

2025

Mamba-Transformer בראייה ממוחשבת: מבט השווואתי

סמינר במבנה מלאכותית
מחבר: שמעון פישמן

תוכן עניינים

3	מבוא	
4	Transformer1
4	הקדמה	
4	ארQUITקטורה	
5	Attention	
5	Multi-Head Attention	
5	Position-wise Feed-Forward Networks	
5	Positional Encoding	
6	סיכון	
7	טרנספורמרים - ראייה ממוחשבת2
7	הקדמה	
7	ניסיונות קודמים למש טרנספורמר עבור ראייה ממוחשבת	
7	TiT	
8	Fine-tuning	
8	ניסויים והשוואות	
9	סיכון	
10	Segment anything in medical images3
10	הקדמה	
10	SAM	
11	MedSAM	
11	ארQUITקטורה	
11	השוואת ביצועים	
13	"על העבודה הרפואי"	
13	סיכון	
14	MAMBA4
14	הקדמה	
14	S4	
15	MAMBA	
15	מנגנון הבחירה	
16	התאמת לחומרה	
17	בלוק ה-MAMBA	
17	ניסויים אמפיריים	
18	העתקה סלקטיבית	
18	DNA Modeling	
19	Speed and Memory Benchmarks	

20	סיכום	
21	Vision Mamba	.5
21	הקדמה	
21	ארQUITטורה	
21	תיאור כללי	
21	בלוק ה-Vim	
22	ניסויים והשוואות	
22	קלאוסיפיקציה של תמונות	
22	סגמנטציה סמנטית	
23	זיהוי אובייקטים וסגמנטציה לפי מופע	
23	זיכרון ומהירות	
24	סיכום	
25	מhiba עבור סגמנטציה של צילומים רפואיים בתלת-ממד SegMamba	.6
25	הקדמה	
25	ארQUITטורה	
25	בלוק מhiba מרחבית באוריינטציה תלת-ממדית (TSMamba) Block	
26	Feature-level Uncertainty Estimation (FUE)	
27	ניסויים והשוואות	
28	סיכום	
29	MambaVision - מודל מhiba-טרנספורמר היברידי עבור ראייה ממוחשבת	7.
29	הקדמה	
29	ארQUITטורה	
29	שלבים 1-2	
30	שלבים 3-4	
31	ניסויים והשוואות	
31	Image classification	
32	זיהוי אובייקטים וסגמנטציה	
32	אימון בקנה מידת גודל על ImageNet-21K	
32	עיצוב בлок ה- <i>honi</i> on MambaVision	
33	דפוא היברידי	
33	סיכום	
34	סיכום	
35	ביבליוגרפיה	

מבוא

עבודה זו מציגה סקירה של מאמרים בנוגע למימוש מודלים מבוססי ארכיטקטורת ה-Transformer וארכיטקטורת-h-Mamba עבור ראייה ממוחשבת. בחלק הראשון של העבודה (פרק 1-3) נועסק בארכיטקטורת הטרנספורמר. פרק 1 יציג סקירה תמציתית של ארכיטקטורת הטרנספורמר כפי שהוצגה במאמר משנת 2017 [1]. פרק 2 מפרט את מודל-h-Vision Transformer [2] שמשמש את ארכיטקטורת הטרנספורמר עבור ראייה ממוחשבת. פרק 3 מציג שימוש Vision Transformer עבור משימה שימושית בעולם האמתי, יצירת כל עבור רפואי לצורכי סגמנטציה של צילומים רפואיים [5].

החלק השני של העבודה (פרק 4-6) מכיל מבנה כמעט סימטרי לראשון. פרק 4 סוקר את ארכיטקטורת הממבה [6], פרק 5 סוקר ארכיטקטורה הממשת את מבנה עבור ראייה ממוחשבת (ViM) [10] ופרק 6 מציג שימוש של מבנה באופן ספציפי עבור סגמנטציה של צילומים רפואיים בתלת-ממד [11].

בחלק האחרון של העבודה (פרק 7) מוצגת ארכיטקטורה היברידית המשלבת מבנה וטרנספורמר לצורכי יצירת עמוד שדרה עבור מודלי ראייה ממוחשבת [12].

העבודה שואפת להנחיל לקורא הבנה יסודית של מודלי הטרנספורמר והממבה ושל דרכי המימוש המגוונות שלהם עבור משימות ראייה ממוחשבת.

Transformer .1

הקדמה

פרק זה נציג את ארכיטקטורת הטרנספורמר כפי שהוצגה במאמר *Attention Is All You Need* משנת 2017 [1]. הטרנספורמר הינו מודל sequence-to-sequence שאומן על משימות תרגום מאנגלית לצרפתית / גרמנית, אולם מחברי המאמר מצינים את הפטנציאלי של המודל בפתרון בעיות נוספות כמו לדוגמה English constituency parsing.

המודלים שקדמו לטרנספורמר התבוססו על קובולוציה, RNN ומנגמוני קשב (Attention), ולעומתם הטרנספורמר משתמש אך ורק במנגמוני קשב ובכך מחליף לחוטין את השיטות האחרות. המודל שהוצע מושג תוצאות תרגום מדייקות יותר בעלות אימון נמוכה יותר באופן שימושי בהשוואה למודלים קודמים.

מאמר זה פרץ את הדרך למודלים השונים של בניית מלאכותית הנפוצים כיום והארכיטקטורה שהוא מציע עומדת בבסיס מודלים של [5][4][2][13], Computer Vision [3], Audio Analysis [1] ועוד. הסיבות לתופוצתו הרחבה של המודל הן הקביליות של הארכיטקטורה שמאפשרת אימוןiesel על שבבים מתאימים, האיכות של התוצאות שהוא מניב והיכולת להתאים את המודל למטרות רבות ומגוונות.

ארכיטקטורה

מבנה הטרנספורמר הוא של מקודד-מפענה (Encoder-Decoder) שמקבל קלט ומחזיר פלט. הסימboleים המתקבלים כקלט מקודדים לייצוג סדרתי, כך שהקידודлокח בחשבון גם את המיקום של תת המחרוזת ביחס לקלט. כל פלט שנוצר מתווסף כקלט לצורך חישוב המשך הפלט, וכך שהמשך הפלט מושפע מהקלט וגם מהפלט שקדם לו.

איור 1 : ארכיטקטורת הטרנספורמר [1].

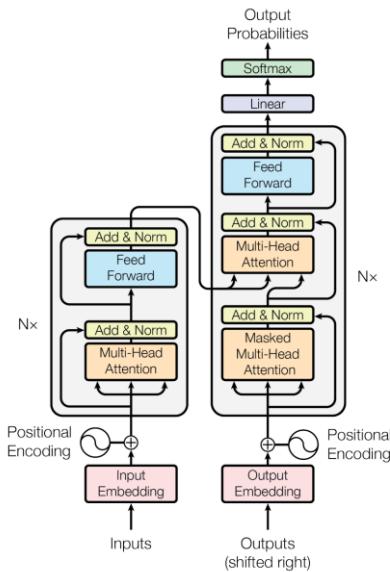


Figure 1: The Transformer - model architecture.

מקודד: המקודד בניו מ-6 שכבות זהות שמבצעות את חישוביין במקביל, כל שכבה מחולקת לשתי תתי-שכבות: Position-wise Feed-Forward Networks ו- Multi-Head Attention.

מפענה: המפענה מורכב בדיאק כמו המקודד ובנוסף מתווסף לכל שכבה תחת שכבה שלישית שמבצעת Multi-Head Attention על הפלט שיצא מהמקודד. בנוסף, ממבצעות בו התאמות על מנת שהתחזיות עבור המחרוזת ה-*o* ישפנו אן ווק מהמחרוזות הידועות שמוקמות לפניו ולא ישפנו מהתחזיות לגבי המחרוזות הבאות.

Attention

את מנגנון ה- Attention ניתן להציג כפונקציה שמקבלת סדרה של וקטורים שמהווה ייצוג חדש לסדרת הקלט, ייצוג שבו בכל וקטור מקודדים גם הקשרים בין בין כל שאר איברי הסדרה.

משוואת 1: הטרנספורמר כפונקציה [1]

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

את הפונקציה מתארים במאמר באמצעות המטריצות **Query**, **Key**, **Value**. המטריצות הללו קשורות לסדרת הקלט כך שכל וקטור מיוצג בשורה מתאימה לו בכל מטריצה. המכפלת Qk^T תהיה מטריצה שבה התא j, i יל את מידת הרלוונטיות של וקטור הקלט x לוקטור הקלט i . על מנת להתמודד עם בעיית הגרדיינט הנעלם מחלקים את תוצאות המטריצה בשורש של d_k שהוא הימד של וקטוריים במטריצות Q ו K. לאחר מכן מבצעים על התוצאה את פונקציית ה-softmax בצדיו לשווות לתוצאה ממך הסתברותי.

לאחר מכן המטריצה שהתקבלת תוכפל במטריצה ה- V כך שהוקטורים במטריצה הפלט יהיו תוצאה ממושקלת של V לפי המשקלים ההסתברותיים שחושבו בתהילך. באופן זה, תתקבל סדרה חדשה של וקטורים כך שבכל וקטור יקיים גם הקשר בין שני חברי. הכנוי שnitן במאמר לוריאנט זה של Attention הוא:

"Scaled Dot-Product Attention"

Multi-Head Attention

על מנת לחשב את החומרה Attention علينا להתחשב בכל המרכיב של המודל. אולם במקומות מסוימים שכלל וקטור יהיה במד של מרכיב הפרמטרים d_{model} ונctrיך לחשב מכפלת מטריצות גדולה מאוד, המודל מבצע הטלה של הוקטורים למרחבים ממשיים ממידת קטן יותר ובמקומות מכפלת מטריצות אחת שמבצעת את החישוב כולל, המודל מפצל את החישוב למספר חישובים שמתבצעים במדת קטן יותר. באופן זה, המודל יכול לבצע חישובים קטנים במקביל ולשרר את התוצאות שמתќבות. כך מתבצע החישוב בצורה מהירה יותר תוך שימוש בעלי במשאים.

Position-wise Feed-Forward Networks

כל שכבה של Encoder או Decoder מכילה תת- שכבה של Position-wise Feed-Forward Networks. שכבה זו מעבירה כל טוקן בנפרד ובאופן זהה דרך רשת נירונים בצד לדיק את הייצוג של הוקטור, וכך שיטמון בתוכו משמעות רחבה שנלמדה על ידי המודל בשלב האימון. נסביר את התהילך לפי המשוואה שהוצגה במאמר:

משוואת 2: [1] Feed Forward Network

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

עבור וקטור x שמייצג טוקן מסוים, מבצעים פעולה כפלי בין בין מטריצה W שמייצגת טוקן שנלמד בתוך המודל. בשלב זה נוצר וקטור במד גודל יותר מהمد המוקורי. לוקטור זה מתבצע חיבור של ערך b_1 שמייצג הטיה שנלמדה לאורך האימון. לאחר מכן מתבצעת הפונקציה ReLU שמאפסת ערכים קטנים מ-0 שמשמעותם בוקטור. בשלב הבא מוכפלת התוצאה במטריצה W שמעבירה את הוקטור חזרה לממד המוקורי שלו ומתווסף וקטור הטיה b_2 .

טהילך זה הינו זהה בין כל הטוקנים באותה שכבה, אולם בין שכבה לשכבה מטריצות המשקלים וvectורי הטיה משתנים.

Positional Encoding

마חר והטרנספורמר אינו מכיל התייחסות לסדר הפנימי של המחרוזות, علينا להוסיף לו רכיב שמקודד לכל וקטור המיצג טוקן מתוך הקלט את מיקומו בתחום הקלט.

משוואה 3: [1] Positional Encoding

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$

הfonקציית הzzאת ממחזירה לכל pozיציה וקטור מממד זהה לזה של הוקטור המיצג את הטוקן, באופן הבא:
עבור הממד הוא מוחזר הערך של הפונקציה הראשונה אם זוגי והערך של הפונקציה השנייה אם ערכו של זוגי. כך יוצאה שוקטור התוצאה הוא מהצורה (בහינתן שהממד של הוקטור זוגי):

משוואה 4: המחתת הייצוג של וקטור Position Embedding (על בסיס משוואה 3).

$$PE(\text{pos}) = \left(\sin\left(\frac{\text{pos}}{10000^{0/d_{\text{model}}}}\right), \cos\left(\frac{\text{pos}}{10000^{0/d_{\text{model}}}}\right), \sin\left(\frac{\text{pos}}{10000^{2/d_{\text{model}}}}\right), \cos\left(\frac{\text{pos}}{10000^{2/d_{\text{model}}}}\right), \sin\left(\frac{\text{pos}}{10000^{4/d_{\text{model}}}}\right), \cos\left(\frac{\text{pos}}{10000^{4/d_{\text{model}}}}\right), \dots, \sin\left(\frac{\text{pos}}{10000^{(d_{\text{model}}-2)/d_{\text{model}}}}\right), \cos\left(\frac{\text{pos}}{10000^{(d_{\text{model}}-2)/d_{\text{model}}}}\right) \right)$$

וקטור זה מתווסף לוקטור שמייצג את הטוקן ובכך מתווסף לייצוג של הטוקן גם מיקומו.

סיכום

פרק זה הציגנו באופן תמציתי את עיקרי החלקים המרכיבים את ארכיטקטורת הטרנספורמר. היכולת של הטרנספורמר לקודד בצורה טובה ובאופן יעיל את הקשרים המתקיים בין החלקים השונים של הרץ המעובד היוויטה חידוש בתחום עיבוד שפה טبيعית שהרואה לראשונה שניתן לבסס מודלי שפה טבעית על מנגןני קשב בלבד. בעוד המאמר ביצע ניסויים על משימות תרגום משפה לשפה, מחקרים שבוצעו אחרים הראו שניתן למשם מודלי שפה גדולים ורחבים על בסיס ארכיטקטורה זו וייתר מכך, ניתן למשם מודלים מבוססי טרנספורמר עבור מגוון רחב של תחומים ולקבל תוצאות מעולות. בפרק הבא נציג מאמר שמאמש את הטרנספורמר עבור ראייה ממוחשבת מבל' לבצע כמעט שום שינוי בארכיטקטורת המקורית כפי שהוצגה בפרק זה.

2. טרנספורמרים - ראייה ממוחשבת

הקדמה

בפרק זה נבצע סקירה של המאמר AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE משנת 2021 [2]. המאמר מציע להשתמש בארכיטקטורה של הטרנספורמר עבור מודלים של ראייה ממוחשבת, זאת בגין מודלים קודמים שהסתמכו על CNN גם אם תוך שילוב מסוים של מנגנוני Attention. לפי אמר זה, המודל המוצע במאמר, ViT, מביב תוצאות טובות יותר ממודלים עכשוויים עבור סט נתונים גדול אולם מטיביו של הטרנספורמר הוא מניב תוצאות טובות ממודלים מבוססי CNN עבור סט נתונים קטן שכן לטרנספורמר אין את התכונות האינדוקטיביות שיש CNN. אולם, עם הבעיה מגיעה הפתרון שכן הטרנספורמר מסוגל לאמן מודלים בגודל עצום ולהפיק תוצאות מעולות.

