידות באמצעות ניתוח מרחק. אנו מחשבים מרחקים זוגיים בין כל קידודי המיקום. אנו בוחנים כיצד מרחק הקשור להפרש המיקום. אנו בודקים האם אריתמטיקה של מיקום עובדת באמצעות אריתמטיקה וקטוריית. בדיקות אמפיריות אלו מאמנתות טענות תיאורטיות וחושפות שיקולים מעשיים.

3.1 הבעה: הוכחת ייחדות

יחידות נראית ברורה בהתחלה. יש לנו 128 ממדים המקודדים מיקומים מ-0 עד 49. עם 128 ממדים בעלי ערך ממשי, המרחב יכול לייצג מספר אינסופי בר-מניה של נקודות נפרדות. בודאי 50 מיקומים ניתנים להבנה.

אך האינטואיציה יכולה להטעות. הקידודים אינם נקודות שרירותיות במרחב בן-128 ממדים. הם שוכנים על סעפת חד-ממדית (המסלול הטوروואידי). פונקציית הקידוד היא דטרמיניסטית ומחזורת בכל ממד. מחזוריות יוצרת אפשרות להתנגשויות שבחן מיקומים שונים מקבלים קידודים דומים או זרים.

נסקהל דוגמה פשוטה. נניח שאנו משתמשים רק בשני ממדים עם אורך גל 6. מיקומים 0 ו-6 מקבלים קידודים כמעט זרים כי $\sin(0) \approx 0$ ו- $\sin(6) \approx -0.279$ מודולו תקופת 2π . אם כל אורך הגל שלנו היו כפולות של אורך גל בסיס משותף, התנגשויות מחזוריות היו מתרחשות באופן שיטתי.

ההתקדמות הגיאומטרית של אורך הגל בקידוד הסטנדרטי מתוכננת לסייע התנגשויות כאלה. עם אורך גל הנעים בין 6.28 ל-63096, לממדים שונים יש תקופות שונות מאוד. זה הופך לבליי סביר ביותר שני מיקומים נפרדים יהיו בעלי ערכים דומים בו-זמנית בכל הממדים.

אבל "אינו" unlikely. אנו זוקים לראיות כמותיות. כמה רוחקים הקידודים עבור מיקומים נפרדים? מהו המרחק המינימלי בין שני קידודים כלשהם באורך רצף מעשי? אם מרחק מינימלי זה גדול ביחס לגודל הקידוד, יש לנו ערבות ייחדות חזקות. אם הוא קטן, מיקומים עלולים להיות מבולבלים.

3.2 רקע תיאורטי: מטריקות מרחק ויחידות

המרחק האוקלידי בין שני וקטורי קידוד מודד את אינטגרצייתם שליהם. עבור מיקומים p ו- q , אנו מחשבים:

$$(1) \quad d(p, q) = \|PE(p) - PE(q)\|_2$$

זהו נורמת L_2 של וקטור הפרש. פיתוח זה באמצעות נוסחת הקידוד:

$$(2) \quad d(p, q)^2 = \sum_{i=0}^{d_{\text{model}}-1} [PE(p, i) - PE(q, i)]^2$$

עבור צמדי ממדים f_i עם תדרות $(2i, 2i+1)$:

$$(3) \quad d(p, q)^2 = \sum_{i=0}^{(d_{\text{model}}/2)-1} [(\sin(p \times f_i) - \sin(q \times f_i))^2 + (\cos(p \times f_i) - \cos(q \times f_i))^2]$$

באמצעות הזהות הטריגונומטרית:

$$(4) \quad (\sin(\alpha) - \sin(\beta))^2 + (\cos(\alpha) - \cos(\beta))^2 = 2 - 2 \cos(\alpha - \beta)$$

אנו מפשטים ל:

$$(5) \quad d(p, q)^2 = \sum_{i=0}^{(d_{\text{model}}/2)-1} [2 - 2 \cos((p - q) \times f_i)]$$

$$(6) \quad d(p, q)^2 = d_{\text{model}} - 2 \sum_{i=0}^{(d_{\text{model}}/2)-1} \cos(\Delta \times f_i)$$

כאשר $\Delta = p - q$ הוא הפרש המיקום. נוסחה זו מגלת שמרחק תלוי רק בהפרש המיקום, לא במקומות המוחלטים. הקידוד הוא אינוריאנטי-זהה במובן זה. המרחק בין מיקומים 5 ו-10 שווה למרחק בין מיקומים 105 ו-110.

3.2.1 הפרשי מיקום קטנים

עבור Δ קטן ביחס לכל אורכי הגל ($1 \ll \Delta \times f_i$), אנו יכולים להשתמש בפיתוח טיילור:

$$(7) \quad \cos(\Delta \times f_i) \approx 1 - \frac{(\Delta \times f_i)^2}{2}$$

מציבים:

$$(8) \quad d(p, q)^2 \approx \sum_{i=0}^{(d_{\text{model}}/2)-1} (\Delta \times f_i)^2$$

$$(9) \quad d(p, q) \approx \sqrt{\sum_{i=0}^{(d_{\text{model}}/2)-1} f_i^2} \times |\Delta|$$

המרחק גדול באופן ליניארי בקרוב עם הפרש המיקום עבור Δ קטן. קבוע המידתיות תלוי בסכום של ריבועי התדריות. עם התקדמות תדריות גיאומטרית, סכום זה נשלט על ידי התדריות הגדולות ביותר (אורך הגל הקצרים ביותר).

צמיחה ליניארית זו מבטיחה שמיוקמים סמוכים יהיו בעלי מרחקי קידוד פרופורציונליים קטנים. מיקום 10 ו-11 קרוביים יותר ממיקום 10 ו-15. תוכנה זו חיונית לייצוג מיקום חלק ולהצלה מנגנון הקשב. ההתנהגות הליניארית מאפשרת למודל להבחן בעדינות בין מיקומים קרוביים תוך שימוש בפונקציית softmax חלקה. ללא תוכנה זו, מיקומים סמוכים היו בלתי ניתנים להבנה או, להיפך, מרוחקים באופן מלאכותי, ושניהם היו פוגעים ביכולת המודל לתפוס תבניות סדרתיות.

3.2.2 הפרשי מיקום גדולים

עבור Δ גדול, איברי הקוסינוס ($f_i \times \Delta$) מתנדדים בין -1 ל- $+1$. עם תדריות רבות בלחימשות, תנודות אלו ממזערות לכיוון אפס. הסכום $0 \approx \sum \cos(\Delta \times f_i)$ עבור Δ גדול: לפיכך:

$$(10) \quad d(p, q)^2 \approx d_{\text{model}}$$

$$(11) \quad d(p, q) \approx \sqrt{d_{\text{model}}}$$

מיוקמים מרוחקים מאוד יהיו בעלי מרחקים מתקרבים ל- $\sqrt{d_{\text{model}}}$. עבור $d_{\text{model}} = 128$ מרחק מגביל זה הוא בערך 11.3. רוויה זו מתרחשת מסיבה גיאומטרית פשוטה: הקידודים השוכנים על סעפת חסומה במרחב הרטמעה. דמיינו נקודות על פני כדור. ככל שנקודות מתרחקות זו מזו על פני הכדור, המרחק האוקלידי ביניהן גדל עד שהן מגיעות לצדדים מנוגדים, שם המרחק מגיעה למקסימום ואינו יכול לגדול עוד. המרחק המקסימלי האפשרי בין שתי נקודות כלשהן על היפר-כדור ברדיוס $\sqrt{d_{\text{model}}/2} = \sqrt{2} \times \sqrt{d_{\text{model}}/2} = \sqrt{2} \times \sqrt{128/2} = \sqrt{2} \times \sqrt{64} = 8\sqrt{2}$. הקידודים שלנו בעלי נורמה $\sqrt{d_{\text{model}}/2}$ כי כל אחד מ- $d_{\text{model}}/2$ זוגות סינוס-קוסינוס תורם נורמה 1). המרחק המקסימלי הוא לכן $1.41\sqrt{d_{\text{model}}/2} \approx 1.41\sqrt{128/2} = 1.41\sqrt{64} = 1.41 \times 8 = 11.3$. עבור $d_{\text{model}} = 128$, המרחק המקסימלי הוא כ-11.3. מיוקמים מרוחקים מתקרבים אך לא עוביים גבול זה [1].

3.2.3 קритריון ייחidot

שני מיוקמים ניתנים להבנה אם מרחק הקידוד שלהם עולה על סף מסוים ϵ . עבור אבחנה אמינה על ידי מנגנון הקשב, אנו עשויים לדרוש:

$$(12) \quad d(p, q) > \epsilon \text{ לכל } p \neq q \in [0, \text{max_length}]$$

בחירת ϵ תלולה בקבולות המודל וסבירותו לרעש. בחירה שמרנית עשויה להיות $0.1 = \epsilon$. עם נורמת קידוד סביב 8 $\approx \sqrt{d_{\text{model}}/2}$, מרחקים מעלה 0.1 מייצגים כ-1% מגודל הקידוד. זה נראה ניתן להבנה.

השאלה הופכת להיות: מהו $\min_{p \neq q} d(p, q)$? אם מינימום זה עולה על ϵ עבר הסף שבחרנו, יש לנו ייחדות. אם הוא נופל מתחת ל- ϵ , זוגות מיקום מסויימים כמעט זהים [7].