ניסויות קודמים למッシュ טרנספורמר עבור ראייה ממוחשבת

כאמור, עבור עיבוד שפה טבעית הטרנספורמר מפצל את המחרוזות לטוקנים שלאחר מכון מקבילים ייצוג וקטורי. לאחר מכן מחשב מכפלות פנימיות של הוקטורים על מנת לקודד את הקשרים שבין הטוקנים. חישובים אלה מתבצעים בעלות חשיבות ריבועית בגודל הקלט.

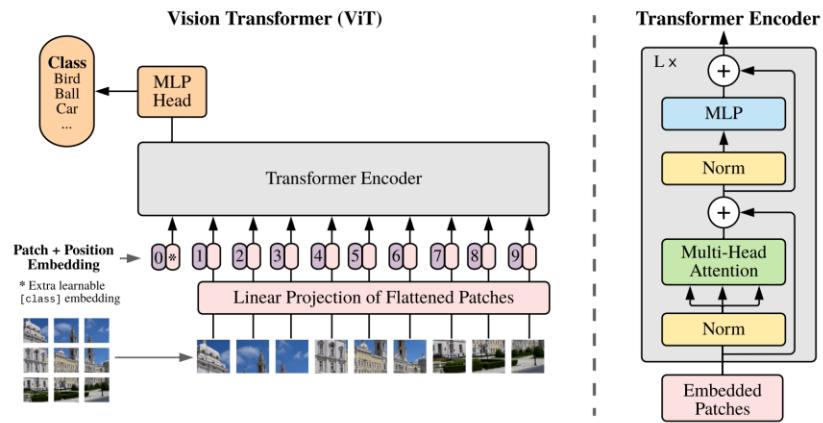
כאשר ניגשים למשם את הטרנספורמר עבור ראייה ממוחשבת עולה השאלה כיצד לייצג את התמונה בתוך המודל. גישה נאיבית תפצל את התמונה לפיקסלים השונים שלה ותחשב את החסכנות בין כל פיקסל לכל שאר הפיקסלים שבתמונה. אולם, כמות הפיקסלים בתמונה והעובדת שפונקציית החסכנות היא ריבועית, הופכת את המשימה לבליי אפשרית עבור תמונה קלה בגודל ריאלי. עקב כך הוצעו גישות שונות להtamודד עם הבעיה. גישה אחת מציעה לבצע Attention רק בתחום אזור מסוים בתמונה, כך שכל פיקסל מחשב רק את הקשר בין פיקסלים אחרים אזור. גישה נוספת מציעה לבצע Attention בין כמות משתנה של פיקסלים, כך שבמקרים קיצוניים ייחסבו עבור פיקסל מסוים רק קשרים בודדים. עם זאת, על אף שגישות אלה מקטינות משמעותית את כמות החישובים תוך הצגת ביצועים מבטחים, הן דוחשות הנדסה מיוחדת של החומרה על מנת לתמוך בפונקציות המיוחדות שכן דורשות. גישה קרובה מאוד לגישה של מאמר זה בחרה לפצל את התמונה לחתיות של 2×2 פיקסלים אולם גישה זו מוגבלת לרוחולזיה נמוכה בעוד הגישה שמצגת במאמר רלוונטיות גם לרוחולזיהBINONIOT. בנוסף המאמר מגדים ביצועים תחרותיים אל מול מודלים מבוססי CNN שהיו הטוביים ביותר עד למאמר זה.

ViT

המאמר מציג מודל מבוססי טרנספורמר שמיועד לבצע משימות של זיהוי תמונות (Image Recognition), אולם שואף להוות בסיס לימוש של מודלים עבור משימות ראייה ממוחשבת נוספות כגון: Image Detection, Image Segmentation ו-Image Segmentation. על מנת למשם זאת, הגישה המוצגת במאמר היא להציגם כמה שייתור למודל המקורי של הטרנספורמר כדי לאפשר שימוש במודלי הטרנספורמר הקיימים ובמערכות הייעילים שלהם. כאמור בפרק 1, הטרנספורמר הסטנדרטי מקבל טקסט כקלט ומפצל אותו לחתיות קטנות שבטורן מייצגות על ידי וקטורים. כתע נראה כיצד ניתן לישם הליך הטעעה דומה עבור תמונות.

המאמר מציע לחלק את תמונה הקלט לחתיות של 16×16 פיקסלים, לשטח את הייצוג שלהן לייצוג חד מימדי, ולאחר מכן לבצע הטלה לינארית באמצעות מטריצה E שנלמדה מראש למרחבumiMD S, שזהו המימד בו יתבצעו כל פעולות הטרנספורמר משלב זה והלאה. בנוסף, מתווסף טוקן [class] לתחילה הקלט לצורך סיוג המחלקה המתאימה לתאר את תוכן הקלט. כמו כן, במהלך ההטעה של התמונה למרחב הייצוג של הקלט מוטמע גם ייצוג המיקום (Positional Embedding) עבור כל אחת מהחתיכות של התמונה. מחברי המאמר מצאו שייצוג דו-מימדי של מיקום החתיכה בתוך התמונה אינו עדיף על ייצוג חד מימדי ולכן סידרו את החתיות באופן סדרתי חד-מימדי. בשלב זה לכל החתיות של התמונה יש ייצוג חד מימדי שמקורו לעיבוד על ידי הטרנספורמר באופן זהה לצורה שבה הטרנספורמר פועל עבור עיבוד שפה טבעית. איור 2 ממחיש את מבנה המודל.

איור 2: ארכיטקטורת ViT, המACHINE [2].



יתרונ בולט של מודל זה לעומת מודלים מבוססי CNN הוא כי בעוד שCNN מבצע הנקודות אינדוקטיביות על הקטלט, המודל הנ"ל כמעט ואין מחלוקת בדבר בכךו לקלט.

מחברי המאמר מציעים לשלב בתהיל' הנטמעה של הקטלט שימוש CNN כדי לקודד את החתיכות השונות של התמונה לפי התכונות שזוהו. אולם, המאמר מוכיח כי ניתן להתבסס באופן בלעדי על Attention לצורך הטעמעה הקטלט ולכן שילוב זה הוא מיותר.

Fine-tuning

כאשר ניגש לאמן מודל ViT נאמן אותו תחיליה על datasets גדולים ולאחר מכן מכתאים את המודל למשימה הספציפית על ידי Fine-Tuning שמתבצע על דатаה סט קטן וממוקד יותר, באופן הבא: ראשית נסיר את Prediction Head שנוצר במהלך האימון המקורי (השכבה האחורה בראשת הנוירונים שמוציאה כקלט את הקלאסIFIKAציה) ונחליף אותה בשכבה חדשה בגודל $K \times D$ כאשר D הוא הממד של ה-Patch Embeddings ו- K הוא מספר המחלקות השונות של המשימה הספציפית. שכבה זו תאוחת כליה לפחות כדי שהקלאסIFIKAציה תלמוד מהدادה סט החדש מבלי הטיה.

לעתים קרובות כדאי לבצע את הטיב על תמונות ברזולוציה גבוהה יותר כדי להשיג תוצאות מדויקות יותר. אולם, במצב זה מתעוררת בעיה. החלוקה לחטיכות של 16×16 פיקסלים בתמונה ברזולוציה גבוהה תגרור יותר חתיכות לכל תמונה ביחס לתמונה ברזולוציה נמוכה ובגודל זהה. עובדה זאת ככלעצמה אינה בעייתית שכן המודל יודע להתמודד עם קלטים באורך משתנה, אך הבעיה נועוצה בזכות Positional Embedding, שכן המיקום המקורי בתמונות ברזולוציות שונות הוא בעל משמעויות אחרות. לכן צריך לבצע המרה של קידוד המיקומים של החטיכות באימון המקדים למיקום האמתי שלהם בתמונה.

המקומות היחידים בהם המודל מניה הנקודות על המבנה הדו-忞ידי של התמונות הם במהלך החלקה לחטיכות ובמהלך ההתאמנה בין הרזולוציות השונות. מלבד בשתי נקודות אלה המודל אינו מניה הנקודות על הקטלט, זאת בגין מודלים מבוססי CNN שמניחים שניtin לחץ תכונות לפי אזורים שונים בתמונה, מה שמהווה הטיה גודלה מאוד על הקטלט.

ניסויים והשוואות

על מנת לבדוק את האפקטיביות של המודל, ביצעו ניסויים השוואתיים בין וריאנטים שונים שלוי לבין מודלים מבוססי CNN שהיו עד אותה עת *State of the art*. ההשוואות בוצעו לפי משימות שונות ודטה סטים שונים כאשר המטרה היא להשוות בין התוצאות לפי רמת דיוק ועולות חישובית כאשר אומנו על אותו>Data.

טבלה 1: השוואת בין מודלי ראייה ממוחשבת SOTA ל-DiT על מבחני קלואסיפיקציה פופולריים [2].

	Ours-JFT (ViT-H/14)	Ours-JFT (ViT-L/16)	Ours-121k (ViT-L/16)	BiT-L (ResNet152x4)	Noisy Student (EfficientNet-L2)
ImageNet	88.55 ± 0.04	87.76 ± 0.03	85.30 ± 0.02	87.54 ± 0.02	88.4/88.5*
ImageNet ReaL	90.72 ± 0.05	90.54 ± 0.03	88.62 ± 0.05	90.54	90.55
CIFAR-10	99.50 ± 0.06	99.42 ± 0.03	99.15 ± 0.03	99.37 ± 0.06	—
CIFAR-100	94.55 ± 0.04	93.90 ± 0.05	93.25 ± 0.05	93.51 ± 0.08	—
Oxford-IIIT Pets	97.56 ± 0.03	97.32 ± 0.11	94.67 ± 0.15	96.62 ± 0.23	—
Oxford Flowers-102	99.68 ± 0.02	99.74 ± 0.00	99.61 ± 0.02	99.63 ± 0.03	—
VTAB (19 tasks)	77.63 ± 0.23	76.28 ± 0.46	72.72 ± 0.21	76.29 ± 1.70	—
TPUv3-core-days	2.5k	0.68k	0.23k	9.9k	12.3k

Table 2: Comparison with state of the art on popular image classification benchmarks. We report mean and standard deviation of the accuracies, averaged over three fine-tuning runs. Vision Transformer models pre-trained on the JFT-300M dataset outperform ResNet-based baselines on all datasets, while taking substantially less computational resources to pre-train. ViT pre-trained on the smaller public ImageNet-21k dataset performs well too. *Slightly improved 88.5% result reported in Touvron et al. (2020).

כפי שניתן לראות בטבלה 1, הוריאנטים של DiT מציגים תוצאות טובות יותר משל מודלים State Of The Art שקדמו להם על benchmarks שונים בעוד שללות האימון שלהם מבחינת ימי עבודה של ליבת מעבד TPUv3 היא נמוכה יותר באופן משמעותי.

סיכום

בפרק זה הצגנו את מודל ה-Vision Transformer שמשמש את הטרנספורמר עבור ראייה ממוחשבת כמעת ללא שינוי. מכורח עלויות החישוב הגבוהות הכרוכות בחישוב Self-Attention בין כל פיקסל לפיקסל בתמונה, במאמר הוצע לחלק כל תמונה לחתי-כטוטות של 16×16 ולממש את מנגנון הקשב בין החתי-כטוטות השונות. באופן זה, עלויות החישוב יורדות באופן דרמטי והשימוש בטרנספורמר כמות שהוא הופך ליריאלי. תוצאות הניסויים שבוצעו עבור מאמר זה עקפו את המודלים הקודמים לו ב מבחנים שונים והדגימו את האפקטיביות של הטרנספורמר גם עבור שימושות של ראייה ממוחשבת. מאמר זה סלל את הדרך לשימוש רחוב במודל הטרנספורמר עבור שימושות ראייה ממוחשבת ועומד בבסיסם של מודלים רבים גם כיום. הפרק הבא מציג מודל שימושי מבוסס DiT המשמש לסגןנטציה של צילומים רפואיים.

Segment anything in medical images .3

הקדמה

סגמנטציה של צילומים רפואיים היא חלק חשוב בעבודתם של רופאים לצורכי אבחון, תכנון הטיפול וניתוח של המחלות. בפרק זה נציג מודל שמשתמש בארכיטקטורת הטרנספורמר ובאופן ספציפי יותר- ב-DiVi שהוצע בפרק הקודם, לצורך בניית מודל אוניברסלי שמייעל סגמנטציה של תמונות רפואיות מסווגים שונים, קר שנייתן להשתמש במודל זה לצורך סגמנטציה של צלומי CT, רנטגן, אנדוסקופיה ועוד של חלקים שונים בגוף האדם. המודל, המכונה MedSAM, אומן על מאגר גדול של צילומים רפואיים ומכיל 1,570,263 זוגות של תמונה והסגמנטציה שלה מכסה 10 סוגים שונים של הדמויות רפואיות ו-30 סוגים שונים של סרטן.

בעוד מודלים קודמים שנעודו לצורך סגמנטציה של צילומים רפואיים מתאימים לצורך שימוש מסוימת ומתוקשים בעת ביצוע משימה אחרת או בעת שימוש בסוג אחר של הדמיה, מודלים לsegueנטציה של תמונות טבעיות ביצעו התקדמות רבה ומוסוגלים לבצע משימות מגוונות בהצלחה. אולם, המודלים שמציגים תוצאות טובות עבור תמונות טבעיות מציגים תוצאות מוגבלות כשותבבקשים לבצע סגמנטציה של צילומים רפואיים עקב השוני הגדול ביניהם. במאמר זה מנוטים המחברים להציג כיצד דומה עבור צילומים רפואיים.

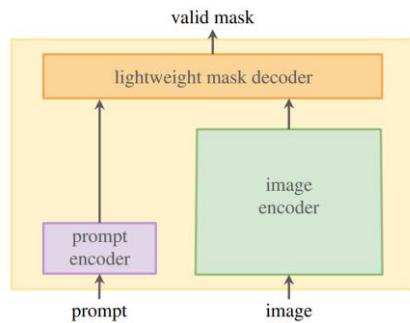
סגמנטציה אוטומטית לחלוּtin היא שאיפה גדולה אך לא פרקטית עבור צילומים רפואיים. סיבה אחת לכך היא השוני בין המשימות הנדרשות. לדוגמה, במקרים של צילום CT של סרטן כבד, רופא עלול לדרוש במקרה אחד סגמנטציה של הגוף בלבד ובמקרה אחר לסמן את הכבד כלו ואת האיברים סביבו, כתלות במצב הקלייני של המטופל. בנוסף, בין סוג ההדמויות השונות קיימים הבדלים המקיימים על סגמנטציה אוטומטית. למשל הדמויות מסוימות מייצרות תמונות תלת ממד ואחרות מייצרות תמונות דו ממד. בהתאם לכך מחברי המאמר לבסו את המודל שלהם על המודל SAM שמבצע סגמנטציה חצי-אוטומטית באופן שמתאים יותר עבור דרישות המודל.

SAM

SAM או Segment Anything Model [4] הוא מודל שהוצע על ידי חוקרים מקבוצת Meta AI Research, שמטרתו להוות foundation model עבור סגמנטציה של תמונות. המודל אומן על דата סט גדול ורחב ונועד לבצע סגמנטציה לפי prompt שהתקבל מהמשתמש בצורה של טקסט, סימן נקודה או מלבן בתמונה.

הארQUITקטורה של SAM מכילה encoder על image encoder שմבוסס על encoder של DiVi ומטרתו להטמע את התמונה בתוך מרחב וקטורי מממד גבוה. בנוסף ישנו רכיב הנקרא prompt encoder שמטמע את prompt שהתקבל ורכיב בשם mask decoder אשר מקבל את הפליטים של image encoder וה-mask decoder (ובנוסף מקבל output token שנלמד מראש על ידי המודל כדי לדיק את התוצאות). ה-mask decoder לוקח את הפליטים שקיבל, מבצע עליהם cross-attention כדי להתייחס אותם יחידה אחת ומחזיר את mask שמהווה את הסגמנטציה הרצiosa (ראה איור 3).

איור 3: ארכיטקטורת SAM [4].



MedSAM

בעוד SAM מציג תוצאות טובות על תמונות "טבעיות", המודל מתקשה לבצע סגמנטציה אינטואטיבית על צילומים רפואיים. במקביל למאמר זה ניסו חוקרים אחרים למשם SAM כモות שהוא על צילומים רפואיים. ניסיונות אלו העלו תוצאות טובות על צילומים בהם הgebenות בין החלקים של התמונהבולטים, ותוצאות פחות טובות על תמונות בהם הgebenות היו מוחכמים פחות. המודל שמצוג במאמר, MedSAM מציג ביצועים משופרים תוך ביצוע Fine-tuning של מודל SAM על דатаה סט גדול ורחב של צילומים רפואיים.

SAM תומך במספר סוגי פרוימטטים (טקסט, נקודה, מלבן)อลם עבור צילומים רפואיים מחברי המאמר העדיף לאפשר רק סימון מלבן ("bbox"). לטענותם, סימון נקודה עלול להפיק מספר סימונים שונים והוא פחות חד-משמעות מסימון מלבני. בנוסף, הסימון המלבני הוא עיל יותר לשיטתם, במיוחד כאשר מסמנים מספר עצמים בתמונה.

ארQUITקטורה

MedSAM עוקב אחר המבנה הארכיטקטורי של SAM ומכל image encoder, prompt encoder ו- mask encoder. בכך לאזן בין יעילות חישובית לבין ביצועי הסגמנטציה המודל משתמש בגרסת הבסיסית של ViT (ViT-base) שכן הגרסאות הגדולות יותר (ViT Large, ViT Huge) נותנות תוצאות משופרות כמעט אך יעילות הרבה פחות מבחינה חישובית.

לטרנספורמר הבסיסי 12 שכבות כל אחת מהן מורכבת מ multi-head self-attention ו- MLP block משולבים על ידי normalization layer. האימון המקדים הzbוצע על ידי fully-supervised modeling (MAE) [14] ולאחריו הzbוצע self-supervised modeling. MAE היא שיטה לאימון מודלים של ראייה מוחשבת באופן של self-supervised. ראשית המודל מכסה אחוז מסוים מהחטויות של התמונה ולאחר מכן הוא מסנה להשלים אותן בלבד. באופן זה המודל לומד את מבנה התמונות באופן עצמאי ולא השגחה ומשפר את עצמו. לאחר מכן, מאמנים את המודל בהשגחה מלאה באמצעות המاجر העצום של SAM שמכל תמונות ותווות של סגמנטציה.

image encoder מקודד את התמונות באופן הבא: כל תמונה בגודל מקורי של 1024×1024 מוחולקת לחתייכות של 16×16 קר שישן 64×64 חתיכות שונות שבתוון מעובדות לייצוג וקטורי על ידי encoder מה שיציר ייצוג חסכווי יותר מאשר 1024×1024 .

prompt encoder מקודד את המלבן המתחום (bounding box) באמצעות הנקודות העליונה-שמאלית והתחתונה-ימנית שמקובלות ייצוג וקטורי בן 256 ממדים כל אחת קר שזוג הווקטורים מייצג את המלבן.

mask decoder בנייתן על מנת לאפשר אינטראקציה בזמן אמת עם המשמש. ראשית, מתבצעת הטמעה חד פעמית של התמונה על ידי image encoder. כאשר ישמן מלבן על התמונה יבצע prompt encoder יוצרת הטמעה של המלבן ואז mask decoder מייצר ייצוג מאוחדר של התמונה והמלבן באמצעות שני בלוקים של טרנספורמר ושתי שכבות של transposed convolution כדי לשפר את הרזולוציה.

השוואת ביצועים

מאמר זה מציג שיפורים בדיקות של סגמנטציה של צילומים רפואיים לפי מספר מבחנים והשוואות. השוואות שבוצעו על בסיס median DSC score בין המודלים DeepLabV3+, U-net, SAM, MedSAM כאשר U- net ו-

אומנו לצורכי המשימות הספציפיות שנבדקו, מצאו כי גם בבדיקות **external validation** וגם בבדיקות **internal validation** מציג את הביצועים הכי פחות טובים מלבד עבור תמונות צבעוניות בהן הוא מציג ביצועים לא רעים. נמצא גם, כי החזון של תוצאות DSC של MedSAM במחנים השונים הוא הגבוה ביותר וגם התפלגות התוצאות שלו היא הקטנה ביותר, מה שמחיש את יכולת השם של המודל להציג תוצאות טובות יותר ממודלים אחרים אף שלא אומן ספציפית עבור כל סוג שנבדק, ועוקף את המודלים - **U-Net** ו-**DeepLabV3+** על אף שאומנו באופן נפרד עבור כל משימה שנמדדה.

איור 4: הערכה כמותית ואיכותית של MedSAM. [5]

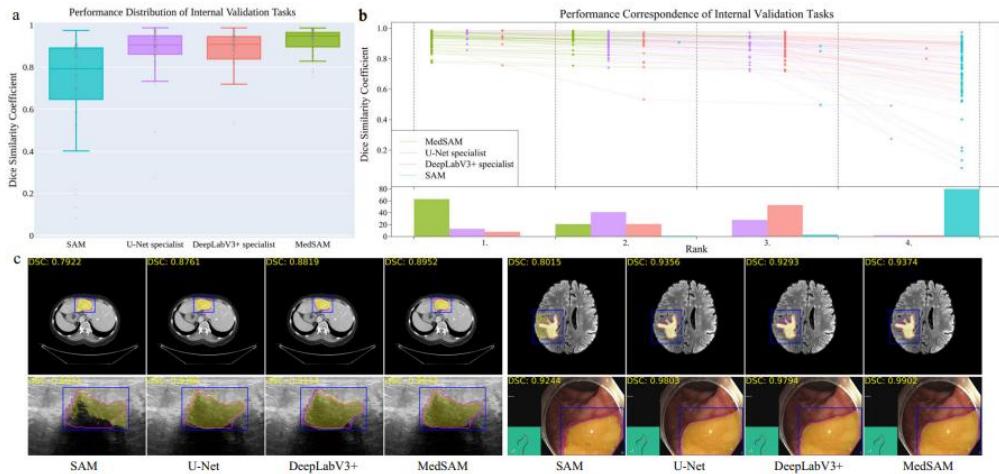


Fig. 3 | Quantitative and qualitative evaluation results on the internal validation set. **a** Performance distribution of 86 internal validation tasks in terms of median dice similarity coefficient (DSC) score. The center line within the box represents the median value, with the bottom and top bounds of the box delineating the 25th and 75th percentiles, respectively. Whiskers are chosen to show the 1.5 of the interquartile range. Up-triangles denote the minima and down-triangles denote the maxima. **b** Podium plots for visualizing the performance correspondence of 86 internal validation tasks. Upper part: each colored dot denotes the median DSC achieved with the respective method on one task. Dots corresponding to identical tasks are connected by a line. Lower part: bar charts represent the frequency of achieved ranks for each method. MedSAM ranks in the first place on most tasks. **c** Visualized segmentation examples on the internal validation set. The four examples are liver cancer, brain cancer, breast cancer, and polyp in computed tomography (CT), (Magnetic Resonance Imaging) MRI, ultrasound, and endoscopy images, respectively. Blue: bounding box prompts; Yellow: segmentation results. Magenta: expert annotations. Source data are provided as a Source Data file.

corresponding to identical tasks are connected by a line. Lower part: bar charts represent the frequency of achieved ranks for each method. MedSAM ranks in the first place on most tasks. **c** Visualized segmentation examples on the internal validation set. The four examples are liver cancer, brain cancer, breast cancer, and polyp in computed tomography (CT), (Magnetic Resonance Imaging) MRI, ultrasound, and endoscopy images, respectively. Blue: bounding box prompts; Yellow: segmentation results. Magenta: expert annotations. Source data are provided as a Source Data file.

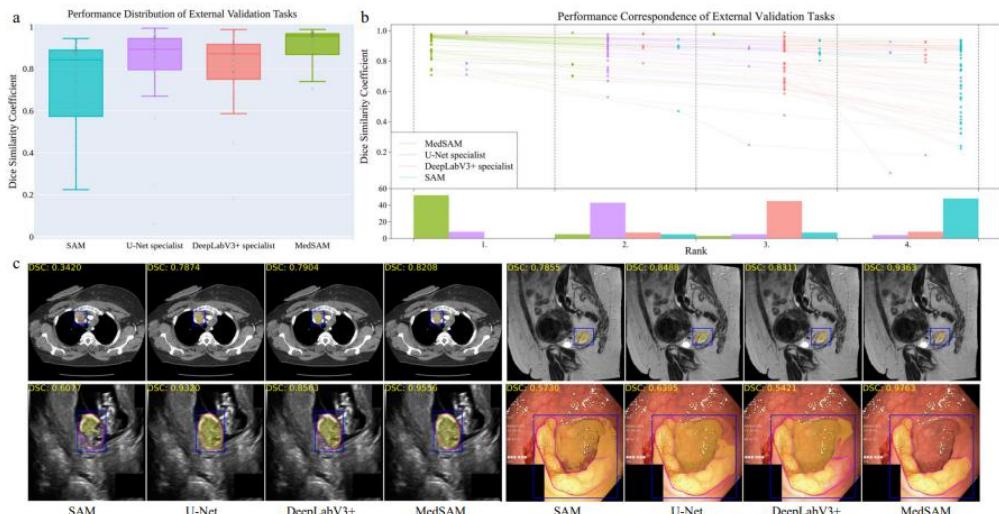


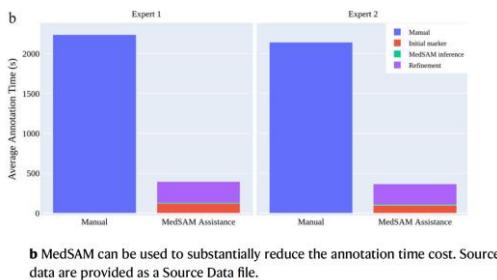
Fig. 4 | Quantitative and qualitative evaluation results on the external validation set. **a** Performance distribution of 60 external validation tasks in terms of median dice similarity coefficient (DSC) score. The center line within the box represents the median value, with the bottom and top bounds of the box delineating the 25th and 75th percentiles, respectively. Whiskers are chosen to show the 1.5 of the interquartile range. Up-triangles denote the minima and down-triangles denote the maxima. **b** Podium plots for visualizing the performance correspondence of 60 external validation tasks. Upper part: each colored dot

denotes the median DSC achieved with the respective method on one task. Dots corresponding to identical tasks are connected by a line. Lower part: bar charts represent the frequency of achieved ranks for each method. MedSAM ranks in the first place on most tasks. **c** Visualized segmentation examples on the external validation set. The four examples are the lymph node, cervical cancer, fetal head, and polyp in CT, MR, ultrasound, and endoscopy images, respectively. Source data are provided as a Source Data file.

יעול העבודה הרפואי

בנוסף, על מנת להראות את פוטנציאל הייעול של העבודה של הרופא בוצע מען ניסוי. הניסוי מدد את משך הזמן שנדרש עבור שני רדיולוגים מנוסים לסמן חתיכות של גידולים בצלומי CT תוך שימוש בMedSAM וכמה זמן נדרש עבורם לבצע זאת באופן ידני לחלוויין. הצלומים שנבחרו אינם חלק מקבוצת האימון של המודל אך שהניסוי מדגים את היעילות של המודל בסגמנטציה של תМОנות שאיננו מכיר מראש. הניסוי ארך שני שלבים. בשלב הראשון נתנו לשני הרופאים בנפרד לסמן באופן ידני את גידולי יתרת הכליה על השכבות השונות של הצלומים התלת-ממדיים והזמן שההתהילך לוקח להם נמדד ונשמר. בשלב השני, לאחר תקופת צינון של שבוע, הרופאים סימנו באמצעות צירים את הקצחות של הגידול מהשכבה העליונה לתחתונה. התהילך זה בוצע פעמי 3 עד 10 שכבות של הצלום. לאחר מכן MedSAM ביצע סגמנטציה על בסיס הסימונים המעתים הללו באופן הבא: עבור שכבה מסוימת יסומנו מלבדים שחווסמים את הצירים שסומנו. עבור שכבה שלא סומנה תבצע אינטראפלציה לינארית שתסמן מלבדים על בסיס השכבות הסמוכות שסומנו. לאחר מכן MedSAM ייקח את התמונות והסימונים ויבצע סגמנטציות לתמונות על בסיס הסימונים. לבסוף, יעברו המומחים על התוצאות ויתקנו אותן.

איור 5: יעול העבודה הרפואי על ידי MedSAM [5].



התוצאות הראו הפחתה של 82.37% - 82.95% בזמן הנדרש עבור שני המומחים בהתאם לכך לבצע את הסגמנטציה הרצויה.

סיכום

המודל MedSAM מהווה כלי לבחוק סגמנטציה של צילומים רפואיים מסוגים שונים וחקלים אוטומטיים מגוונים מבלי לדרש אימון ספציפי לכל שימוש. הכללי משלב בין אוטומציה לבין התאמת אישית ומפחית את זמן העבודה מבלי לוותר על הדיקוק. המודל מעלת על מודלים כליליים שהיו עד כה SOTA ואף מציג ביצועים תחרותיים ולעיתים עדיפים על מודלים המתמחים במשימות ספציפיות. MedSAM מציג פרדיוגרמיה מוכחת לאימוץ מודלים של תמונות טבעיות עבור תמונות שאין טבעיות כמו לדוגמה סגמנטציה של תאים בתמונות ממיקרוסקופ או סגמנטציה של אברונים בתמונה של מיקרוסקופ אלקטרוניים.