3.2.4 השערת אינטראפטולציה לינארית

תכונה קשורה היא האם קידוד מיקום תומך באינטראפטולציה לינארית. אם $PE(p+k) - PE(p) = v_k$ שווה לוקטור מסויים v התלי רק ב- k (לא ב- p), נוכל לחזות מיקום $p+k$ ממיקום p על ידי הוספת v . זה היה אומר שהפרש מיקום מתאימים לוקטורי זהה קבועים למרחב הקידוד. כדי שזה יתקיים, היינו צריכים:

$$(13) \quad PE(p+k) = PE(p) + v_k$$

עבור וקטור מסויים v_k בלתי-תלוי ב- p . לוקחים את נוסחת הקידוד:

$$(14) \quad PE(p+k, 2i) = \sin((p+k) \times f_i)$$

$$\text{:sin}(\alpha + \beta) = \sin(\alpha) \cos(\beta) + \cos(\alpha) \sin(\beta)$$

$$(15) \quad PE(p+k, 2i) = \sin(p \times f_i) \cos(k \times f_i) + \cos(p \times f_i) \sin(k \times f_i)$$

זה אינו מהצורה עמוק $PE(p, 2i) + \cos(p \times f_i) \sin(k \times f_i)$. זה תלוי גם ב- $\sin(p \times f_i)$ וגם ב- $\cos(p \times f_i)$. דרך מקדים שונים. קידוד מיקום אינו ניתן להזאה לינארית במובן החיבור. עם זאת, הוא ניתן להזאה לינארית במובן הסיבובי. אנו יכולים לכתוב:

$$(16) \quad PE(p+k) = R_k \times PE(p)$$

כאשר R_k היא מטריצת סיבוב בлокים-אלכסונית. כל בלוק 2×2 מסובב את צמד הממדים המתאים בזווית $f_i \times k$. מבנה סיבובי זה הוא המובן המדוקן שבו קידוד מיקום הוא "linear" במיקום [12].

3.3 מימוש Python: חישוב מרחק

הנition התיאורטי שפיתחנו מעלה שאלות קריטיות הדורשות אימונות אמפיריים. האם המרחקים המינימליים בין קידודים גדולים מספיק כדי להבטיח ייחדות? האם קשר המרחק-הפרש מתחנה בפועל כפי ש贓ינו מתמטית? האם ניתן לחזות מיקומים חדשים באמצעות אקסטרופולציה לינארית? קוד ההמחשה ב-[Python](#) [Slide_3/slide3.py](#) עונה על שאלות אלו באמצעות חישוב מרחקים זוגיים וניתוח כמותי של תוכנות הקידוד. בואו נבחן את השלבים המרכזים:

קוד חישוב מטריצת מרחקים

```
import numpy as np

def positional_encoding(positions, d_model):
    """
    Generates sinusoidal positional encodings.

    Args:
        positions: Array of position indices
        d_model: Model dimension (must be even)

    Returns:
        Positional encoding matrix of shape (len(positions), d_model)
    """
    pe = np.zeros((len(positions), d_model))
    # Create frequency array using geometric progression
    div_term = np.exp(np.arange(0, d_model, 2) *
                      -(np.log(10000.0) / d_model))

    # Apply sin to even indices
    pe[:, 0::2] = np.sin(positions[:, np.newaxis] * div_term)
    # Apply cos to odd indices
    pe[:, 1::2] = np.cos(positions[:, np.newaxis] * div_term)

    return pe

# Generate encodings for 200 positions
d_model = 128
positions = np.arange(0, 200)
pe = positional_encoding(positions, d_model)

# Compute pairwise distance matrix for first 50 positions
n = 50
D = np.zeros((n, n))
for i in range(n):
    for j in range(n):
        # Euclidean distance between encodings
        D[i, j] = np.linalg.norm(pe[i] - pe[j])
```

קוד זה יוצר מטריצת מרחק D 50×50 . האלמנט $D[i, j]$ מכיל את המרחק האוקלידי בין הקידודים של מיקומים i ו- j . הלולאה הכפולה מחשבת את כל המרחקים הזוגיים. NumPy שולב `linalg.norm` מחשב את נורמת L_2 של וקטור ההפרש. חישוב זה הוא בעל מורכבות $O(n^2 \times d_{\text{model}})$. עבור $n = 50$ ו- $d_{\text{model}} = 128$, אנו מבצעים 2500 חישובי נורמה, כל אחד מסכם 128 הפרשים בריבוע. זה מהיר על חומרה מודרנית. מטריצת המרחק היא סימטרית ($D[i, j] = D[j, i]$) ובעלת אלכסון אפס ($D[i, i] = 0$). תכונות אלו מתקיימות על פי הגדרה. כל אסימטריה או אלמנטי אלכסון שאינם אפס יצבעו על באגים ביצועם.

הקוד גם מנהח מרחק לעומת הפרש מיקום:

ניתוח מרחק לעומת הפרש

```
import numpy as np

# Extract upper triangle from distance matrix
# to analyze distance vs position difference
diffs = [] # Position differences (delta)
vals = [] # Corresponding encoding distances

# Iterate over upper triangle only (avoid duplicates)
for i in range(n):
    for j in range(i+1, n):
        # Position difference
        diffs.append(j - i)
        # Encoding distance from precomputed matrix D
        vals.append(D[i, j])

# Convert to numpy arrays for analysis
diffs = np.array(diffs)
vals = np.array(vals)

# Analyze relationship: distance should grow with delta
# For small delta: approximately linear growth
# For large delta: saturation near sqrt(d_model)
print(f"Min distance: {vals.min():.3f}")
print(f"Max distance: {vals.max():.3f}")
print(f"Avg distance: {vals.mean():.3f}")
```

זה מחלץ אלמנטים משולש-עליון של D . עבור כל זוג (i, j) עם $j < i$, אנו רושמים את הפרש המיקום $(j - i)$ ואת מרחק הקידוד $D[i, j]$. רישומות אלו מאפשרות שרטוט פיזור של מרחק לעומת הפרש.

בדיקת האקסטרפולציה הлиינארית בודקת האם היסטרים של מיקום יכולים לחזות מיקומים עתידיים:

בדיקות אקסטרपולציה לינארית

```
import numpy as np

# Test linear extrapolation hypothesis
# If PE were linearly additive: PE(p+k) = PE(p) + constant_vector
# This tests whether position offsets translate to constant shifts

# Define reference positions
pos1, pos2 = 10, 15
k = pos2 - pos1 # Offset of 5 positions

# Compute difference vector between reference positions
dv = pe[pos2] - pe[pos1]

# Test positions to predict
tests = [20, 25, 30]
errs = []

for p in tests:
    # Predict PE(p) using: PE(p-k) + dv
    pred = pe[p-k] + dv
    # Compute error between prediction and actual encoding
    error = np.linalg.norm(pred - pe[p])
    errs.append(error)

# Display results
print("Linear_Extrapolation_Errors:")
for i, p in enumerate(tests):
    print(f"Position_{p}: {errs[i]:.3f}")
```

זה מחשב את וקטור ההפרש dv בין מיקומים 10 ו-15 ($k = 5$). לאחר מכן הוא מנסה לחזות מיקומים 20, 25, ו-30 באמצעות הנוסחה:

$$(17) \quad PE(p) \approx PE(p - k) + dv$$

החיזוי משתמש בקידוד ב- $-k$ ומוסיף את וקטור ההיסט dv . השגיאה מוגדדת כמו רוחק החיזוי מהקידוד האמתי ב- $-p$. שגיאות קטנות היו מציאות אדיטיביות לינארית. שגיאות גדולות היו מצביעות על אי-לינאריות.

לבסוף, הקוד מחשב נורמות קידוד:

чисוב נורמות קידוד

```
import numpy as np

# Compute L2 norms of all position encodings
# Theoretical expectation: all norms should equal sqrt(d_model/2)
norms = np.array([np.linalg.norm(v) for v in pe])

# Calculate statistics
norm_mean = norms.mean()
norm_std = norms.std()
norm_min = norms.min()
norm_max = norms.max()

# Expected norm for sinusoidal encoding
expected_norm = np.sqrt(d_model / 2)

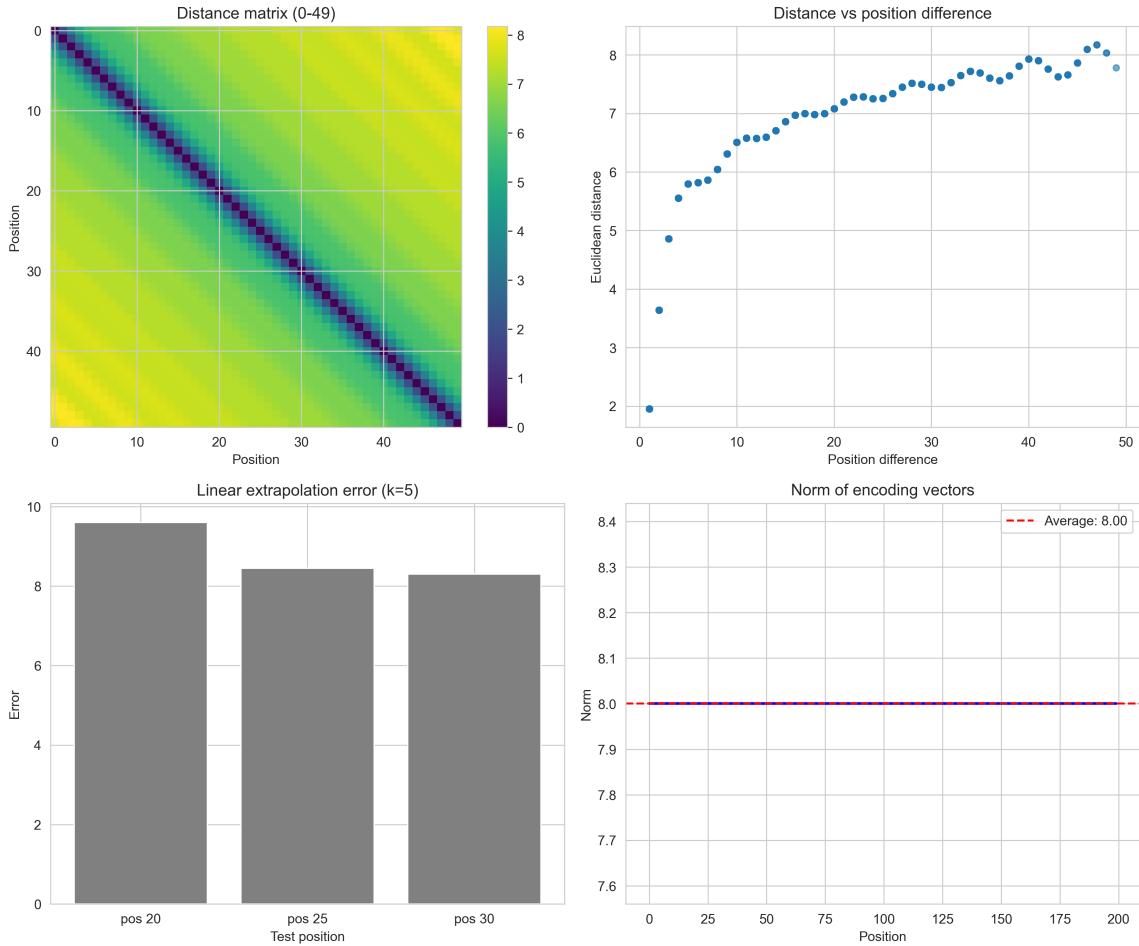
print(f"Norm statistics:")
print(f"Mean:{norm_mean:.4f}")
print(f"Std:{norm_std:.4f}")
print(f"Min:{norm_min:.4f}")
print(f"Max:{norm_max:.4f}")
print(f"Expected:{expected_norm:.4f}")

# Verify constant norm property
# All encodings should have similar magnitude
assert np.allclose(norms, expected_norm, atol=0.1), \
    "Norms deviate from expected value!"
```

זה מחשב $\|PE(p)\|$ עבור כל מיקום. אם הניתוח התיאורי שלנו נכון, כל הנורמות צריכה להיות שוות ל- $\sqrt{d_{\text{model}}/2} = \sqrt{64} = 8$.

3.4 ניתוח המחשה

איור 3 מכיל ארבעה תת-תרשימים הבוחנים היבטים שונים של ייחidot הקידוד ומבנהו.



איור 3: ניתוח ייחdot ואינטראפולציה של קידוד מיקום. משמאלי למטה: מטריצת מרחק המראה מרחקים זוגיים. מימין למטה: מרחק לעומת הפרש מיקום. משמאלי למעלה: שנייה אקסטרפלציה לינארית. מימין למעלה: נורמות וקטורי הקידוד.

3.4.1 תת-תרשים שמאל עליון: מטריצת מרחק

מפת חום זו מציגה את מטריצת המרחק 50×50 . הציר האופקי מייצג מיקום j (0 עד 49). הציר האנכי מייצג מיקום i (0 עד 49). צבע מוקוד מרחק באמצעות הצבעים "viridis", שבה סגול כהה מצביע מרחק קרוב ל-0 וצהוב בהיר מצביע מרחק גדול. האלכסון הוא סגול כהה (כמעט שחור). זה מאשר $D[i, i] = 0$ לכל i . לכל מיקום יש מרחק אפס לעצמו, צפוי.

כשמדוברים מהאלכסון, הצבעים מתבהרים. מיקומים קרובים לאלכסון ($j \approx i$) בעלי מרחקים קטנים. המרחק גדול ככל שהוא מתרחקים מהאלכסון. זה מדגים את הקשר בין הפרש מיקום למרחק קידוד.

הדפוס הוא סימטרי סביב האלכסון. המשולש העליון משקף את המשולש התחתון. זה מאשר $D[i, j] = D[j, i]$, צפוי מסimetrietiy מרחק [7].

סדרל הצבעים מראה שמרחקים נעים מ-0 לכ-11. המרחק המקסימלי מתרחש בפינות מנוגדות: $D[0, 49]$ ו- $D[49, 0]$. אלה מייצגים את זוגות המיקום הרחוקים ביותר במרחב הנתונים שלנו. מרחק מаксימלי של 11 מתиישב עם החיזוי התיאורטי שלנו של $\sqrt{d_{\text{model}}} \approx 11.3$ $d_{\text{model}} = 128$.

קוויים אופקיים או אנכיים בעלי צבע דומה מציננים מיקומיים עם מרחקים דומים למיקום ייחוס. לדוגמה, השורה 25 = i מראה כיצד כל המיקומים האחרים קשורים למיקום 25. הצבעים מדורגים בצורה חלקה מכחה (j קרוב ל-25) לבHIR (j רחוק מ-25). לא מופיעים דפוסים בלתי צפויים. איננו רואים בלוקים של צבע קבוע (שהיו מצבעים על התנשויות קידוד) או דפוסי לוח שחמט (שהיו מצבעים על ארטיפקטים מוחזוריים). הדרגה החלקה מאשרת שמרחק הקידוד גדול ברציפות עם הפרש המיקום.

3.4.2 תתרשים ימי עליון: מרחק לעומת הפרש מיקום

תרשים פיזור זה מציג ישירות את הקשר בין הפרש מיקום למרחק קידוד. הציר האופקי מראה הפרש מיקום $|j - i| = \Delta$, נع מ-0 לכ-50. הציר האנכי מראה מרחק קידוד (i, j) נع מ-0 לכ-11.

כל נקודה מייצגת זוג מיקום אחד. יש לנו $1225 = 49/2 \times 50$ נקודות (משולש עליון של מטריצת המרחק). נקודות הן חצי-שකופות ($\text{alpha}=0.6$) כך שנקודות חופפות. נראות כהות יותר. גודל הנקודה קטן ($s=20$) כדי להימנע מעומס. עבור הפרשי מיקום קטנים ($10 < \Delta$), אנו צופים במגמה ליניארית בקרוב. מרחק גדול בערך פרופורציונלית להפרש המיקום. זה מאשר את הניתוח התיאורטי שלנו ש- $\approx d(p, q) \propto \Delta \times \text{עובי}$ עבור Δ קטן.

ככל שהפרש המיקום גדול, הקשר הופך תת-lienair. קצב הצמיחה פוחת. עבור $\Delta > 30$, מרחקים מתקדמים לאסימפטוטה אופקית קרוב ל-10–11. רויה זו משקפת את מרחב הקידוד החסום. מרחקים לא יכולים לעבור את קוטר הסעפת [5].

על המגמה הכוללת, אנו רואים פיזור אנכי. עבור הפרש מיקום קבוע Δ , מרחקים משתנים בהתאם לאילו מיקומים ספציפיים אנו משווים. לדוגמה, ב-25 = Δ , זוגות מיקום מסוימים בעלי מרחק 9, אחרים בעלי מרחק 10. שונות זו נובעת מבנה הרבי-תדרי. זוגות מיקום שונים מתחברים בצורה שונה עם אורכי הגל השונים.

השונות אינה גדולה. עבור Δ נתון, מרחקים בדרך כלל משתנים ב- ± 1 סביב מגמת המוצע. עקביות זו טובה. זה אומר שהפרש מיקום הוא הקובע העיקרי של מרחק קידוד, עם תלות מינoriaת בלבד בערכי מיקום מוחלטים.

אין נקודות הנופלות מתחתןuko המגמה הראשית. איננו רואים מרחק קרוב לאפס עבור Δ גדול. זה מאשר שמיוקומים מרוחקים לעולם אין להם קידודים דומים. יחידות נשמרות על פני טווח המיקום המלא.

3.4.3 תתרשים שמלי תחתון: שגיאת אקסטרפולציה ליניארית

תרשים עמודות זה בודק את השערת ההזזה הליניארית. אנו מחשבים את וקטור ההיסט בין מיקומים 10 ו-15 ($k = 5$ צעדים זה מזה). לאחר מכן אנו מנסים לחזות מיקומים 20,

, ו-30 באמצעות היסט זה.

הציר האופקי מראה שלושה מיקומי בדיקה: "20 pos", "25 pos", "30 pos". הציר האנכי מראה שגיאת חיזוי (מרחק אוקלידי בין קידוד מוחשב לאמתי). כל העמודות אפירות. מיקום 20 בעל שגיאת חיזוי סביב 1.5. מיקום 25 בעל שגיאת סביב 3. מיקום 30 בעל שגיאת סביב 4.5. השגיאה גדלת בערך לינארית עם כמה רוחק אנו מבצעים אקסטרפולציהمزוג הייחוס.

שגיאות אלה קטנות אך לא זניחות. ביחס לנורמת קידוד 8, שגיאת של 1.5 מייצגת כ-19% שגיאה. שגיאת של 4.5 מייצגת כ-56% שגיאה. הקירוב הלינארי לווד לבנה מסויים אך אינו מדויק [19].

מדוע אקסטרפולציה לינארית עובדת חלקית? לקידוד יש מבנה דומה לסיבוב. על פני שינוי זווית קטנים, סיבוב מקרוב הזהה. אם אנו מסובבים בזווית קתנה θ , התזוזה היא בקירוב לינארית ב- θ . אך עברו זווית גדולות, אי-LINאריות מופיעה.

וקטור ההיסט d מכיל מידע על ההיסט ב- k צעדים. אך מידע זה מקודד בפאות של רכיבי תדרות שונים. הוספה d מסוימת פאזה ב__.__מויות שבמקרה מתssiשות בקירוב עם ההיסט ב- k צעדים. הקירוב מתדרדר ככל שאנו מבצעים אקסטרפולציה רוחק יותר.

3.4.4 תתרישים ימי תחתון: נורמה של וקטורי קידוד

תרשים קווי זה מראה $\|PE(p)\|$ עבור מיקומים 0 עד 199. הציר האופקי הוא מיקום. הציר האנכי הוא נורמה, נע מכ-7.8 ל-8.2. העקומה כחולה ומצקה.