אולם המודל מוגבל. בעיה אחת שלו היא חוסר האיזון בין סוג הצלומים הרפואיים שעולם להטאות את התוצאות לפי הנלמד מסוג הצלומים השכיחים יותר מה שעלול לפגום בפונCTION של תצלומים מסווגים פחות שכיח כמו מוגרפיה. אולם, לאחר והמודל השיג הבנה רחבה מຕור הדאטה עליה אומן יהיה ניתן לשפר את הביצועים שלו בנקודות החולשה שלו בקלות על ידי fine-tuning.

MAMBA .4

הקדמה

מודלים גדולים המאמנים מראש על DATA סט גדול במיוחד ומתקנים מותאמים לביצוע משימות רבות ומגוונות, המכונים Foundation Models, מציגים ביום תוצאות טובות. מודלים אלה מבוססים בעיקר על ארכיטקטורת הטרנספורמר ושבבת הself-attention שלהם, המיקום שלהם ועוד. אולם, לארכיטקטורה זו ליצר DATA מרכיב המציג את הטוקנים השונים, הקשרים ביניהם, המיקום שלהם ועוד. אולם, לארכיטקטורת context window סופי, כך שקיימת מגבלה על אורך הסדרה שניתן לבצע עליה self-attention. בנוסף, כל חישוב בתחום context window גודל כלשהו, מתרחש בסיבוכיות ריבועית בגודל context window. מגבלות אלו הובילו לניסויות רבות למשת את מבנה self-attention בצורה ייעילה יותר אך ניסויות אלו באו על חישוב האפקטיביות של המנגנון ועד כה לא נמצא מימוש אפקטיבי במידה שווה ויעיל יותר מהקיים.

Mamba: Linear-Time Sequence Modeling with Selective State Spaces [6] לריעונות אחרים לבסיס Foundation Models ייעילים יותר ואפקטיבייםstructured state space sequence models באופן שווה ואףלו יותר. קודם למאמר התפרסמו מודלים מסוג (S4) שמצילים לייצג קשרים ארכוי טוח בסיבוכיות לנארית או קרובה לכך והציגו תוצאות מבטיחות במספר תחומיים. מודלים אלה מצילים בתחוםים שמתקיים בהם רציפות כגון אודיו וראיה, אך מתקשים בתחוםים בידיהם כמו טקסט. המחברים מציגים במאמר ארכיטקטורה משופרת בשם Selective State Space model שבסיסת על S4 אך מוסיפה מנגנון בחירה המאפשר למודל להתמקד במידע מסוים בתחום הרצף ובכך להניב תוצאות משופרות. ראשית נציג בקצרה מעט רקע למאמר.

S4

בשנת 2022 פרסמו אלברט גו ושותפים מאמר [8] בו הם מציגים אפשרות לימוש יעיל של SSM לצורך בניית sequence models שמסוגלים לבצע מגוון משימות מנעד רחב של תחומי הדורשים עיבוד של תליות ארכוי טוח. הארכיטקטורה שהוצעה מכונה S4 Structured State Space Sequence Models.

SSM מוגדר לפי המשוואות הבאות [8]:

משוואת 5: המשוואות הבסיסיות של SSM [8].

$$\begin{aligned} x'(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned}$$

כאשר x הוא המצב החבוי בנקודת הזמן הנוכחי $(-t)$, x הוא המצב החבוי בנקודת הזמן הבאה בתור), (t) n הקולט בנקודת הזמן הנוכחי, $A B C D$ הם פרמטרים שנלמדו על ידי Gradient descent ו- y (t) הוא הפלט הנגזר בנקודת הזמן הנוכחי.

S4 מציג מספר שיטות ביחס לSSM פשוטים ובכך מאפשר לשימוש ברעיון של SSM עבור מודלים גדולים, מה שאינו ריאלי עבור רשתות RNN פשוטות. ראשית, הרעיון המרכזי של S4 הוא לאפשר חישוב מקבילי של sequence ובכך לעקוף את המגבילות של רשתות RNN סטנדרטיות, כמפורט במשוואות 6-7:

משוואת 6-7: S4 בתצורת קונבולוציה [8].

$$\begin{aligned} y_k &= \overline{CA}^k \overline{B} u_0 + \overline{CA}^{k-1} \overline{B} u_1 + \cdots + \overline{CAB} u_{k-1} + \overline{CB} u_k \\ y &= \overline{K} * u. \end{aligned} \tag{6}$$

$$\overline{K} \in \mathbb{R}^L := \mathcal{K}_L(\overline{A}, \overline{B}, \overline{C}) := \left(\overline{CA}^i \overline{B} \right)_{i \in [L]} = (\overline{CB}, \overline{CAB}, \dots, \overline{CA}^{L-1} \overline{B}). \tag{7}$$

משמעות אלה מתארות את \bar{y} לפי מכפלת מטריצות ידועה מראש בהינתן \bar{K} (שיטת החישוב של \bar{K} באופן עיל היא לא פשוטה מבחינה טכנית וmoצגת בהרחבה במאמר[8]). בכך, מתאפשר אימון מקבילי וביסוס מודול גדול על ארכיטקטורה זו באמצעות GPU הופך לרייאלי, בניוגד לרשותה RNN פשוטות שלא ניתן לאימון מקבילי.

בנוסף, S4 נועד בשיטת העבודה של HiPPO [7] במהלך המימוש של המטריצות של ה- SSM על מנת לתת משקל מועט לעבר ומשקל גדול יותר למצבים האחרונים ולהפיק תוצאות מדויקות יותר.

MAMBA

בעוד S4 מתגבר על בעיית המקבילות של RNN ומישם יכולת המאפשרת להוסיף את הדיכרונו לפי זמן העדכון, קיימת מגבלה נוספת שמנענת ממודלים כדוגמת S4 להציג תוצאות תחרותיות. מנגנון השכחה של המודלים אינו תליי קלט אלא קבוע עבור כל המודל, מגבלה שగורעת מהיכולת של המודל לזכור את המידע הרלוונטי לקלט ולשכוח את המידע שאינו רלוונטי. משום כך, מחברי המאמר [6] משנת 2023 בנו מנגנון בחירה ("Selection Mechanism") שמעצב את המטריצות של SSM לפי הקלט ובכך מאפשר בחירה התוליה גם בקלט ולא רק במודל הנלמד. אולם, שנייה זה עשוי לגרום אליו עלויות חישוב גבוהות מבחינה תיאוריתית ופרקטיות כאחד, שכן מודלי SSM קודמים הציגו עלויות כמעט לינאריות רק תודות לכך שהו "יעיריים" לקלט ולזמן. עובדה זו הציבה אתגר למחברי המאמר, אותו הם פתרו באמצעות אלגוריתם מותאם לחומרה.

מנגנון הבחירה

מנגנון הבחירה הוא רעיון רחב שנitinן למשמש עבור רשותות RNN ו- CNN סטנדרטיות, ובעבור פרמטרים שונים של הארכיטקטורה וכן ניתן למשמש אותו באמצעות טרנספורמציות שונות. לדוגמה, מנגנון הבחירה של MAMBA מוסיף את העדכנים של המצב החבוי לפי מידת הרלוונטיות של הקלט באופן הבא: מידת הרלוונטיות של הקלט במצב הנוכחי נלמדת על ידי המודל. המודל מחשב את המטריצה \bar{A} כתלות בקלט הנוכחי. \bar{A} מעדכנת את הפיטרים A-B כך שלאחר דיסקרטיזציה \bar{A} ו- \bar{B} יסנו את הקלט בהתאם למידת הרלוונטיות שלו.

נפרט מספר השפעות של מנגנון הבחירה של MAMBA:

סלקטיביות מאפשרת לסנן טוקנים של רושע שעולים להופיע בין טוקנים רלוונטיים. טוקנים כאלה מופיעים לעיתים קרובות בסוגי DATA שונים, לדוגמה השימוש בהברות כמו "אה..." במהלך דיבור. הסלקטיביות מאפשרת לסנן לחלוטין הברות כאלו עבר קטיעי אודיו של דיבור, ולחסוך את המקום שבו דרושת בתוך המצב החבוי.

השפעה חיובית נוספת של מנגנון הבחירה היא האפשרות לבחור את הקונטקט הרלוונטי עבור הקלט. כבר הבחינו [9] של מרבית העיקרונות הסוברים כי הגדלת חלון ההקשר משפרת את התוצאות, נמצא כי מודלים רבים לא השתפרו בעת שהגדילו את חלון ההקשר. הסיבה יכולה להיות מסוימת שלעויות יש צורך להעתם מkonceptstein מסוים בתוך החלון ומודלים אלה אינם יודעים לבצע זאת. אולם, מנגנון הבחירה של מבנה מאפשר למודל לאפס את הקונטקט לפי הצורך ובאופן זה להקשר הרלוונטי.

בנוסף, כאשר מספר רצפים שאינם קשורים אחדו לכדי רצף אחד, מודלים כמו SSM אינם מסוגלים לבצע את הפרדה ועלולים לאבד לחלוטין מידע רלוונטי עקב הציגו לרצפים הקודמים. לעומת זאת, מנגנון הבחירה מאפשר לאפס את המצב וכך לשמר את כל המידע הרלוונטי עבור הרצף הנוכחי.

ນזין CUT את המשמעות של הוספת מנגנון הבחירה לפרמטרים השונים:

הפרש של \bar{A} : דلتא מוסיף עד כמה להתקדם בקלט הנוכחי ועד כמה להעתם ממנו. ערך שואף לאינסוף של דلتא יגרום לכך שההיסטוריה תשכח כמעט לחלוטין וה המצב יושפע באופן כמעט בלעדי מהקלט הנוכחי. ערך אפסי של דلتא יגרום לכך שהקלט הנוכחי כמעט כmut ולא ישנה את המצב.

הפרש של A: על אף SCN להוסיף סלקטיביות לא עצמוני, מאוחר והמודל משתמש בא רק דרך מפגש עם דلتא המחברים מניחים כי הוספת סלקטיביות לא לא תשיג שיפור משמעותי.

הפרש של B ו- C: הוספת מנגנון בחירה לא תווסת את ההשפעה של הקלט על המצב החבוי והוספת מנגנון זה ל- C תווסת את השפעת המצב החבוי על הפלט.

התאמת לחומרה

כאמור, הוספת מנגנון הבחירה גורמת לכך שאBAB MAMBA Linear Time Invariance איננו מימושים היעילים של SSM קודמים שדורשים AT&T אינם רלוונטיים עבור מודל זה. לעומת זו דרושת חישיבה מחודשת לגבי אופן המימוש של מודל המשלב מנגנון זה.

לצורך חישיבה זו הציגו המחברים הבדיקה מעניינת. הם הבוחינו כי לאחר והסיבוכיות החישובית של RNN היא $O(BLDN)$ עם מוקדם קבוע נמוך ושל CNN היא $O(BLD \log(L))$ עם מוקדם קבוע גבוה אליו בהינתן שמדד המצב N נמוך מספיק, RNN ידרשו פחות פעולות חישוב. לכן, מבחינה תיאוריתנית ניתן למשם מנגנון בחירה עבור SSM (בדומה ל-RNN) באופן יעיל לא פחות ואף יותר מאשר מימוש בצורה של קונבולוציה.

אולם, קיימים קשיים טכניים בימוש הרעיון הזהה. ראשית, התלות בזמן ובקלט לא מאפשר שימוש במשאים באופן מקבלי על מנת להאיץ את החישוב. בנוסף, שבירת המטריציות תלויות הקלט עבור כל שלב בקלט דורשת זיכרון רב שמאט את תהליך החישוב ולא מאפשר לבצע את החישוב ברמות המהירות של המעבד בהם קיימים זיכרון מצומצם מאוד.

לפתרון הבעיה השנייה הציעו המחברים פתרון משולב שמותאם למעדים מודרניים. ראשית, המצב יוחשב רק בرمות היעילות יותר של המעבד תוך כדי תנועה. כך יחסכו קריאות וכתיות יקרות בין החלקים של המעבד. בנוסף, רוב הפעולות מוגבלות על ידי הזמן של העברת המידע בין חלק המעבד ולא על ידי זמן החישוב עצמו, שכן ביצוע פעולות רבות באותו רמה של המעבד ייחסוק זמן רב. למשל, במקרה לביצוע DISCRETIZATION של הפרמטרים של SSM בrama האיטית ולהעביר את התוצאה לחישוב בrama מהירה, נעביר את הפרמטרים כמו שהם לrama מהירה ונבצע את החישובים שם בצורה יעליה יותר. בנוסף לכך, מבצעים kernel fusion כך שמספר פעולות הופכות לפעולה אחת ונחסך זמן קריאה וכתייה רב, למשל על ידי התכה של פעולות חישוב ופעולות כתיבה לזכרון לכדי פעולה אחת.

לפתרון בעיית המקבילות, הציעו להשתמש באלגוריתם סריקה במקום בחישוב סדרתי שכן על אף שהישוב recurrence באופן סדרתי איננו בהכרח לינארי וכן ניתן להשתמש בטכניקות של מודלים שהם Linear Time Invariant, מאמריהם קודמים הראו שניתן לחשב אותו באמצעות סריקה מהירה על ידי שימוש בטכניקות של scan , prefix sum, parallel scan , למשל המאמר [15].

לבסוף, נדרש שלא לשמר את כל מצב הביניים של רשף הנירונים, על אף שהם נדרשים לצורך חישוב backpropagation, שכן הם דורשים מקום רב שלא מאפשר לבצע את החישובים בرمות המהירות של המעבד. את זאת ממשים באמצעות חישוב מחדש שמתבצע במהלך הbackward כאשר הקלט עבר מהrama האיטית בrama מהירה. כתוצאה לכך זיכרון של שכבת הסריקה הסלקטיבית זהות לדרישות של הטרנספורמר האופטימלי שמנוהש באמצעות FlashAttention. איור 6 ממחיש את הפעולות שנעשות מעבר במצב הבא. המצבים החבויים מחושבים במלואם רק בrama מהירה של המעבד שמסומנת בכתבום [6].

איור 6: אילוסטרציה של מנגנון הבחירה עם חישובים ברמות החומרה השונות [6].