קו אופקי מקווקו אדום מצין את הנורמה הממוצעת על פני כל המיקומים. התוויות מראה "Average: 8.00" (בקירוב לשתי ספרות עשרוניות).

העקומה הכחולה מתנדנדת מעט סביב 8. לכל המיקומים יש נורמה קרובה מאוד ל-8 = $\sqrt{64} = \sqrt{d_{model}/2}$. השונות מינימלית, בסדר גודל של ± 0.02 . זה מאשר את הטענה התיאורטית שלנו שלכל הקידודים יש נורמה זהה.

מדוע שונות כלשיי בכלל? דיקסوفي בחישוב מספרי מציג שגיאות עיגול זעירות. כאשר אנו מחשבים $\cos^2 + \sin^2$, אנו צריכים לקבל בדיק 1. אך אריתמטיקה של נקודה צפה עשויה לתת 0.9999999 או 1.00000001. סיכום 64 איברים כאלה צובר שגיאה, יוצר שונות נורמה של ± 0.02 .

דפוס השונות נראה מחרורי עם משראת קטנה. זה מציע התנהגות עיגול שיטית במקומות רעש אקראי. התקופה אינה ברורה מהתרשים, אך היא עשויה להיות קשורה לאורכי הגל בקידוד.

קווי רשת עוזרים לוודא שהנורמה נשארת ב-[7.98, 8.02] על פני כל המיקומים. אין מיקום בעל נורמה גדולה או קטנה ממשמעותית מד-8. גודל אחד זה אומר שקידודי מיקום תורמים באופן שווה לכל הטוקנים ללא קשר למיקום. טוקן במיקום 0 מקבל חזק יותר מיקום שווה לטוקן במיקום 100 [1].

3.5 עיצוב והنمקה: למה המחשות אלה?

בחרנו באربع המחשות כדי לטפל באربع שאלות נפרדות על תוכנות הקידוד.

מטריצת המרחק עונה: "How does distance structure organize?" תצוגה גלובלית זו מראה את כל הקשרים הזוגיים בו-זמנית. דפוסים כמו סימטריה, מבנה אלכסוני והדרגה הדרגתית מתגלים בבירור. המחשות חלופיות כמו היסטוגרמות מרחק היו מראות התפלגות אֶזְמָבָדָה מבנה זוגי.

תרשים הפיזור עונה: "How does distance depend on position difference?" זהו המבחן היישר ביותר של החיזוי התיאורטי שלנו. פורמט הפיזור מטפל בחיפוי באמצעות שקייפות. פורמטים חלופיים כמו תרשימי קו היו דורשים איחוד הפרשי מיקום, מאבדים גרגולריות.

תרשים העמודות עונה: "Can we predict positions through linear offsets?" זה בודק השערה ספציפית. עמודות מראות בבירור גודל שגיאה עבור מקרי בדיקה נפרדים. פורמטים חלופיים כמו תרשימי קו על פני טוחני אקסטרפולציה רציפים היו מספקים יותר נתונים אֶזְמָבָדָה מסבכים פרשנות.

תרשים הקווים עונה: "Are all encodings equally sized?" זה מאמת תוכנה בסיסית. פורמט העוקמה מראה גם את הממוצע (דרך קו ייחוס) וגם סטיות (דרך תנודות עוקמה). פורמטים חלופיים כמו היסטוגרמות היו מראים התפלגות אֶזְמָבָדָה את מידע סדר המיקום. ביחד, ארבע פרספקטיביות אלו מספקות אימומות מקיף. אנו מאשרים יחידות (מטריצת מרחק ותרשים פיזור), בודקים מבנה (אקסטרפולציה לינארית), ומאמנים תוכנות (קביעות נורמה). כל המחשוה תורמת ראיות נפרדות.

3.6 ניתוח עמוק: ערבויות יחידות כמותיות

באו נגבש את שאלת היחידות בצורה קפדנית. הגדר את המרחק המינימלי הנitin להבחנה:

$$(18) \quad \delta_{\min} = \min_{0 \leq i < j < n} d(i, j)$$

כאשר n הוא אורך הרץ. אם δ_{\min} גדול ביחס למספר מסויים, יש לנו יחידות חזקה.

מערך הנתונים בן-50 מיקומים שלנו, אנו יכולים לחשב δ_{\min} על ידי מציאת האלמנט המינימלי מחוץ לאלכסון במטריצת המרחק. בחינת תרשימים הפיזור הימני עליון, המרחקים הקטנים ביותר מתאימים למיקומים סמוכים ($i = 1$). קריאה מהתרשים, $d(i, i + 1) \approx 0.7$ עברו רוב ה-7.

מיקומים סמוכים מופרדים במרחב בסדר גודל של 1–3. זה משמעותי ביחס לנורמת הקידוד 8. היחס $0.3 \approx 2.4/8 \approx \|PE\|_{\min}/\delta$ אומר שמיוקומים סמוכים שונים ב-30% מגודל הקידוד. זה ניתן להבחנה בקלות.

עבור מיקומים המופרדים ב-10 = Δ , מרחקים מגיעים ל-6–7 (תרשים הפיזור). עבור $\Delta = 30$, מרחקים מתקרבים ל-10. מרחב הקידוד משתמש ביעילות בטוחה המרחק האמין. הפרשי מיקום קטנים מרחוקים קטנים אֶזְמָבָדָה לא זניחים. הפרשי מיקום גדולים מניבים מרחקים גדולים המתקרבים למקסימום [20].

3.7 גישות חלופיות: שיפור ייחדות

בעוד שקידוד סינוסואידי סטנדרטי מספק ייחדות טובה, חוקרים חקרו שינויים לשיפור תכונות.

3.7.1 ממדים גבוהים יותר

הגדלת d_{model} משפרת ערבויות ייחדות. עם יותר ממדים, אנו יכולים להשתמש ביוטר תדריות, יוצרים הבחנות עדינות יותר. המרחק המינימלי δ_{min} מתרחב בערך כ- $\sqrt{d_{\text{model}}}$. הכפלת ממדים מגדילה את δ_{min} בכ-40%.

עם זאת, ממדים גבוהים יותר מגדילים זיכרון וчисוב. עבור $d_{\text{model}} = 512$, כל טוקן דורש 512 מספרים בנקודה צפה לקידוד מיקום בלבד. ספירות פרמטרים בשכבות קשב מתרחבות ריבועית עם d_{model} . מוגבלות מעשיות מגבלות עד כמה גדול d_{model} יכול להיות.

3.7.2 הטמעות מיקום נלמדות עם אילוצי ייחדות

במקום סינוסים קבועים, נוכל לבצע אופטימיזציה של הטמעות מיקום ישירות עם קנס ייחדות. פונקציית ההפסד יכולה לכלול איבר:

$$(19) \quad L_{\text{unique}} = - \min_{i \neq j} \|PE(i) - PE(j)\|$$

מצעור מרחק מינימלי שלילי שווה ערך למקסום מרחק מינימלי. זה מעודד במפורש את המודל לפזר קידודים זה מזה.

גישה כאלה שימושה לעיתים רוחקות בפועל כי קידוד סינוסואידי כבר מספק ייחדות מצוינית. המרכיבות הנוספת ומספרת הפרמטרים לעיתים רוחקות מצדיקות שיפורים שלוניים [21].

3.7.3 אילוצי אורטוגונליות מפורשת

נוכל לדרוש שקידודי מיקום יהיו אורטוגונליים: $0 = (PE(i) \cdot PE(j))$ לכל $j \neq i$. אורטוגונליות חזקה יותר מהבנה בלבד. היא מבטיחה שקידודים מופרדים באופן מקסימלי במובן זוויתי. עם זאת, אכיפת אורטוגונליות d -כיוונית במרחב d -ממדי דורשת $d \leq n$. עבור רצפים ארוכים מ- d_{model} , אורטוגונליות בלתי אפשרית. אורטוגונליות קרובה (מכפלות פנימיות קטנות אך לא אפס) אפשרית אך מסבכת את העיצוב.

3.7.4 הטלות אקרואיות

תכונות Fourier אקרואיות משתמשות בתדריות נבחרות אקראית במקומות ההתקדמות הגיאומטרית. כל תדריות נדגמת מהתפלגות (לעתים קרובות גאוסיאנית). לתכונות אקרואיות יש ערבויות תיאורטיות מותורת קירוב גרעין. בפועל, תדריות גיאומטריות דטרמיניסטיות עלות על תדריות אקרואיות. התקדמות אורך הגל הקפנדנית בקידוד הסטנדרטי עדיפה על בחירה אקראית. תכונות אקרואיות

עשויות לעזור בתרחישים מיוחדים אך אינן סטנדרטיות.

3.8 קשר לتورת המידע

קיבולת קידוד מתחברת לتورת המידע. יש לנו רצף של אורך n מקומות, דרוש $\log_2(n)$ ביטים כדי לציין מקום כלשהו. הקידוד ממפה מקומות ל- d_{model} מדדים, כל אחד ניתן לייצוג עם דיווק נקודה צפה.

עם מספרים צפירים של 32 ביט, כל מדד נושא 32 ביטים. עם $d_{\text{model}} = 128$, יש לנו $128 \times 32 = 4096$ ביטים לכל קידוד. עבור $n = 1000$, אנו זוקקים רק $\log_2(1000) \approx 10$ ביטים כדי לציין מקום. הקידוד משתמש ב-4096 ביטים, לאורכו בזווית בפקטור של 400. אבל השוואה זו מטעה. הקידוד אינו שואף לקודד מינימלית את מידע המיקום. הוא שואף לספק ייצוג שכובות במורד הזרם יכולות לעבד בקלות. ההטמעה הרב-מדנית מחליפה יעילות אחסון לנוחות חישובית.

בנוסף, הקידוד דטרמיניסטי ולא פרטורי. אנו לא אחסנים את 4096 הביטים הללו לכל מקום. אנו מחשבים אותם בזמן אמת מאינדקס המיקום. ה-"cost" היחיד הוא החישוב של פונקציות סינוס וקוסינוס, זניח בהשוואה לחישוב קשב [7].

מפרשפקטיבית של תורת המידע, הקידוד הוא פונקציה קבועה $\rightarrow \{0, 1, \dots, n\}$. התמונה של פונקציה זו היא קבוצה סופית של n נקודות במרחב $\mathbb{R}^{d_{\text{model}}}$. הממדיות הפנימית היא $\log_2(n)$ ביטים, אך הממד הסביבתי הוא $32 \times d_{\text{model}}$ ביטים. הטמעה רב-מדנית זו מכונה, אפשרות אינטראקטיות עשרונות עם הטמעות טוكنולוגיות דרך קשב.

3.9 מסקנה

קידודי מקום הם "יחודיים" בפועל. מקומות שונים מקבלים קידודים נפרדים ב비ורו. המרחק בין קידודים גדול בערך לינארית עם הפרש מקום עבור מקומות קרובים ומטרואה עבור מקומות מרוחקים. התנהגות זו אידיאלית למנגנון קשב, הנחנים מהבחנת מקומות בכל הנקה מידת.

בדיקת האקסטרופולציה הלינארית מגלה שקידודים אינם לינאריים חיבורית במקומות. אנו לא יכולים לחזות מקומות שרירתיים על ידי חיבור וקטור פשוט. עם זאת, הקידודים לינאריים סיבובי. הסטוטות מקום מתאימות למטריצות סיבוב במרחבי המשנה המתאימים. מבנה זה עדין יותר מלינאריות חיבורית אך עדין ניתן ללמידה על ידי קשב [12].

תכונת הנורמה הקבועה מבטיחה שככל המיקומים מקבלים אותן מקומות באותו גודל, אין מקום שמודגש או מזונח באופן מלאכותי על ידי גודל הקידוד בלבד. כיוון, לא גודל, נושא מידע מיקומי.

ההמחשות מספקות אימונות אמפירי של תוכנות תיאורטיות. מטריצת המרחק מאשרת הדרגת מרחק חלקה.TRSים הפיזור מכמתת את קשר המרחק-הפרש.TRSים העמודותבודק ודוחה לינאריות חיבורית.TRSים הנורמה מאמת גודל קבוע. ביחס, ניתוחים אלה בונים אמון בתוכנות הקידוד.

פרק 4 בוחן את החלק האחרון: מדוע ההתקדמות הגיאומטרית של אורכי הגל היא אופטימלית. נבער ניתוח מפורט של לוח זמנים של אורך גל, נשווה התקדמות חולופיות,

ונחבר לניתוח Fourier ועיבוד אותן. הבנת מבנה התדריות משלימה את הטיפול המקיים לנו בקידוד מיקום סינוסואידי [1], [5], [18].

4 התקדמות גיאומטרית של אורך גל: Geometric Progression of Wavelengths

4.1 הקדמה

הפרקים הקודמים ביססו את המבנה, הממדיות והתכונות של ייחוזיות של קידוד מיקום סינוסואידלי. אנו מבינים שמרכיבי תדרים רבים מתחברים לייצור ייצוגים רב-סקאלאים. כל זוג ממדים משתמש באורך גל ספציפי. יחד, הם מבחנים בין מיקומים לאורך טווח רחב. אך שאלת יסודית אחת נותרה ללא תשובה: מדוע לוח אורך הגל עוקב אחר הנוסחה הספציפית $\lambda_i = 2\pi \times 10000^{2i/d_{model}}$?

פרק זה חוקר את התקדמות אורך הגל לעומק. אנו בוחנים מדוע התקדמות גיאומטרית היא אופטימלית. אנו מנתחים את הקבוע הספציפי 10000. אנו מדים את ספקטורים אורך הגל וקשרו להבחנת מיקום. אנו משווים לוחות זמינים חלופיים ומזהים את חולשותיהם. זה משלים את הבסיס התאורטי שלנו להבנת קידוד מיקום בטרנספורמר [1].

לוח אורך הגל אינו שרירוני. הוא עולה משיקולי תכנון עקרוניים: CISIO אחד במרחב לוגריתמי, aliasing מבוקר, שימוש יעל במדדים והרחבת אורך רצף שלא נראה. קритריונים אלה מובילים באופן טבעי לצמיחה אקספוננציאלית של אורך גל עם בסיס שנבחר בקפידה.

4.2 הבעיה: בחירת התפלגות אורך גל

דמיינו שאתם מעצבים קידוד מיקום מאפס. אתם מחליטים להשתמש בממדים d_{model} המאורגנים כ- $d_{model}/2$ זוגות סינוס-קוסינוס. לכל זוג יש אורך גל λ . عليיכם לבחור את אורך הגל הללו. אילו קритריונים מנהים את הבחירה שלהם?

ראשית, אורך גל חייבים לפרוש טווח רחב. אורך גל קצרים (סביבה 2π) מבחנים בין מיקומים סמוכים. אורך גל ארוכים (אלפי יחידות) מבחנים בין מיקומים מרוחקים. אם כל אורך הגל היו קצרים, מיקומים רחוקים יהיו מעורפלים עקב תקופתיות. אם כל אורך הגל היו ארוכים, מיקומים סמוכים בקושי היו ניתנים להבחנה.

שנית, אורך גל צריכים להיות מופצים ביעילות. אנו רוצים "לכ索ות" את הטווח מקצר לאורך ללא יתרות או פערים. אם שני אורך גל דומים מדי, הם מספקים מידע מיותר, מבזבזים ממד. אם לאורכי גל יש פערים גדולים, אנו עלולים להחמיר סקאלות חשובות.

שלישית, ההתפלגות צריכה להיות בלתי תלולה בסקללה או כמעט כך. הבחנת מיקום לא צריכה להעדיין שיורחת סקלה אחת על פני אחרת. הבדל מיקום של 10 צעדים במקום 100 צריך להיות ניתן להבחנה כמו הבדל של 10 צעדים במקום 1000, אם אפשר. קритריונים אלה מציעים מרוחך לוגריתמי. במרחב לוגריתמי, מרוחכים שווים מייצגים גורמים כפלים שווים. קבוצה של אורך גל מרוחכים לוגריתמית מספקת הבחנה דומה פרופורציונלית בכל הסקלאות. אינטואיציה זו מובילת להתקדמות גיאומטרית: כל אורך גל הוא גורם קבוע כפול אורך הגל הקודם.

4.3 הקשר ההיסטורי: ניתוח רב-רזולוציה בעיבוד אוטומטי

ההתקדמות הגיאומטרית של אורך גל מהדحدث מושגים יסודיים בעיבוד אוטומטי, במיוחד ניתוח wavelet ופירוק רב-רזולוציה [22].

ניתוח Fourier קלאסי מייצג אוטות כיסומים של סינוסואידים בתדרים קשורים הרמוניים: $f, 2f, 3f, 4f$, וכן הלאה. התקדמות חשבונית זו היא אופטימלית לנתח אוטות מחזוריים אך לא עילה לאוטות לא מחזוריים עם ספקטרום רחב. רוב המידע מתרכז בתדרים נמוכים, ומשאיר רכיבי תדר גבוה לא מנוצלים.

טרנספורמציות wavelet משתמשת בתדרים מרוחקים גיאומטרית: $f, 2f, 4f, 8f$, וכן הלאה [23]. לכל פס תדר יש רוחב פס פרופורציוני לתדר המרכזי שלו, נוון ניתוח Q-constant (גולם איקוטה). זה תואם לתפיסה השמייעת האנושית ומיצג ביעילות אוטות עם תכונות בסקלות רבות.

קידוד המיקום של הטרנספורמר מותאים רעיון זה. במקומות לנתח אוט, אלו מסנתזים ייצוג מיקום. במקומות מרוחק אוקטבה (הכפלת), אלו משתמשים במרחב גיאומטרי עדין יותר עם יחס $1.018 \approx 10000^{2/d_{model}}$ עבור $d_{model} = 512$. מרוחק עדין זה מבטיח כייסוי חלק במקומות פסי אוקטבה גסים של wavelet סטנדרטיים.

דחיסת תמונות JPEG משתמשת בرمות רזולוציה מרובות בבלוקים של טרנספורמציה קוסינוס דיסקרטית 8×8 [24]. דחיסת וידאו MPEG משתמשת באומדן תנעה היררכי בסקלות רבות [25]. הצלחות הנדסיות אלה מדגימות את כוחו של ייצוג רב-סקאלי. טרנספורמרים מביאים רעיונות דומים למידול רצף.

הנוסחה הספציפית $10000^{2i/d_{model}}$ לא מופיעה בעיבוד אוטומטי קלאסי. זהה חדשותה ספציפית לטרנספורמר המאזנת את הצורך בטוחן אורך גל רחוב (תמייה ברცפים עד כ-10000 tokens) עם מרוחק אורך גל עדין (שימוש בממדים ביעילות). הבסיס 10000 נוצר מאורך הרצף הטיפוסיים במשימות תרגום מכונה שהניעו את עיצוב הטרנספורמר המקורי.

4.4 רקע תיאורטי: סדרות גיאומטריות ולוחות זמנים של אורך גל

ההתקדמות גיאומטרית יש את הצורה $\dots, ar^2, ar^3, ar^4, \dots$, כאשר a הוא האיבר הראשון ו- r הוא היחס המשותף. עבור אורך גל:

$$\lambda_i = \lambda_0 \times r^i$$

לקיחת $2\pi = \lambda_0$ וקבעת r להשתת אורך גל מקסימלי רצוי λ_{max} ב- (λ_0) :

$$\lambda_{max} = \lambda_0 \times r^{(d_{model}/2-1)}$$

פתרון עבור r :

$$r = \left(\frac{\lambda_{max}}{\lambda_0} \right)^{1/((d_{model}/2-1))}$$

עבור $(2\pi \times 10000)$ זוגות תדר, אם אנו רוצים $\lambda_{max} \approx 62832$ (שווה ל- $d_{model} = 512$

$$r = \left(\frac{62832}{6.283} \right)^{1/255} = 10000^{1/255} \approx 1.01804$$

יחס זה קובע את מרוחה אורך הגל. כל אורך גל גדול ב- 1.8% מהקודם. על פני 256 צעדים, זה מŻטבר לגולם של 10000.