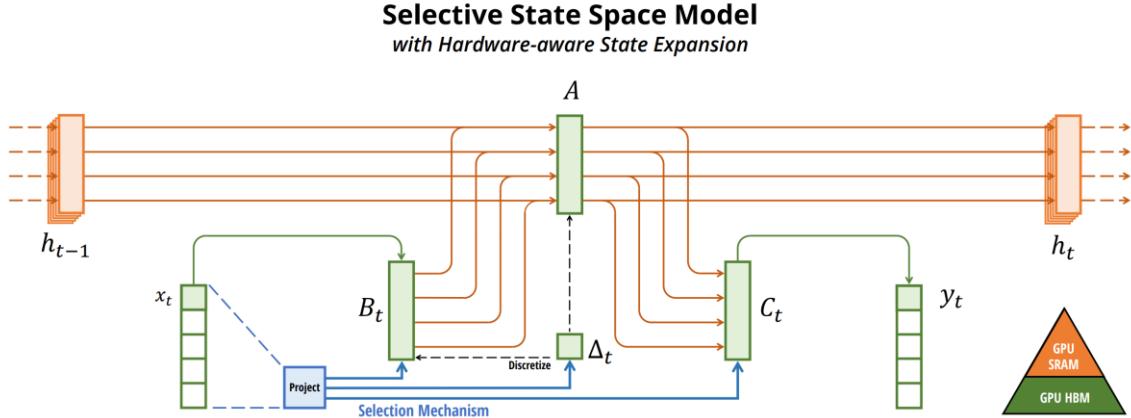


Figure 1: (Overview.) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN , times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the (Δ, A, B, C) parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

בלוק ה-MAMBA

בלוק הממבה בנוי באופן הבא:

ראשית הקלט משוכפל לשני עותקים, נסמן z, x . x עובר הטלה לינארית ולאחר מכן דרך דרכ קונבולוציה, ואז דרך פונקציית אקטיבציה ויוצר את $'x$. בשלב זה $'x$ עובר דרך מנגנון ה-SSM הסלקטיבי המתואר בפסקאות הקודמות וייצר את $'z$. במקביל, z עובר הטלה לינארית ואז דרך פונקציית אקטיבציה וייצר את $'z$. כעת, $'x$ ו- $'z$ עוברים דרך מנגנון שער (Gating Mechanism) בו נקבע כמה מהקלט המקורי לשמר וכמה לשנות. לבסוף התוצאה עוברת הטלה לינארית שייצרת פלט מודד של הקלט שנכנס לבלוק הממבה. איור 7 ממחיש את מבנה הבלוק הממבה.

איור 7: בלוק ה-Mamba [6]

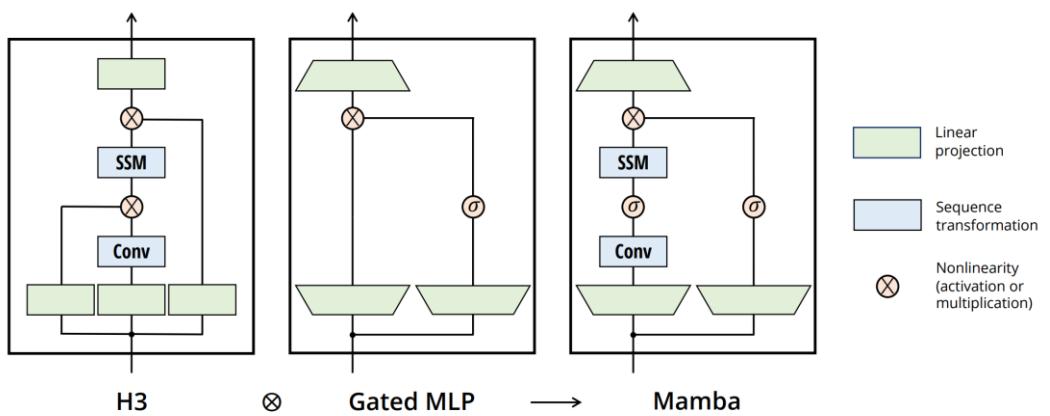


Figure 3: (Architecture.) Our simplified block design combines the H3 block, which is the basis of most SSM architectures, with the ubiquitous MLP block of modern neural networks. Instead of interleaving these two blocks, we simply repeat the Mamba block homogeneously. Compared to the H3 block, Mamba replaces the first multiplicative gate with an activation function. Compared to the MLP block, Mamba adds an SSM to the main branch. For σ we use the SiLU / Swish activation (Hendrycks and Gimpel 2016; Ramachandran, Zoph, and Quoc V Le 2017).

ניסויים אמפיריים

לצורך הערכת יכולות של המודל בוצעו מספר ניסויים והשוואות, בפרק זה נציג חלק מהם.

העתקה סלקטיבית

מבחן סינטטי נפוץ עבור מודלים הוא מבחן העתקה, בו נבדק אם המודל מצליח להעתיק טקסט מתוך הקלט. אולם, מבחן זה ניתן לפתרה בקלות על ידי זכירה של רצפים באורך העתקה. לכן, נוצר מבחן העתקה הסלקטיבית, בו נדרש המודל להעתיק טקסט עם דילוגים אקראיים. ניתן לראות בטבלה 2 כי מנגנון הבחירה מאפשר לבצע העתקה צזו בצורה מיטבית כאשר משלבים אותו בתוך ארכיטקטורות שונות, ובפרט כאשר הוא משולב בתוך בלוק MAMBA (S6 הוא S4 בשילוב מנגנון הבחירה):

טבלה 2: מבחן העתקה סלקטיבית [6].

MODEL	ARCH.	LAYER	Acc.
S4	No gate	S4	18.3
-	No gate	S6	97.0
H3	H3	S4	57.0
Hyena	H3	Hyena	30.1
-	H3	S6	99.7
-	Mamba	S4	56.4
-	Mamba	Hyena	28.4
Mamba	Mamba	S6	99.8

Table 1: (**Selective Copying**)
Accuracy for combinations of architectures and inner sequence layers.

DNA Modeling

עבור המאמר, נעשו מספר ניסויים שונים לבדוק את הביצועים שմבנה מציגה עבור מידול DNA. ראשית נבדק השיפור של מבנה בעת הגדלת מספר הפרמטרים של המודל ובעת הגדלת חלון ההקשר של המודל. הניסוי הראה כי מבנה משתפרת בשני המקרים עם העלייה בגודל, זאת בניגוד למתחמות כפי שנitinן לראות בגרף הבא [6]:

איור 8: השפעת גודל המודל ואורך הרץ על ביצועי מודלים במשימת ניבוי DNA [6].

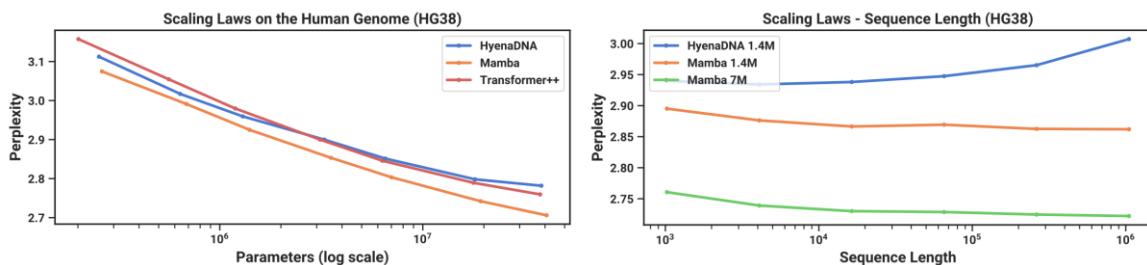


Figure 5: (**DNA Scaling Laws**.) Pretraining on the HG38 (human genome) dataset. (*Left*) Fixing short context length $2^{10} = 1024$ and increasing size from $\approx 200K$ to $\approx 40M$ parameters, Mamba scales better than baselines. (*Right*) Fixing model size and increasing sequence lengths while keeping tokens/batch and total training tokens fixed. Unlike baselines, the selection mechanism of Mamba facilitates better performance with increasing context length.

בנוסף נעשה ניסוי במשימות קלאסיפיקציה של DNA של קופי אדם שגם בו הציגה מבנה שיפור בביטויים בהשוואה למודל שלו הושוו בעת הגדלת חלון ההקשר ומספר הפרמטרים של המודל. יש לציין כי הניסויים של הגדלת חלון ההקשר לא הושוו מול מודל מבוסס טרנספורמר שכן שימוש בטרנספורמר עבור חלון ההקשר גדול כל כך אינו רלוונטי.

איור 9: קלאסייפיקציה של DNA של קוף אדם [6].

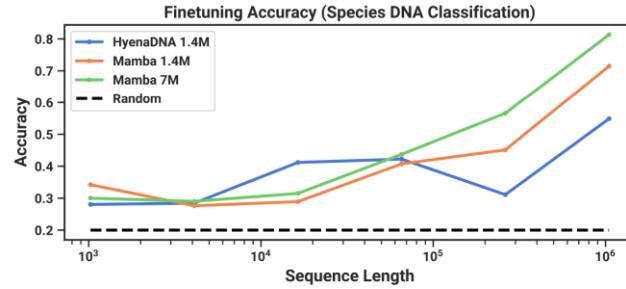


Figure 6: (**Great Apes DNA Classification.**) Accuracy after finetuning on sequences of length $2^{10} = 1024$ up to $2^{20} = 1048576$ using pretrained models of the same context length. Numerical results in Table 13.

Speed and Memory Benchmarks

המחקר בדק את הביצועים של מ מבנה במונחים של מהירות תפוקה במהלך inference ביחס לשוואות למודל טרנספורמר, ובמונחים של מהירות ביצוע פעולות החישוב המרכזית על קולט באורך משתנה, שהוא פועלות החיבור Scan, ומשוואה להיטון FlashAttention, קונבולוציה, ולפעולות ScanPyTorch של ספריית PyTorch. הביצועים מוצגים באיור 10.

איור 10: השוואת יעילות חישובים במהלך האימון ומהירות תפוקה במהלך Inference [6].

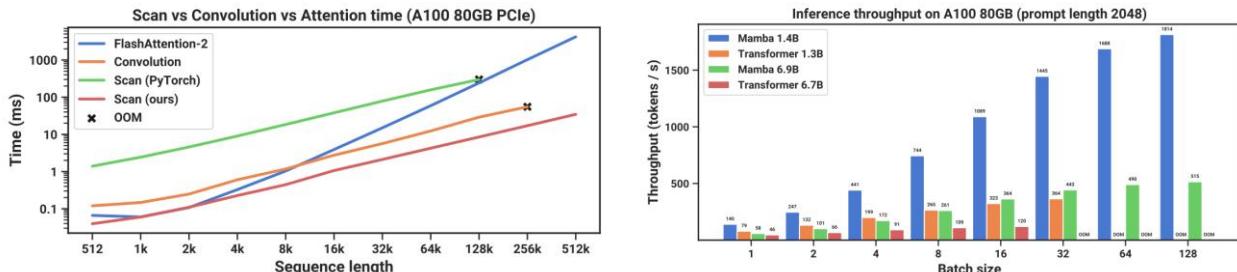


Figure 8: (**Efficiency Benchmarks.**) (*Left*) Training: our efficient scan is 40× faster than a standard implementation. (*Right*) Inference: as a recurrent model, Mamba can achieve 5× higher throughput than Transformers.

גם במונחים של שימוש יעיל בזיכרון מ מבנה דרישת כמות זיכרון גבוהה רק במעט על הטרנספורמרים האופטימליים ביותר כגון 38.2GB עבור size של 32 בהשוואה ל 34.5GB שהטרנספורמר דורש. הנתונים מוצגים בטבלה 3.

טבלה 3: השוואת דרישות זיכרון מ מבנה וטרנספורמר [6].

Table 15: (**Memory benchmark.**) Mamba’s memory footprint is comparable to the most optimized Transformer. Results for 125M models.

Batch size	Transformer (w/ FlashAttention-2)	Mamba
1	4.6GB	4.8GB
2	5.2GB	5.8GB
4	6.9GB	7.3GB
8	11.5GB	12.3GB
16	20.7GB	23.1GB
32	34.5GB	38.2GB

סיכום

ארכיטקטורת MAMBA שואפת להוות תחליף שנותן מענה לחולשות של הטרנספורמר, ובינהה- חלון הקשר מוגבל, סיבוכיות חישובית ריבועית ומגבילות בעת עיבוד מידע רציף. כפתרון לכך, פנו מחברי המאמר למודלים מבוססי SSM, והוסיפו להם מנגנון בחירה שמאפשר לשמר את המידע הרצוי ולסנן את המידע הלא רלוונטי. אולם מנגנון זה גורר קשיים למיושן יעל של המודל ולכן הוצגו במאמר התאמות שנעודו לנצל בצורה מיטבית את משאבי החישוב המודרניים ובכך להשיג יעילות חישובית גבוהה. המאמר מראה כי מJAVA משיגה תוצאות שימושיות ואף עוקפות את הטרנספורמר במגוון רחב של תחומיים ובמיוחד בתחוםים בהם נדרש קונטיקסט ארוך במיוחד כמו גנטיקה, ועודיו. הכותבים מאמינים כי MAMBA מסוגל להיות עמוד השדרה הגנרי של Sequence Models ולהחליף בכך את הטרנספורמר.

Vision Mamba .5.

הקדמה

ארכיטקטורת MAMBA מציגה תוצאות מבטיחות עבור מידול של רצפים ארכיטקטוניים כפי שצינו בפרק הקודם הקודם, אולם מימושה עבור משימות ראייה מוחשבת ניצב בפניו אתגר משמעותי. אתגר ראשון היא העבודה שמבנה מעבד את הקלט באופן סדרתי בעוד חלק התמונה השונים אינם יכולים קשר סדרתי אלא גלובל. אתגר שני הוא בכך שלמבנה אין מודעות למקום של כל טוקן בתוך הקלט, בעוד המיקום של כל פיקסל או חתיכה ביחס לתמונה יכולה להיות להבנת התמונה וחלקיה השונים. המאמר [10] שנמצא בפרק זה פותר את הבעיה הראשונה על ידי בLOC ממהה דו-כיווני שמעניק הקשר גלובלי לכל חלק של התמונה, ואת הבעיה השנייה על ידי הוספת Position Embedding לחلكי הקלט השונים.

המחקר הראה תוצאות משופרות בהשוואה למודל מבוסיס הטרנספורמר DeiT במספר מובנים: במונחי דיוון עבור משימות מגוונות, במונחי מהירות (פ' 2.8) וזיכרון (חישך 86.8% זיכרון מעבד). لكن, תוצאות המאמר מציגות פוטנציאל גדול לביסוס Foundation Models לראייה מוחשבת על בסיס MAMBA.

ארכיטקטורה

תיאור כללי

מודל המבנה הסטנדרטי עובד על רצפים חד ממדים, אך על מנת לעבד את התמונות מתבצעת באופן ראשוןי המירה של החתיכות השונות של התמונה לייצוג חד ממד. לאחר מכן מתבצעת הטלה לינארית לממד S של המצב החבוי ומתווסף קידוד שמייצג את המיקום של כל חתיכה בתוך התמונה (Position Embedding). עבור קלאסיפיקציה, בדומה לנעשה עבור BERT [3] וכשם שנעשה עבור ViT [2], מושגים טוקן מחלקה (Class Token) שמייצג את כל החתיכות. לאחר מכן המועובדים עברים ל-MLP והפלט שלו עובר לשכבה Vision Mamba Encoder נוספת וכן הלאה עד לשכבה האחורונה. לאחר מכן לוחכים את טוקן המחלקה שנפלט מהשכבה האחורונה, מבצעים עליון נורמליזציה ומעבריהם אותו דרך בLOC MLP שማפיק את היחסוי. איור 11 ממחיש את הארכיטקטורה המוצגת במאמר.