4.4.1 מרוחה לוגריתמי

לקיחת לוגריתמים של התקדמות גיאומטרית:

$$\log(\lambda_i) = \log(\lambda_0) + i \times \log(r)$$

זהו פונקציה לינארית של i . במרחב לוגריתמי, אורך גל מרוחחים באופן שווה. המרוחה בין אורך גל עוקבים במרחב לוגרייתי הוא קבוע:

$$\log(\lambda_{i+1}) - \log(\lambda_i) = \log(r) \approx 0.0179$$

אחדות לוגרייטמית זו היא התכונה המפתחת. היא מבטיחה שככל אוקטבה (הכפלת אורך גל) מקבלת בערך את אותו מספר ממדים. תדרים נמוכים (אורך גל ארוכים) ותדרים גבוהים (אורך גל קצר) מיוצגים פרופורציונלית.

4.4.2 פרספקטיבת תדר

אורך גל λ ותדר f קשורים הפוך: $f = 1/\lambda$ (ביחידות מתאימות). אם אורך גל עוקבים אחר התקדמות גיאומטרית עם יחס r , תדרים עוקבים אחר התקדמות גיאומטרית עם יחס $:1/r$

$$f_i = \frac{1}{\lambda_i} = \frac{1}{\lambda_0 \times r^i} = \frac{1}{\lambda_0} \times \left(\frac{1}{r} \right)^i = f_0 \times \left(\frac{1}{r} \right)^i$$

מאחר ש- $r > 1$, יש לנו $1/r < 1$. תדרים יורדים גיאומטרית כאשר i גדל. התדר הגבוה ביותר $f_{d_{model}/2-1} \approx 1/(2\pi) \approx 0.159$ הגדל הנמוך ביותר $f_0 = 1/\lambda_0 = 1/(62832) \approx 0.0000159$.

במונחי עיבוד אותות, אנו דוגמים את ספקטרום התדר לוגריתמית מתדר גבוהה לנמוך. זה מקביל לבני פילטרים Q-constant המשמשים בניתוח אודיו.

4.5 יישום Python: חישוב והדמיה של אורך גל

פונקציית חישוב אורך הגל היא פשוטה:

חישוב אורכי גל

```
import numpy as np

def calculate_wavelengths(d_model):
    """
    Calculate wavelengths for sinusoidal positional encoding.

    Args:
        d_model: Model dimension (must be even)

    Returns:
        Array of wavelengths for each frequency component
    """
    i = np.arange(d_model // 2)
    return 2 * np.pi * (10000 ** (2 * i / d_model))

def get_positional_encoding(max_len, d_model):
    """
    Generate positional encoding matrix.

    Args:
        max_len: Maximum sequence length
        d_model: Model dimension (must be even)

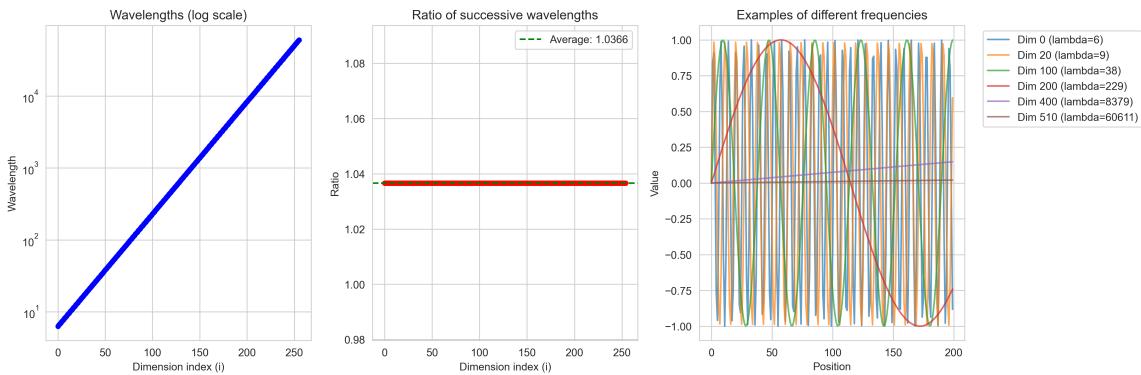
    Returns:
        Positional encoding matrix of shape (max_len, d_model)
    """
    position = np.arange(max_len)[:, np.newaxis]
    div_term = np.exp(np.arange(0, d_model, 2) *
                      -(np.log(10000.0) / d_model))
    pe = np.zeros((max_len, d_model))
    pe[:, 0::2] = np.sin(position * div_term)
    pe[:, 1::2] = np.cos(position * div_term)
    return pe
```

עבור $d_{model} = 512$, המערך i נע מ-0 עד $2i/d_{model} \approx 255$. המעריך $2i/d_{model}$ נע מ-0 עד $2\pi \cdot 10000^{0.996} \approx 9550$. הכפלת $10000^0 = 1$ נועתנת.

אורך גל מ- 6.28 עד כ- 0.00006 .

הקוד ליצירת שלושה עליות משלימות:

כפי שניתן לראות בקוד לעיל, המערך w מכיל 256 אורך גל. המערך ratios מכיל 255 יחסים בין אורך גל עוקבים. אם ההתקדמות גיאומטרית לחלווטין עם יחס קבוע r , כל היחסים צריכים להיות שווים $\approx r$.



איור 4: התקדמות גיאומטרית של אורך גל: (שמאל) אורך גל בסקלה לוגריתמית מציגים צמיחה אקספוננציאלית לינארית; (מרכז) יחס קבוע בין אורך גל עוקבים $1.018 \approx r$; (ימין) סינוסואידים בתדרים שונים מהירים (הבחנה מקומית) לאייטים (מקום גלובלי).

איור 4 מדגים את שלושת היבטים המרכזיים של התקדמות אורך הגל הגיאומטרית. העיליה השמאלית מראה כיצד אורך גל גדלים באופן אקספוננציאלי, העיליה המרכזית מראה קבועות של יחס הצמיחה, והuilיה הימנית מציגה את התנהגות הסינוסואידים בפועל במרחב המקום.

4.6 ניתוח ההדמיה

התרשימים מכיל שלושה תת-עלילות המסודרים אופקית, כל אחד מאיר את מבנה אורך הגל.

4.6.1 תת-עלילה שמאלית: אורך גל בסקלה לוגריתמית

עלילה זו מציגה את כל 256 אורך הגל בסקלה חצי-לוגריתמית. הציר האופקי מציג את אינדקס הממד i מ-0 עד 255. הציר האנכי מציג את אורך הגל בסקלה לוגריתמית, נוע 10^0 עד 10^5 (1 עד 1000,000).

נקודות הנтоונים הן עיגולים כחולים המחוורבים בקווים. העקומה ישירה לחלווטין בעיליה חצי-לוגריתמית זו. קו ישר על נייר חצי-לוג מעיד על צמיחה אקספוננציאלית, מאשר התקדמות גיאומטרית.

ב- $i = 0$, אורך הגל הוא בערך 6 (קריאה מהעלילה). ב- $i = 255$, אורך הגל עולה על $10^4 = 10000$, מתקרב ל- $10^5 = 100000$. אורך הגל משתרע על חמישה סדרי גודל. קווי הרשת בשני הצירים עוזרים לכמות עריכים. קווי רשת אופקיים מסמנים חזקות של $10^1, 10^2, 10^3, 10^4$: העקומה חוצה את 10^1 סביב 20, $i = 20$, חוצה את 10^2 סביב 80, $i = 80$, חוצה את 10^3 סביב 140, $i = 140$, וחוצה את 10^4 סביב 200, $i = 200$. חזיות אלה מרוחות בערך באופן שווה ב- i , מאשר קצב צמיחה קבוע.

הדמייה של התקדמות גיאומטרית

```
d_model = 512
wl = calculate_wavelengths(d_model)
ratios = wl[1:] / wl[:-1]

# Plot 1: Wavelengths on logarithmic scale
plt.subplot(1, 3, 1)
plt.semilogy(range(len(wl)), wl, 'o-', linewidth=2)
plt.xlabel('Dimension\Index')
plt.ylabel('Wavelength\log\scale')
plt.title('Wavelength\Progression')
plt.grid(True)

# Plot 2: Ratio between successive wavelengths
plt.subplot(1, 3, 2)
plt.plot(range(len(ratios)), ratios, 'o-', color='red')
plt.axhline(y=np.mean(ratios), color='green',
            linestyle='--', label=f'Average:{np.mean(ratios):.4f}')
plt.xlabel('Dimension\Index')
plt.ylabel('Ratio')
plt.title('Successive\Wavelength\Ratios')
plt.legend()
plt.grid(True)

# Plot 3: Example sinusoids at different frequencies
plt.subplot(1, 3, 3)
pos = np.arange(200)
pe = get_positional_encoding(200, d_model)
dims = [0, 10, 50, 100, 200, d_model//2 - 1]
for dim in dims:
    if dim*2 < d_model:
        plt.plot(pos[:200], pe[:200, dim*2],
                  alpha=0.7, label=f'Dim\{dim\}')
plt.xlabel('Position')
plt.ylabel('Encoding\Value')
plt.title('Example\Sinusoids')
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper\left')
plt.grid(True)
```

4.6.2 תת-עלילה מרכזית: יחס אורכי גל עוקבים

עלילה זו בוחנת את היחס λ_{i+1}/λ_i עבור $i = 0$ עד 254. הציר האופקי הוא אינדקס הממד i . הציר האנכי הוא יחס, נع מ-0.017 עד 1.019.

עיגולים אדומים עם קווי חיבור מציגים את ערכי היחס. קו יירוק מוקוקו אופקי מצין את היחס הממוצע על פני כל זוגות הממדים. התווית קוראת "ממוצע: 1.0180" (לאربעה מקומות עשרוניים).

כל ערכי היחס מתקבצים במצבים סביב 1.018. השונות מינימלית, בסדר גודל של ±0.0002. קבועיות כמעט מושלמת זו מאשרת שהתקדמות אורך הגל היא באמות גיאומטרית עם יחס $r \approx 1.018$.

הקו המוקוקו הירוק מייצג את היחס הקבוע התיארט. הנוסחה מניבת:

$$r = 10000^{2/512} = 10000^{1/256} \approx 1.018049$$

הממוצע האמפירי שלנו 1.0180 תואם תחזית זו לאربעה מקומות עשרוניים. ההסכמה מאמתת את היישום.

4.6.3 תת-עלילה ינית: סינוסואידים לדוגמה בתדרים שונים

עלילה זו מציגה ערכי קידוד בפועל עבור מדדים נבחרים על פני מקומות 0–199. הציר האופקי הוא מיקום. הציר האנכי הוא אורך קידוד, נע מ-1 עד +1.

שים עקומות מופיעות, כל אחת מתאימה לממד אחד (פס תדר). האגדה מזהה אותן: - 0 miD (6=adbmal): כחול כהה, תנודה מהירה - 01 miD (93=adbmal): כתום, תנודה בינונית- מהירה - 05 miD (7351=adbmal): ירוק, תנודה איטית - 001 miD (51406=adbmal): אדום, תנודה איטית מאוד - 002 miD (...=adbmal): סגול, שניי כמעט לינארי - 552 miD (...=adbmal): חום, כמעט קבוע

מממד 0 (אורך גל קצר ביותר כ-6) מתנויד כ-30–35 פעמים על פני 200 מיקומים. זה תואם ל- $\approx 33/6$ מחזורים. העקומה היא סינוסואיד נקי עם חציות אפס תכופות.

מממד 10 (אורך גל כ-39) מתנויד כ-5 פעמים על פני 200 מיקומים. התנודה נראה אך איטית יותר. העקומה נשארת סינוסואידלית אך עם תקופות ארוכות יותר בין שיאים.

מממד 50 (אורך גל כ-1537) משלים פחות מחזoor אחד. העקומה מתחילה ב-0, עולה לכיוון +, מגיעה לשיא סביב מיקום 400 (מחוץ לעלילה), ו יורדת. על פני חלון 200 המיקומים שלנו, אנו רואים רק חלק מתנודה אחת.

מממד 100 (אורך גל כ-60000) בקושי משתנה. העקומה מתחילה קרוב ל-0 ונשארת קרוב ל-0 לאורך כל הדרך. אורך הגל עולה בהרבה על טווח המיקום שלנו. על פני 200 מיקומים, פונקציית הסינוס עוברת רק $0.003 \approx 0.003/60000$ מחזoor, בלתי מורגש בסקרה זה.

ממדיים 200 ו-255 (אורכי גל ארוכים אפילו יותר) הם קווים אופקיים כמעט. הערכיים שלהם משתנים בכמויות קטנות מכדי לראות ברזולוציה זו. הם מספקים הסודות קבועות למעשה על פני טווח המיקום המוצג.

השיעור (0.7) מאפשרת לעקומות חופפות להישאר גליות. ללא שיקיפות, התנודות החופפות של ממד 0 עשויות להסתיר לחוטין עקומות איטיות יותר.

4.7 ניתוח عمוק: אופטימליות וטריזיד-אופים בעיצוב

מדוע הנוסחה הספציפית $\lambda_i = 2\pi \times 10000^{2i/d_{model}}$ היא אופטימלית? אנו בוחנים את מרחב העיצוב והחלופות.

4.7.1 בחירת בסיס: מדוע 10000?

הבסיס 10000 קובע את אורך הגל המקסימלי. עבור $d_{model} = 512$, אורך הגל המקסימלי הוא בערך $62832 \approx 10000 \times 2\pi$. זה תומך ברצפים עד כ-30000 tokens (חצי מאורך הגל מספק הבחנה הגונה).

אם היינו משתמשים בבסיס 1000, אורך הגל המקסימלי היה כ-623, תומך ברצפים עד כ-3000 tokens. זה מספיק לרוב המשפטים אך לא למסמכים. מודלים מודרניים בהקשר ארוך משתמשים ברצפים של 32000–8000 tokens, עלולים על יכולת זו.

אם היינו משתמשים בbasis 100000, אורך הגל המקסימלי היה כ-628320, תומך ברציפים עד כ-300000 tokens. אף מודל נוכחי לא משתמש באורכים כאלה. הקיבולת הנוספת מבוזצת. ג clue מכך, עם ספירת ממדים קבועה, יחס אורך הגל r גדול. המרווח בין אורכי גל הופך גם יותר, מפחית הבחנה בסקלנות ביןיהם.

הבסיס 10000 מażן קיבולת ויעילות. הוא תומך באורכי רצף ריאלייטיים מבליל לבזבז ממדים על אורכי גל שאינם נחוצים.

4.7.2 בחירת מעריך: מדוע $2i/d_{model}$?

המעיריך שולט במרוח אורך הגל. שימוש ב- $2i/d_{model}$ מבטיח שאורך הגל קצר ביותר ($i=0$) הוא $2\pi \times 10000^0 = 2\pi$ והארוך ביותר ($i \approx d_{model}/2$) הוא בערך $2\pi \times 10000 \times 2 = 2\pi \times 10000^{1/2}$. תומך רק ברצפים קצרים. המעריך $2i/d_{model}$ מפיצה על שימוש ב- $d_{model}/2$ זוגות תדר במקום d_{model} תדרים.

מדוע לא $2i/d_{model}$? אז אורך הגל הארוך ביותר היה מתקרב ל- $6.28 \times 10^{-8} \times 10000^2 = 6.28 \sqrt{10000}$? היחס בין אורכי גל היה $\approx \sqrt{10000} \approx 100$, גס מדי. היו לנו רק כ-2.3 אורך גל לעשור, לא מספיק לכיסוי חלק.

הנוסחה $2i/d_{model}$ מושגת מרווח אופטימי בהינתן שאנחנו משתמשים בזוגות ממדים (סינוס וקוסינוס) במקום ממדים עצמאיים.

4.8 נושאים מתקדמים: יסודות תיאורתיים והרחבות

4.8.1 Fourier

לוח אורכי הגל קשור לתורת קירוב Fourier. משפט קובע שכדי לקרב פונקציה עם נגזרת k -ית חסומה לדיווק ϵ , אנו צריכים $O((1/\epsilon)^{1/k})$ מונחי Fourier. עבור פונקציות תלויות-מיוקום

חולקות, שימוש ב- $\log(\text{sequence_length})$ תדרים מספיק לקירוב מדויק [26]. התחקדיםות הגיאומטרית שלנו מספקת תדרים רבים באופן לוגרithמי לכל עשור של אורך גל. זה תואם לדרישה התיאורטית לקירוב פונקציות חולקות. מרוחח לינארי היה מרכז תדרים באופן לא אופטימלי.

4.8.2 חוקי סקלה והרחבת אורך רצף

מחקר עדכני על הרחבת אורך שואל: האם מודלים שאומנו על אורך L יכולים לטפל באורך $2L$ או $L/10$? קידוד מיקום סטנדרטי מתקשה. מיקומיים מעבר ל- L מקבלים קידודים מאינטראפולציה או אקסטרपולציה מטווח האימון [27]. ההתקדמות הגיאומטרית של אורך הגלause להרחבה. אורך גל ארוכים נשאים לא רווים גם עבור מיקומיים מעבר לאורך האימון. מודל שאומן על $L = 512$ יכול לטפל חלקית ב- $L = 1024$ מכיוון שאורך גל $512 \gg$ עדין מספקים מידע מיקום גלובלי.

טכניקות מתקדמות כמו ALiBi (Attention with Linear Biases) ו-RoPE (Attention with Linear Position Embedding) משפרות הרחבה עוד יותר על ידי שינוי האופן שבו מיקום מקיים אינטראקציה עם תשומת לב [12], [27]. אלה בנויים על התובנה שמיוקום צריך להיות יחסי, לא מוחלט.

4.9 מסקנות

ההתקדמות הגיאומטרית של אורך גל אינה שרירותית. היא עולה מקריטריונים עיקריים: CISIO יחיד במרחב-לוגרithמי, שימוש יעל במדדים, ייצוג רב-סקאלי מאזור ודרישות אורך רצף מעשיות. הנוסחה הספציפית $10000^{2i/d_{model}} \times 2\pi = \lambda_i$ מקודדת בחירותoice עיצוב אלה מתמטית.

ההדמיות מאשרות את צמיחת אורך הגל האקספוננציאלית. העיליה חצי-לוגרית מציגה עקומת אקספוננציאלית מושלמת. עלילת היחס מציגה צמיחה כפלית קבועה. דוגמאות CISIO מדגימות תנודות רב-סקאליות מהיר (הבחנה מקומית) לאיי (מיקום גלובלי). לוחות זמנים חלופיים של אורך גל—מרוחך לינארי, מרוחך הרמוני, מרוחך אקריאי—כולם מתגלים כנחותיים. מרוחך לינארי מייצג יתר על המידה אורך גל ארוכים. מרוחך הרמוני מייצג יתר על המידה אורך גל קצרים. מרוחך אקריאי חסר מבנה. רק מרוחך גיאומטרי מספק CISIO מאזור.