איור 11: ארכיטקטורת Vim [10].

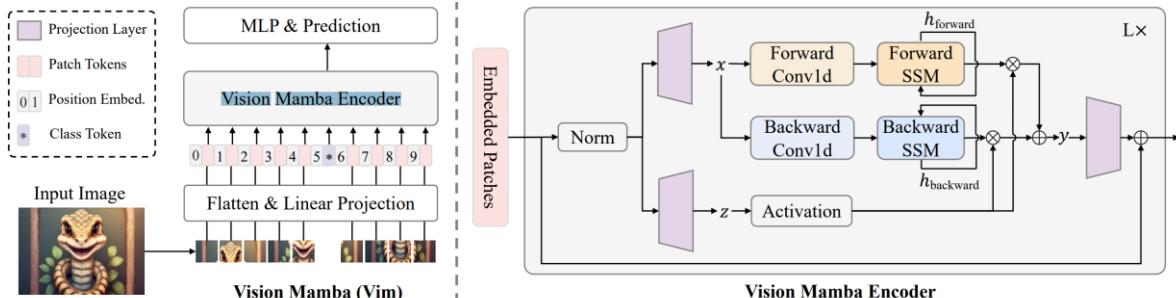


Figure 2: The overview of the proposed Vim model. We first split the input image into patches, and then project them into patch tokens. Last, we send the sequence of tokens to the proposed Vim encoder. To perform ImageNet classification, we concatenate an extra learnable classification token to the patch token sequence. Different from Mamba for text sequence modeling, Vim encoder processes the token sequence with both forward and backward directions.

בלוק ה-Vim

נתאר כעת את מהלך העבודה של בLOC המתוואר בצד ימין של איור 11. הקלט המוטמן (התמונה לאחר השטחה לחד-ממך, הטלה לינארית ל-S ממד והוספת Class token-Position Embedding) עובר נורמליזציה בצד לייצב את החישוב (מניעת היעלמות או התפוצצות של הגרדיאנט) ולאחר מכן עובר הטלה לממד E אל הוקטורים x ו-z. הוקטור x עובר קונבולוציה והופך ל x' . x' עובר הטלה למטריצות של B_0 , C_0 , Δ_0 . לאחר מכן מושמת לצורך יצירת המטריצות הסלקטיביות המותאמות לקלט: A_0 ו- \bar{A}_0 . לאחר

מן מתבצע החישוב של הפלט על ידי שימוש במטריצות שנוצרו בהתאם למנגנון של SSM. הפלט עובר מנגנון של Gating עם הוקטור z לאחר שעבר אקטיבציה. תהליך זה קורא עבור x ו- z עם ה-SSM שרך מההתחלת לסוף עם ה-SSM שרך מהסוף להתחלה. הפלט של שני היכיוונים עבר Gating והפלט של שער זה עבר הטילה לינארית חוזרת לממד D , עובר שער נוסף עם הקלט Embedded של תחילת השכבה הנוכחית של Vim block (skip connection) והפלט של תהליך זה עבר לשכבה הבאה של הבלוק.

ניסויים והשוואות

קלאסיפיקציה של תמונות

טבלה 4: השוואת Top-1 Accuracy ב-*Vim* לבני מודלים מתחרים [10].

Method	image size	#param.	ImageNet top-1 acc.
Convnets			
ResNet-18	224^2	12M	69.8
ResNet-50	224^2	25M	76.2
ResNet-101	224^2	45M	77.4
ResNet-152	224^2	60M	78.3
ResNeXt50-32 \times 4d	224^2	25M	77.6
RegNetY-4GF	224^2	21M	80.0
Transformers			
ViT-B/16	384^2	86M	77.9
ViT-L/16	384^2	307M	76.5
DeiT-Ti	224^2	6M	72.2
DeiT-S	224^2	22M	79.8
DeiT-B	224^2	86M	81.8
SSMs			
S4ND-ViT-B	224^2	89M	80.4
Vim-Ti	224^2	7M	76.1
Vim-Ti †	224^2	7M	78.3 <small>+2.2</small>
Vim-S	224^2	26M	80.3
Vim-S †	224^2	26M	81.4 <small>+1.1</small>
Vim-B	224^2	98M	81.9
Vim-B †	224^2	98M	83.2 <small>+1.3</small>

Table 1: Comparison with different backbones on ImageNet-1K validation set. † represents the model is fine-tuned with our long sequence setting.

טבלה 4 מציגה את תוצאות הניסוי במשימת קלאסיפיקציה של K-1-ImageNet. הניסוי הציג שיטת אימון דומה לכל האפשר עבר המודלים השונים ומצא כי המודלים מבוססי המ מבה הציגו תוצאות טובות יותר במנוחי -Top 1 accuracy מהמודלים המקבילים להם מבחינת גודל. בנוסף נמצא כי כאשר מודלי המ מבה עברו Fine-tuning עבר עיבוד רצפים ארוכים הם הציגו תוצאות טובות אף יותר כך שהמודול S-Vim הגיע לתוצאות קרובות למודל הגדל יותר באופן משמעותי DeiT-B.

טבלה 5: מבחן סגמנטציה סמנטית Vim ומודלים מתחרים [10].

סגמנטציה סמנטית

Method	Backbone	image size	#param.	val mIoU
DeepLab v3+	ResNet-101	512^2	63M	44.1
UperNet	ResNet-50	512^2	67M	41.2
UperNet	ResNet-101	512^2	86M	44.9
UperNet	DeiT-Ti	512^2	11M	39.2
UperNet	DeiT-S	512^2	43M	44.0
UperNet	Vim-Ti	512^2	13M	41.0
UperNet	Vim-S	512^2	46M	44.9

Table 2: Results of semantic segmentation on the ADE20K val set.

כפי שניתן לראות בטבלה 5 המודלים מבוססי Vim עוקפים את מודלי הטרנספורמר הדומים להם בגודל במונחי $mIoU$ (הຂיתור של החיזוי והסימון הנכון חלקו האחד שלהם) בכ- $0.9mIoU$, ואת מודלי Vim הם מושגים עם מודל בעל מספר פרמטרים נמוך במחצית.

זיהוי אובייקטים וסגמנטציה לפי מופע

טבלה 6: מבחן זיהוי אובייקטים וסגמנטציה לפי מופע Vim ומתחרים [10].

Backbone	AP ^{box}	AP ^{box} ₅₀	AP ^{box} ₇₅	AP ^{box} _s	AP ^{box} _m	AP ^{box} ₁
DeiT-Ti	44.4	63.0	47.8	26.1	47.4	61.8
Vim-Ti	45.7	63.9	49.6	26.1	49.0	63.2
Backbone	AP ^{mask}	AP ^{mask} ₅₀	AP ^{mask} ₇₅	AP ^{mask} _s	AP ^{mask} _m	AP ^{mask} ₁
DeiT-Ti	38.1	59.9	40.5	18.1	40.5	58.4
Vim-Ti	39.2	60.9	41.7	18.2	41.8	60.2

Table 3: Results of object detection and instance segmentation on the COCO *val* set using Cascade Mask R-CNN (Cai & Vasconcelos, 2019) framework.

טבלה 6 מדגימה את העדיפות של Ti-Ti Vim על Ti-Ti DeiT גם במנוחה סימון אובייקטים (בקופסה או בפיקסלים). עובדה נוספת שנובעת מהמחקר היא כי כאשר מדובר במשימות סימון של אובייקטים גדולים במיוחד, Vim-Ti עוקף את Ti-Ti DeiT במספרים משמעותיים עוד יותר, מה שבבליט את העדיפות של Vim-Ti במקרה בהם נדרש הקשר גודל יותר. איור 12 מדגימה את העובדה הזאת.

איור 12: השוואת ייזואלית של סגמנטציה של אובייקט גדול [10].



Figure 5: Visualization comparison of DeiT-Ti (Touvron et al., 2021b) and our Vim-Ti on the Cascade Mask R-CNN (Cai & Vasconcelos, 2019) framework. Thanks to the long-range context learning of SSM, we can capture the very large object in the image, which the DeiT-Ti counterpart fails to perceive.

זיכרון ומהירות

במנוחה מהירות ניתנת לראות כי מודל הממבה מציג ביצועים שווים ערך לטרנספורמר עבור תמונות רחולוציה של 512×512 אולם ככל שמדוברת הרחולוציה נעשים החישובים של המודל מבוססים הממבה מהירים יותר משימושותיתvr כר שבעבור רחולוציה של 1248 מציג המודל ביצועים מהירים יותר פי 2.8 כפי שניתן לראות באיור 13.

איור 13: השוואת מספר תמונות בשניה לפי רחולוציות בין Vim ל-DeiT [10].

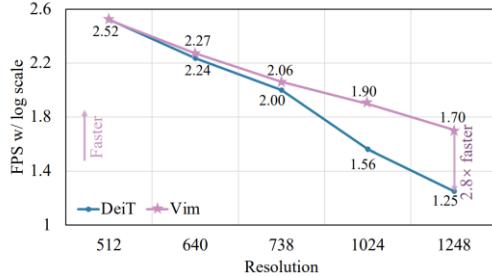


Figure 3: FPS comparison between DeiT-Ti (Touvron et al., 2021a) and our Vim-Ti on the commonly used downstream framework. We perform batch inference and benchmark the log-scaled FPS on the architecture with the backbone and FPN. Vim achieves comparable performance to DeiT with a small resolution, i.e., 512×512 . As the input image resolution increases, Vim has a higher FPS.

לסיוום נראה כי גם במנוחה שימוש בזיכרון המודל מבוסס המ מבנה הוא חסכווי יותר מהמודל מבוסס הטרנספורמר, חסכו שמתגבר עם הגברת הרזולוציה של התמונות המעובדות, כפי שניתן לראות באירור 14.

אייר 14: השוואת יעילות GPU בין Vim ל-[10]DeiT

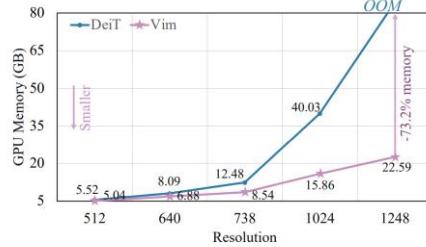


Figure 4: GPU memory efficiency comparison between DeiT-Ti (Touvron et al., 2021a) and our Vim-Ti on the commonly used downstream framework. We perform batch inference and benchmark the GPU memory on the architecture with the backbone and FPN. Vim requires comparable GPU memory to DeiT with a small resolution, i.e., 512×512. As the input image resolution increases, Vim will use significantly less GPU memory.

סיכום

בפרק זה סקרנו את המודל Vim הממש ארכיטקטורת MAMBA עבור ראייה ממוחשבת במגוון משימות על מנת להוות עמוד שדרה גנרי לראייה ממוחשבת. המודל הציג תוצאות עדיפות על מודלי הטרנספורמר שנבדקו במנוחי דיק, מהירות וזיכרון, והראה יכולת הסתגלות טוביה יותר עבור קלטים ארוכים הן במנוחי דיק והן במנוחי מהירות וזיכרון.

6. מ מבה עבר סגנטציה של צילומים רפואיים בתלת-ממד

הקדמה

בעשור האחרון חלו התקדמות משמעותית בשימוש במודלים מובוסים טרנספורמר למשימות סגנטציה של הדמויות רפואיות בתלת-ממד. עם זאת, החיסרון המרכזי של מודלים אלו טמון בעלות החישובית הגבוהה שלהם, הנובעת מהסיבוכיות הריבועית של מנגנון ה-self-attention. בפרק זה נסקור את מודל SegMamba [11] — גישה חדשנית לsegueנטציה של הדמויות תלת-ממד רפואיות, המבוססת על ארכיטקטורת Mamba. מכיוון שהדמויות תלת-ממד מורכבות מרצפים של פרוסות דו-ממדיות, יש צורך יכולת לłączן קשרים מרוחבים-דרציפים הנפרשים על פני שכבות רבות. ארכיטקטורת Mamba מאפשרת זאת באופן יעיל, באמצעות לכידת תלות ארוכת טווח ללא העומס החישובי הכרוך במודלים מסורתיים. כך מגדים המודל את יתרונותיה של Mamba ביצוע משימות מתוגרות, בהן ארכיטקטורות אחרות נתקלו בבעיות.

ארכיטקטורה

איור 15: המasha של ארכיטקטורת SegMamba.[11]

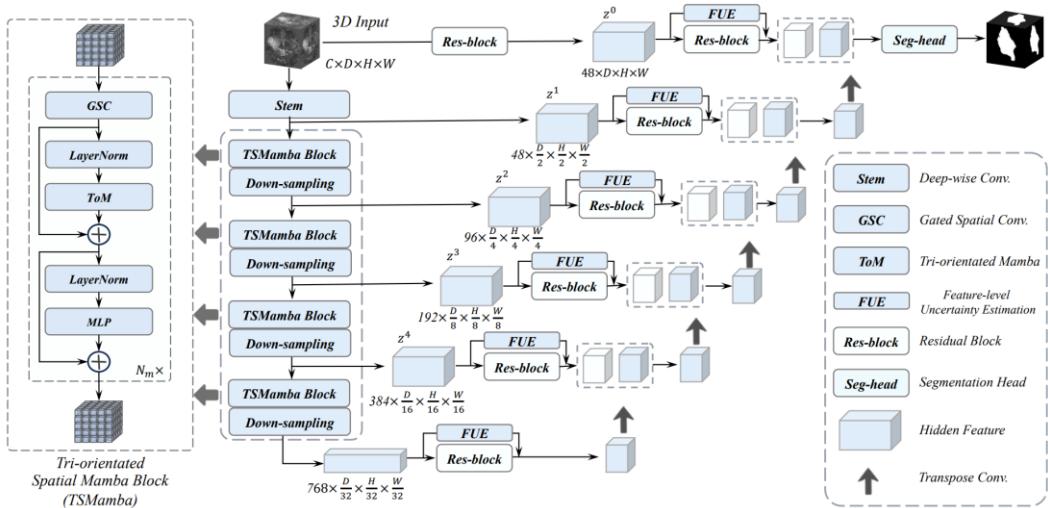


Fig. 1. An overview of the proposed SegMamba. The encoder comprises a stem layer and multiple TSMamba blocks designed to extract multi-scale features. Within each TSMamba block, a gated spatial convolution (GSC) module models the spatial features, and a tri-orientated Mamba (ToM) module represents global information from various directions. Furthermore, we develop a feature-level uncertainty estimation (FUE) module to filter multi-scale features, facilitating more robust feature reuse.

המודול מורכב בעיקרו משלושה חלקים. מקודד, מפונח ו-skip-connections. המקודד מורכב ראשית מבlok Stem שנועד לכוצץ ולזקק את הקלט. לאחר מכן הוא מורכב מספר בלוקים הנקראים TSMamba blocks. נפרט את המבנה של בלוקים אלו.