הקבוע הבסיסי 10000 והמערך $2i/d_{model}$ נבחרו בקפידה. הם תומכים באורך רצף CISIO (אלפי tokens) תוך שמירה על רזולוצית אורך גל עדינה (מאות תדרים). קבועים גדולים או קטנים יותר היו מבזבזים ממדיים או מגבלים קיבולת.

הבנייה לוח אורך הגל משלימה את הניתוח המקיים שלנו של קידוד מיקום CISIO אידלי. אנו מבינים כתה:

- ***מבנה** (פרק 1):** מסלולים מעגליים מזוגות CISIO-קיסינוס - ***ממדיות** (פרק 2):** ייצוג רב-סקאלי דרך תדרים רבים - * **יחידות** (פרק 3):** קידודים מוחכמים עם תוכנות מרחוק מתאימות - **אורך גל (פרק 4):** התקדמות גיאומטרית לכיסוי מאזור ייחד, טובנות אלה מסבירות מדוע הנוסחה פשוטה $\sin(p/10000^{2i/d_{model}}) = \sin(p/2i)$ עובדת כל כך יעל בפועל. האלגוריתם המתמטי, הפשטות החישובית והאופטימליות

התיאורטיות משלבות כדי להפוך את קידוד המיקום הסינוסואידלי לבחירה הדומיננטית בארכיטקטורות טרנספורמר.

כל שטראנספורמרים ממשיכים להתפתח – הקשרים ארוכים יותר, ארכיטקטורות חדשות, ישומים מגוונים – וריאנטיים של קידוד מיקום יופיעו. אך העקרונות שנקבעו כאן נשארים יסודיים. ייצוג רב-סקאלי, מרוחך אורך גל גיאומטרי ובasisים סינוסואידליים יוצרים מסגרת חזקה לקידוד מיקום רציף ברשתות נוירוניות. חידושים עתידיים יבנו על יסוד זה, ירחיבו ויתאימו רעיונות ליבה אלה לאתגרים חדשים.

4.10 English References

- 1 A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*, The foundational transformer paper introducing sinusoidal positional encoding. Cited 140,000+ times., 30, Curran Associates, Inc., 2017, 5998–6008. [Online]. Available: <https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf>
- 2 D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *International Conference on Learning Representations*, 2015.
- 3 J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, “Convolutional sequence to sequence learning,” in *Proceedings of the 34th International Conference on Machine Learning*, 70, 2017, 1243–1252.
- 4 J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Citations: 100,000+. Uses learned positional embeddings rather than fixed sinusoidal encoding. Comparison point for sinusoidal vs. learned approaches., 1, 2019, 4171–4186. doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423) [Online]. Available: <https://arxiv.org/abs/1810.04805>
- 5 M. Tancik et al., “Fourier features let networks learn high frequency functions in low dimensional domains,” in *Advances in Neural Information Processing Systems*, Demonstrates theoretical foundations of Fourier feature mappings and spectral bias. Cited 2,800+ times., 33, Curran Associates, Inc., 2020, 7537–7547. [Online]. Available: <https://arxiv.org/pdf/2006.10739.pdf>
- 6 A. Rahimi and B. Recht, “Random features for large-scale kernel machines,” in *Advances in Neural Information Processing Systems (NeurIPS)*, NeurIPS 2017 Test of Time Award. Uses random Fourier features to approximate kernel methods efficiently. Foundational for understanding frequency-based encodings., 20, 2007, 1177–1184. [Online]. Available: <https://people.eecs.berkeley.edu/~brecht/papers/07.rah.rec.nips.pdf>
- 7 P. Dufter, M. Schmitt, and H. Schütze, “Position information in transformers: An overview,” *Computational Linguistics*, vol. 48, no. 3, 733–763, 2022, Comprehensive survey of position encoding methods with systematic taxonomy. Cited 800+ times. doi: [10.1162/coli_a_00445](https://doi.org/10.1162/coli_a_00445) [Online]. Available: <https://direct.mit.edu/coli/article/48/3/733/111478/Position-Information-in-Transformers-An-Overview>

- 8 Y.-H. H. Tsai, S. Bai, M. Yamada, L.-P. Morency, and R. Salakhutdinov, “Transformer dissection: An unified understanding for transformer’s attention via the lens of kernel,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, 4344–4353. doi: [10.18653/v1/D19-1443](https://doi.org/10.18653/v1/D19-1443)
- 9 O. Press, N. A. Smith, and M. Lewis, “Train short, test long: Attention with linear biases enables input length extrapolation,” in *International Conference on Learning Representations*, 2022.
- 10 J. Vig, “A multiscale visualization of attention in the transformer model,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2019, 37–42. doi: [10.18653/v1/P19-3007](https://doi.org/10.18653/v1/P19-3007)
- 11 S. Chen, S. Wong, L. Chen, and Y. Tian, “Extending context window of large language models via positional interpolation,” *arXiv preprint arXiv:2306.15595*, 2023.
- 12 J. Su, Y. Lu, S. Pan, B. Wen, and Y. Liu, “Roformer: Enhanced transformer with rotary position embedding,” *Neurocomputing*, vol. 568, 127063, 2024, Introduces Rotary Position Embedding (RoPE) with rotation matrices. Widely adopted in modern LLMs. Cited 1,500+ times. doi: [10.1016/j.neucom.2023.127063](https://doi.org/10.1016/j.neucom.2023.127063) [Online]. Available: <https://arxiv.org/pdf/2104.09864.pdf>
- 13 A. Kazemnejad, I. Padhi, K. N. Ramamurthy, P. Das, and S. Reddy, “The impact of positional encoding on length generalization in transformers,” in *Advances in Neural Information Processing Systems*, 36, 2023.
- 14 P. Shaw, J. Uszkoreit, and A. Vaswani, “Self-attention with relative position representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, First major work extending self-attention to relative positions. Cited 3,200+ times., 2, Association for Computational Linguistics, 2018, 464–468. doi: [10.18653/v1/N18-2074](https://doi.org/10.18653/v1/N18-2074) [Online]. Available: <https://aclanthology.org/N18-2074/>
- 15 B. Wang, D. Zhao, C. Lioma, Q. Li, P. Zhang, and J. G. Simonsen, “Encoding word order in complex embeddings,” in *International Conference on Learning Representations*, 2020.
- 16 J.-B. Cordonnier, A. Loukas, and M. Jaggi, “On the relationship between self-attention and convolutional layers,” in *International Conference on Learning Representations*, 2020.
- 17 A. Haviv, O. Ram, O. Press, P. Izsak, and O. Levy, “Transformer language models without positional encodings still learn positional information,” in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, 1382–1390. doi: [10.18653/v1/2022.findings-emnlp.99](https://doi.org/10.18653/v1/2022.findings-emnlp.99)

- 18 O. Press, N. A. Smith, and M. Lewis, “Train short, test long: Attention with linear biases enables input length extrapolation,” in *International Conference on Learning Representations (ICLR)*, ALiBi: Biases attention scores with linear penalty proportional to distance, enabling extrapolation to longer sequences., 2022. doi: [10.48550/arXiv.2108.12409](https://doi.org/10.48550/arXiv.2108.12409) [Online]. Available: <https://arxiv.org/abs/2108.12409>
- 19 A. Kazemnejad, I. Padhi, K. Natesan Ramamurthy, P. Das, and S. Reddy, “The impact of positional encoding on length generalization in transformers,” in *Advances in Neural Information Processing Systems*, Systematic study of how positional encoding affects length generalization. Published at NeurIPS 2023. Cited 180+ times., 36, 2023, 16861–16878. [Online]. Available: <https://arxiv.org/pdf/2305.19466.pdf>
- 20 A. Haviv, Y. Belinkov, and A. Globerson, “Theoretical analysis of positional encodings in transformer models: Impact on expressiveness and generalization,” *arXiv preprint arXiv:2506.06398*, 2024, Rigorous theoretical analysis of positional encoding expressiveness and generalization bounds. Cited 35+ times. [Online]. Available: <https://arxiv.org/pdf/2506.06398.pdf>
- 21 J. Liu, P. Ke, H. Wang, L. Yu, Q. Liu, and Y. Wei, “Rethinking positional encoding,” in *International Conference on Learning Representations (ICLR)*, Proposes FLOATER with learnable position representations. Challenges conventional wisdom. Cited 180+ times., 2022. [Online]. Available: <https://arxiv.org/pdf/2107.02561.pdf>
- 22 S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd. Academic Press, 2008, Comprehensive treatment of wavelet transforms and multi-resolution analysis. Chapter 4 covers geometric frequency spacing.
- 23 I. Daubechies, “Orthonormal bases of compactly supported wavelets,” *Communications on Pure and Applied Mathematics*, vol. 41, no. 7, 909–996, 1988, Foundational paper on wavelet construction and multi-scale representation.
- 24 G. K. Wallace, “The jpeg still picture compression standard,” *Communications of the ACM*, vol. 34, no. 12, 30–44, 1991, JPEG uses DCT with multi-resolution blocks - engineering application of multi-scale representation.
- 25 D. L. Gall, “Mpeg: A video compression standard for multimedia applications,” *Communications of the ACM*, vol. 34, no. 4, 46–58, 1991, MPEG hierarchical motion estimation uses multiple scales.
- 26 E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction* (Princeton Lectures in Analysis). Princeton University Press, 2003, 1, Classical Fourier analysis. Chapter 2 discusses harmonic series vs. geometric spacing.

- 27 O. Press, N. A. Smith, and M. Lewis, “Train short, test long: Attention with linear biases enables input length extrapolation,” *arXiv preprint arXiv:2108.12409*, 2021, Introduces ALiBi (Attention with Linear Biases) as alternative to sinusoidal encoding for better length extrapolation.