(בלוק ממבה מרחבית באוריינטציה) תלת-ממדית (Tri-oriented Spatial Mamba (TSMamba) Block

בעוד הטרנספורמר מצטיין בתפיסת הקשר גלובלי, הצלות של פעולה זו היא גבוהה מאוד לעומת אשר מדובר ברצפים ארוכים במיוחד, כמו במקרה של צילומי תלת-ממד רפואיים. לכן, שיטות מבוססות טרנספורמר כמו UNETR בחרו לכוצץ כל ממד של הקלט חלק 16. אולם, גישה זו מגבילה את יכולת לקולט תוכנות

הנפרשות על פני מספר שכבות של הצילום וכן תוצאות הסגמנטציה מוגבלות בהתאם. מטרת הבלוק היא לפיך, לתפוס קונטיקסט גלובלי ותכונות הפרשות על מספר תמונות ולקודד אותן בצורה מיטבית ובאופן יעיל מבחינה חישובית. תהליך החישוב של הבלוק ה- m מתואר לפי משווהה 8.

משווהה 8: TSMamba

$$\hat{z}_m^l = GSC(z_m^l), \quad \tilde{z}_m^l = ToM(\text{LN}(\hat{z}_m^l)) + \hat{z}_m^l, \quad z_m^{l+1} = \text{MLP}(\text{LN}(\tilde{z}_m^l)) + \tilde{z}_m^l$$

נפרט את הרכיבים השונים של הבלוק.

Gated Spatial Convolution (GSC)

אחר ו-MAMBA מודיע את התכונות של התמונה לאחר השטחה של הקלט מתלת-ממד לחד-ממד, עלולים להתפסס הקשרים שנפרשים על פני הממדים השונים. לכן, לפני שמתבצעת ההשטחה מבצעים קודם כל GSC.

בלוק GSC מורכב הבא: הקלט z עובר במקביל לשני בלוקים של קונבולוציה, האחד מבצע חילוץ תכונות תלת ממדיות לפי קבוצה של $3 \times 3 \times 3$ פיקסלים והשני מבצע חילוץ תכונות עבור פיקסל אחד. כל בלוק קונבולוציה מורכב משכבה נורמליזציה, קונבולוציה, ופונקציה לא-לינארית. הפלטים של שני הבלוקים מוכפלים וועורבים בלוק קונבולוציה של $3 \times 3 \times 3$ על מנת להתייר אותו לפט אחד. לאחר מכן מ被执行 חיבור של פט זה עם הקלט המקורי z על מנת להשתמש מחדש בתכונות של הקלט המקורי (skip connection). איור 16(a) ממחיש את המבנה המתואר של GSC.

איור 16: קונבולוציה מרחבית עם מנגן שער (GSC), בלוק ממבה עם אוריינטציה תלת-ממדית (ToM)

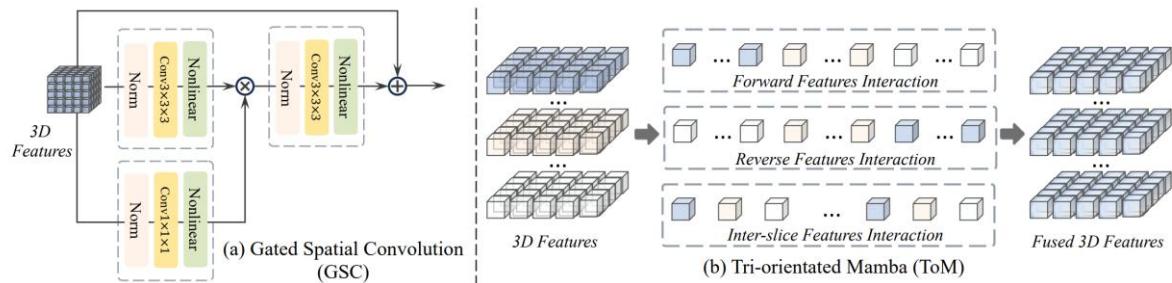


Fig. 2. (a) The gated spatial convolution. (b) The tri-orientated Mamba.

Tri-orientated Mamba (ToM)

הקשרים המתקיימים בין חלק הצלום התלת-ממדי מתקיימים לאורך מספר כיוונים וקשה לתפוס אותם במהלך סריקה חד-כיוונית, כמו שבЛОק המבנה הקלאסי עובד. לכן, בלוק הממבה המוצע במאמר זה משפט את התכונות התלת-ממדיות ל-3 רצפים: z_f , z_r , z_s (forward, reverse, inter-slice). כאשר כל אחד מהם יעבד על ידי בלוק ממבה מסוון ובסופו יבוצע חיבור של שלוש הרצפים כדי להשיג את כל התכונות שחולצו מהקשרים שמתקיימים בכיוונים השונים. איור 16(b) ממחישים את אופן הפעולה של בלוק ה- M .

משווהה 9: ToM [11]

$$ToM(z) = Mamba(z_f) + Mamba(z_r) + Mamba(z_s)$$

Feature-level Uncertainty Estimation (FUE)

הפלטים היוצאים מ-TSMamba מכילים הרבה תכונות שמידת הוודאות לגבין היא מוגבלת (לדוגמה: מידת הוודאות שאזורה מסוימת בתמונה הוא רקע). במאמר [11] מוצגת שכבה שמתווספת לחיבורים השינויים של המודול כדי להפחית את מידת ההשפעה של תכונות עם מידת ודאות נמוכה ולהגבר את ההשפעה של אלו שווודאותן גבוהה. החישוב מורכב כך: עבור מספר ערוצים C ממבצע לכל פיקסל ממוצע של הערכים בערוצים

השוניים כך: $\bar{z}^i = \sigma\left(\frac{1}{C^i} \sum_{c=1}^{C^i} z_c^i\right)$ ולאחר מכן מבצעים על התוצאה פונקציית אקטיבציה כך ש-
 $u_i = -\bar{z}_i \log(\bar{z}_i)$ כך ש- u הוא מידת אי-הווודאות של התוכנות של הפיקול
 כאשר ערך גבואה מצבע על מידת אי-הווודאות גבואה. לבסוף הפלט של FUE הוא:
 $(u^i - 1) \cdot z^i + z^i \cdot \tilde{z}^i = z^i + z^i \cdot (1 - u^i)$ כאשר z הוא הקלט שערכו קרוב לפעם הראשונה
 ואותו גבואה וערכו ערך הקלט אם ודאותו נמוכה. ראה משווהה 10.

משווהה 10: [11] FUE

$$u^i = -\bar{z}^i \log(\bar{z}^i), \text{ where } \bar{z}^i = \sigma\left(\frac{1}{C^i} \sum_{c=1}^{C^i} z_c^i\right). \quad (1)$$

$$\tilde{z}^i = z^i + z^i \cdot (1 - u^i) \quad (2)$$

ניסויים והשוואות

מאגר צילומי סרטן המעי הגס (CRC-500)

סרטן המעי הגס היוו השלישי בתפוצתו בקרב נשים וגברים כאחד, וגורם המות הגדול ביותר בקרב סרטני מערכת העיכול. למרות זאת הדטה סטימ הקיימים של סגמנטציה של סרטן המעי הגס הינם קטנים ומרובים פרטיאים. לכן אספו עבור מאמר זה מאגר של 500 צילומים של סרטן המעי הגס עם סימוני סגמנטציה שנוצרו כל אחד על ידי רופא מומחה ו עברו אימונות וטיבע על ידי רופא מומחה אחר. הצילומים הינם צילומי CT שנאספו בין השנים 2008-2020 ונוקה מהם כל מידע רפואי רגיש של המטופלים. איור 17 הוא דוגמה מתוך הדטה סט החדש.

איור 17: דוגמה ויזואלית של דатаה מאגר CRC-500.

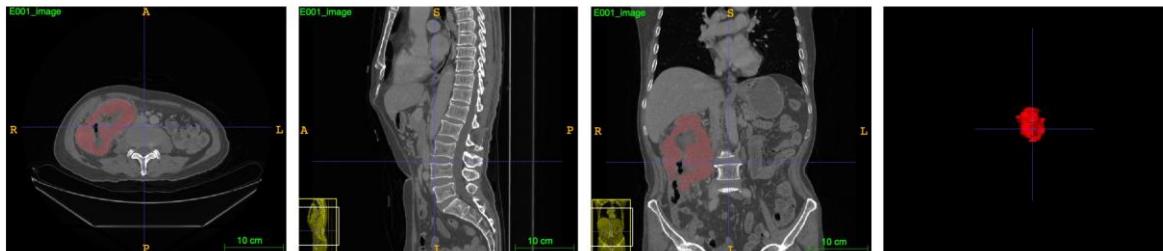


Fig. 3. The data visualization for CRC-500 dataset.

בנוסף למאגר החדש, נבדקו הביצועים של SegMamba בהשוואה למודלים אחרים על דטה סטם נוספים
 ולפי מספר מבחני סגמנטציה שונים. טבלה 7 מציגה את השוואות הביצועים של המודלים על הדטה סטם e
 BraTS2022 של המדימות MRI של גידולים במוח ו-2023 של Tumor Segmentation 2023 של דרכי אויר וטבלה 8 מציגה את
 ההשוואה על הדטה סט החדש.

טבלה 7: השוואת כמותית בין SegMamba למתחרים על BraTS2023-AIIB2023.

Table 2. Quantitative comparison on the BraTS2023 and AIIB2023 datasets. The bold value indicates the best performance.

Methods	BraTS2023								AIIB2023		
	WT		TC		ET		Avg		Airway Tree		
	Dice ↑ HD95 ↓	IoU ↑ DLR ↑ DBR ↑									
SegresNet [18]	92.02	4.07	89.10	4.08	83.66	3.88	88.26	4.01	87.49	65.07	53.91
UX-Net [11]	93.13	4.56	90.03	5.68	85.91	4.19	89.69	4.81	87.55	65.56	54.04
MedNeXt [20]	92.41	4.98	87.75	4.67	83.96	4.51	88.04	4.72	85.81	57.43	47.34
UNETR [6]	92.19	6.17	86.39	5.29	84.48	5.03	87.68	5.49	83.22	48.03	38.73
SwinUNETR [5]	92.71	5.22	87.79	4.42	84.21	4.48	88.23	4.70	87.11	63.31	52.15
SwinUNETR-V2 [7]	93.35	5.01	89.65	4.41	85.17	4.41	89.39	4.51	87.51	64.68	53.19
Our method	93.61	3.37	92.65	3.85	87.71	3.48	91.32	3.56	88.59	70.21	61.33

טבלה 8: השוואת כמותית בין SegMamba למתחרים על CRC-500.

Table 3. Quantitative comparison on the CRC-500 dataset.

Methods	Dice ↑ HD95 ↓	
SegresNet [18]	46.10	34.97
UX-Net [11]	45.73	49.73
MedNeXt [20]	35.93	52.54
UNETR [6]	33.70	61.51
SwinUNETR [5]	38.36	55.05
SwinUNETR-V2 [7]	41.76	58.05
Our method	48.46	28.52

ניתן לראות כי התוצאות שמציג המודל SegMamba עדיפות על המודלים האחרים על פני כל המבחנים שנוטו ולאורך כל הדadata טיטם. טבלה 9 מדגימה כי רכיב המודול הגלובלי TSMamba עדיף על השיטות המתחרות במנוחי זיכרון אימון וזמן היסק ואף במנוחי זיכרון היסק הוא מציג ביצועים סבירים.

טבלה 9: השוואת בין רכיבי מודול גלובלי במנוחי זיכרון אימון וזמן היסק [11].

Table 5. Ablation study for different global modeling modules. TM denotes training memory, IM denotes inference memory, IT denotes inference time, and OOM represents out of memory.

Methods	Core module	Input resolution	Sequence length	TM (M)	IM (M)	IT (case/s)	Is Global
M5	Large-kernel convolution	128 ³	262144	18852	5776	1.92	✗
M6	SwinTransformer	128 ³	262144	34000	9480	1.68	✗
M7	Self-attention	128 ³	262144	OOM	-	-	✓
Our method	TSMamba	128 ³	262144	17976	6279	1.51	✓

סיכום

המודל SegMamba מציג מספר שכליים למודל המבנה המקורי שמתאים אותו לצורכי עיבוד של צילומי תלת-ממד רפואיים. המודל מציע לתפוסף תכונות תלת-ממדיות באמצעות קונבולוציה מרחבית עם מנגןון שער (GSC) לפניו שמתבצעה השטחה של הקלט לחדר-ממד לשולואה וקטורים שונים לצורך עיבוד על ידי בלוק מבנה תלת-כיווני (MoT). בנוסף, המחבר מציע בלוק חדש (FUE) שתפקידו להציג את התכונות שציהויים ודאייחסית ולעומם את התכונות שציהויים אינם מובהק. תרומה נוספת של המחבר היא ביצירת דאטא סט חדש ואיכותי של 500 צילומים של סרטן המעי הגס עם סימוני סגמנטציה שנוצרו על ידי מומחים. לבסוף, מוצגים במאמר ניסויים השוואתיים רבים שנעשו על מנת להציג את ההישג של המודול החדש לעומת אלטרנטיבות שונות בביצוע סגמנטציה של תמונות תלת-ממד רפואיות מגוון דатаה-סטים ומגוון אבחונים רפואיים.

מודל מomba-טראנספורמר היברידי עבור MambaVision .7

ראייה ממוחשבת

הקדמה

ארQUITקטורת Mamba היא אוטו-רגרסיבית ומבצעת את החישובים צעד אחר צעד. עובדה זו מקשה על עיבוד תמונה, שכן בין חלקי התמונה השונים מתקיים קשרים מקומיים וגלובליים כאחד שאינם נוחים לעיבוד בצורה סדרתית. לפtron בעיה זו הוצעו ארQUITקטורות שונות שאות חלקן הצגנו בפרקם הקודמים [10][11]. אולם, להצעות אלה ישן מספר בעיות. MiVi [10] לדוגמה, מציע בלוק Mamba דו-כיוני לצורך חילוץ הקשר אלובי, אך הבלוק הדו-כיוני מאט באופן משמעותי את החישוב מכורח הצורך לחשב את כל הקלט לפני שמייצרים את התחזית לפולט. בנוסף, המרכיבות של המודל יוצרת אתגר סביב אימונו, מסתכנת-overfitting ולא תמיד מייצרת תשובה מדויקות. משומם כך, מודלים מבוססי טראנספורמר או קונבולוציה עדין מובילים על MiVi במגוון MERCHANTABILITYS ראייה ממוחשבת.

פרק זה נסקרו מאמר משנת 2024 [12] שפורסם על ידי חוקרים מחברת Nvidia, בו מוצגים מספר ניתוחים במודלים המשלבים בלוקים של Mamba עם בלוקים של טראנספורמר (Self-Attention). החוקריםניסו לשלב בלוקים של טראנספורמר לפני המmba, אחרי המmba, באמצעות המmba או כל אחד שכבות. הממצאים שלהם הראו כי שילוב בלוקים של self-Attention בסוף התהילה לאחר בלוקים של Mamba מוגביר את יכולת לתפוס תלויות מרוחקות טווח וקונטקסט גלובלי. בנוסף, שימוש בארכיטקטורה היברידית מגביר את קצב התפוקה של המודול ביחס למודול המmba והטראנספורמר הטהורים. בשלל MERCHANTABILITYS שנבדקו על מספר DATA סטם נמצא כי המודול היברידי MambaVision מציג תוצאות טובות יותר משל המודלים המתחרים.

ארQUITקטורה

איור 18: המראה של ארQUITקטורת המודול היברידי [12] MambaVision.

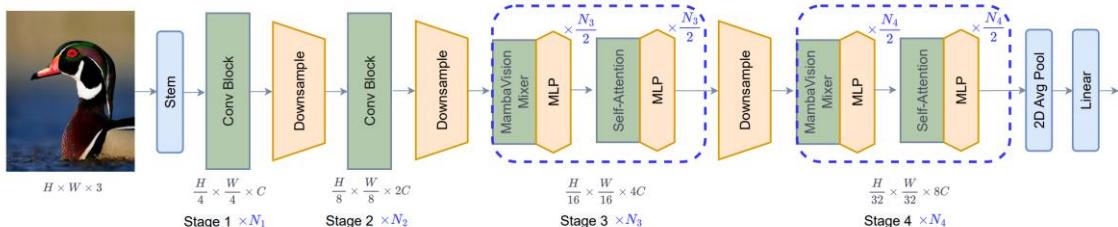


Figure 2 – The architecture of hierarchical MambaVision models. The first two stages use residual convolutional blocks for fast feature extraction. Stages 3 and 4 employ both MambaVision and Transformer blocks. Specifically, given N layers, we use $\frac{N}{2}$ MambaVision and MLP blocks, which are followed by additional $\frac{N}{2}$ Transformer and MLP blocks. The Transformer blocks in the final layers allow for recovering lost global context and capturing long-range spatial dependencies.

הארQUITקטורה שמצועת במאמר מסודרת באופן הבא:

שלבים 1 ו-2

ראשית, שכבת **Stem** מחלצת תכונות ראשוניות מהתמונה כך שקלט בגודל $3 \times H \times W$ הופך לוקטור בגודל $C \times \frac{W}{4} \times \frac{H}{4}$ כאשר C הוא מספר כלשהו של ערוצים. לאחר מכן עובר הקלט דרך בלוק קונבולוציה ובלוק **Downsample** פעמיים לסירוגין (שלבים 1 ו-2), כאשר בלוק הקונבולוציה מוגדר על פי משווה 11 באופן הבא:

משווה 11 : בלוק MambaVision [12]

$$\hat{\mathbf{z}} = \text{GELU}(\text{BN}(\text{Conv}_{3 \times 3}(\mathbf{z}))),$$

$$\mathbf{z} = \text{BN}(\text{Conv}_{3 \times 3}(\hat{\mathbf{z}})) + \mathbf{z},$$

הקלט z עובר פילטר קונволוציה של 3×3 , נורמליזציה, אקטיבציה, וatz שוב נורמליזציה וחיבור שיורי עם הקלט המקורי של שכבה זו z ויוצר את הפלט לשכבה הבאה. באופן זה, מחלצות תכונות מהקלט אך בזכות החיבור השורי לא נאבדות שאר התכונות.

בלוק **Downsample** מבצע שתי קונволוציות רצופות שכל אחת בתורה מקטינה בחצי את הרזולוציה של הקלט. באופן זה שכבות 1 ו-2 המתווארות באירור 18 מוצאות חילוץ מהיר של תכונות מהתמונה המקוריות ומאפשרות לעיבוד לעיבוד של התמונה ברזולוציה נמוכה יותר בשכבות הבאות.

שלבים 3-4

שלבים 3 ו-4 מביצים ($i = 3, 4$) פעולות בהתאם אשר מחציתן הראשונה בלוקים של **MambaVision** וממחציתן השנייה בלוקים של **Self-Attention**. אחרי כל בלוק מעבר הקלט דרך MLP לצורך העשרה ודיק של הפלט. נתאר כעת את בלוק **the-h**-**MambaVision** ונשメיט את הפירוט של בלוק **the-h**-**Self-Attention** אותו תיארנו בפרק 1.

בלוק **the-h**-**MambaVision**

איור 19: בלוק **the-h**-**MambaVision** .[12]

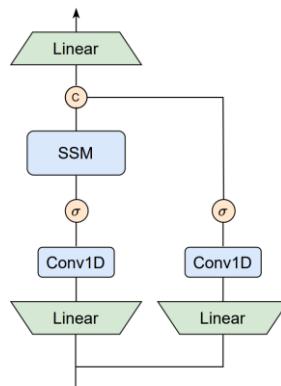


Figure 3 – Architecture of MambaVision block. In addition to replacing causal Conv layer with their regular counterparts, we create a symmetric path without SSM as a token mixer to enhance the modeling of global context.

בלוק הממבה המתואר במאמר זה מבצע מספר התאמות לבלוק הממבה המקורי על מנת להתאים יותר למשימות ראייה ממוחשבת. ראשית, בבלוק הממבה המקורי ישנה קונволוציה סיבתית שמוססתת את ההשפעה של הטוקנים השונים זה על זה כך שטוקן מאוחר לא ישפיע על טוקן קודם. אולם עבר תמנונות, אין אפשרות לאפשר לטוקנים מאוחרים להשפיע על טוקנים מוקדמים שכן לא קיים קשר סדרתי ביניהם. משום כך, הקונволוציה שנוסף לבלוק **the-h**-**MambaVision** תהיה קונבולוציה רגילה. שני המסלולים של הבלוק יהיו זהים מלבד זאת שאחד מהם לא יכול את בלוק **SSM**, זאת כדי לשמור מידע משלבי החישוב הקודמים. הבלוק בנוי לפי הסדר הבא:

1. הקלט עובר הטלה לינארית מממד C לממד $\frac{C}{2}$.
2. הקלט עובר שכבת קונולוציה רגילה.
3. הקלט עובר דרך פונקציית אקטיבציה SiLU.
4. באחד המסלולים יעבור הקלט דרך בלוק **SSM** סלקטיבי [6].
5. שני המסלולים ישורשו ויצרו קלט מממד C .
6. הפלט המשורשר יעבור הטלה לינארית ווועבר כקלט לשכבה הבאה.

המחברים מצינו כי עיצוב הבלוק באופן זה מוביל לייצוג תכונות עשיר יותר, הכללה טוביה יותר ולביצועים משופרים על משימות ראייה ממוחשבת. החוקרים אשרו את האפקטיביות של הבחירה השונות שלהם בעיצוב הבלוק בהשוואה לאפשרויות אחרות על ידי ניסויים נרחבים.

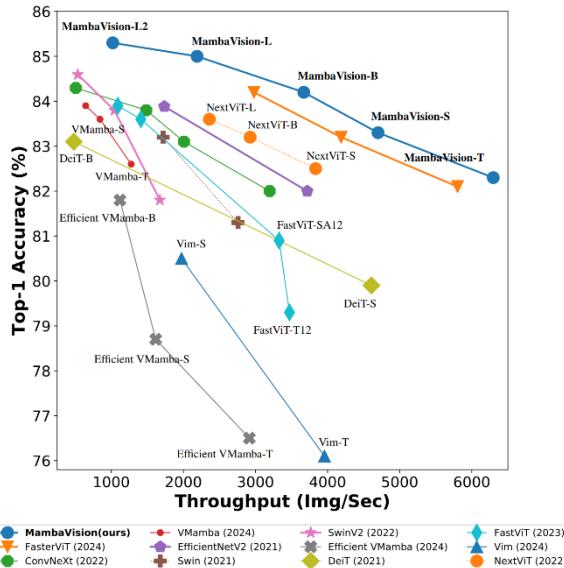
ניסויים והשוואות

לצורך בוחינת המודל נעשו מספר גדול של ניסויים והשוואות, נציג אוטם כאן בקצרה.

Image classification

בטבלה מס' 9 מוצגות התוצאות של הניסוי בקלאסיפיקציה של תמונות. הניסוי השווה בין תוצאות של מודלים מבוססי קונבולוציה, טרנספורמר, מובה והמודל היברידי MambaVision, ומוצא כי MambaVision מציג את התוצאות הטובות יותר במנוחה Top 1 accuracy וגם במנוחה מהירות תפוקה בהשוואה למודלים בעלי גודל דומה. בנוסף, נמצא כי במנוחה עיבוד (FLOPs) המודל היברידי חסכו יותר במשאבים בהשוואה למודלים בגודל דומה. אירור 20 מציג השוואת בין המודלים במנוחה Top 1 accuracy ותפוקה.

אייר 20: השוואת בין מודלי ראייה ממוחשבת במנוחה Top-1 Accuracy ו מהירות תפוקה [12].



טבלה 9: השוואת מבנים קלאסיפיקציה בין מגוון מודלי ראייה ממוחשבת [12].

Table 1 – Comparison of classification benchmarks on ImageNet-1K dataset [4]. Image throughput is measured on A100 GPU with a batch size of 128.

Model	Image Size (Px)	#Params (M)	FLOPs (G)	Throughput (Img/Sec)	Top-1 (%)
Conv-Based					
ConvNeXt-T [23]	224	28.6	4.5	3196	82.0
ConvNeXt-S [23]	224	50.2	8.7	2008	83.1
ConvNeXt-B [23]	224	88.6	15.4	1485	83.8
RegNetY-040 [26]	288	20.6	6.6	3227	83.0
ResNetV2-101 [33]	224	44.5	7.8	4019	82.0
EfficientNetV2-S [27]	384	21.5	8.0	1735	83.9
Transformer-Based					
Swin-T [21]	224	28.3	4.4	2758	81.3
Swin-S [21]	224	49.6	8.5	1720	83.2
Swin-V2-T [22]	256	28.3	4.4	1674	81.8
Swin-V2-S [22]	256	49.7	8.5	1043	83.8
Swin-V2-B [22]	256	87.9	15.1	535	84.6
TNT-S [9]	224	23.8	4.8	1478	81.5
Twins-S [3]	224	24.1	2.8	3596	81.7
Twins-B [3]	224	56.1	8.3	1926	83.1
Twins-L [3]	224	99.3	14.8	1439	83.7
DeiT-B [28]	224	86.6	16.9	2035	82.0
DeiT3-L [29]	224	304.4	59.7	535	84.8
PoolFormer-M58 [38]	224	73.5	11.6	884	82.4
Conv-Transformer					
CoatL-Lite-S [36]	224	19.8	4.1	2269	82.3
CrossViT-S [1]	240	26.9	5.1	2832	81.0
CrossViT-B [1]	240	105.0	20.1	1321	82.2
Visformer-S [2]	224	40.2	4.8	3676	82.1
NextViT-S [17]	224	31.7	5.8	3834	82.5
NextViT-B [17]	224	44.8	8.3	2926	83.2
NextViT-L [17]	224	57.8	10.8	2360	83.6
EfficientFormer-L1 [18]	224	12.3	1.31	6220	79.2
EfficientFormer-L3 [18]	224	31.4	3.9	2845	82.4
EfficientFormer-L7 [18]	224	82.2	10.2	1359	83.4
MaxViT-B [30]	224	120.0	23.4	507	84.9
MaxViT-L [30]	224	212.0	43.9	376	85.1
FasterViT-1 [10]	224	53.4	5.3	4188	83.2
FasterViT-2 [10]	224	75.9	8.7	3161	84.2
FasterViT-3 [10]	224	159.5	18.2	1780	84.9
Mamba-Based					
Vim-T [40]	224	7.0	-	3957	76.1
Vim-S [40]	224	26.0	-	1974	80.5
EfficientVMamba-T [25]	224	6.0	0.8	2904	76.5
EfficientVMamba-S [25]	224	11.0	1.3	1610	78.7
EfficientVMamba-B [25]	224	33.0	4.0	1482	81.8
SiMBA-S [24]	224	15.3	2.4	826	81.7
SiMBA-B [24]	224	22.8	4.2	624	83.5
VMamba-T [20]	224	30.0	4.9	1282	82.6
VMamba-S [20]	224	50.0	8.7	843	83.6
VMamba-B [20]	224	89.0	15.4	645	83.9
MambaVision					
MambaVision-T	224	31.8	4.4	6298	82.3
MambaVision-T2	224	35.1	5.1	5990	82.7
MambaVision-S	224	50.1	7.5	4700	83.3
MambaVision-B	224	97.7	15.0	3670	84.2
MambaVision-L	224	227.9	34.9	2190	85.0
MambaVision-L2	224	241.5	37.5	1021	85.3

טבלה 10: תוצאות סגמנטציה סמנטית ומתחרים [12].

Backbone	Param (M)	FLOPs (G)	mIoU
DeiT-Small/16 [28]	52	1099	44.0
Swin-T [21]	60	945	44.5
ResNet-101 [12]	86	1029	44.9
Focal-T [37]	62	998	45.8
MambaVision-T	55	945	46.0
Swin-S [21]	81	1038	47.6
Twins-SVT-B [3]	89	-	47.7
Focal-S [37]	85	1130	48.0
MambaVision-S	84	1135	48.2
Swin-B [21]	121	1188	48.1
Twins-SVT-L [3]	133	-	48.8
Focal-B [37]	126	1354	49.0
MambaVision-B	126	1342	49.1

Table 3 – Semantic segmentation results with UperNet [34] model using ADE20K dataset. All models are trained using a crop resolution of 512 × 512.

טבלה 11: מבחני זיהוי אובייקטים וסגמנטציה לפי מופע, MambaVision לעומת מתחרים [12].

Backbone	Params (M)	FLOPs (G)	AP ^{box}	AP ^{box} ₅₀	AP ^{box} ₇₅	AP ^{mask}	AP ^{mask} ₅₀	AP ^{mask} ₇₅
DeiT-Small/16 [28]	80	889	48.0	67.2	51.7	41.4	64.2	44.3
ResNet-50 [12]	82	739	46.3	64.3	50.5	40.1	61.7	43.4
Swin-T [21]	86	745	50.4	69.2	54.7	43.7	66.6	47.3
ConvNeXt-T [23]	86	741	50.4	69.1	54.8	43.7	66.5	47.3
MambaVision-T	86	740	51.1	70.0	55.6	44.3	67.3	47.9
X101-32 [35]	101	819	48.1	66.5	52.4	41.6	63.9	45.2
Swin-S [21]	107	838	51.9	70.7	56.3	45.0	68.2	48.8
ConvNeXt-S [23]	108	827	51.9	70.8	56.5	45.0	68.4	49.1
MambaVision-S	108	828	52.3	71.1	56.7	45.2	68.5	48.9
X101-64 [35]	140	972	48.3	66.4	52.3	41.7	64.0	45.1
Swin-B [21]	145	982	51.9	70.5	56.4	45.0	68.1	48.9
ConvNeXt-B [23]	146	964	52.7	71.3	57.2	45.6	68.9	49.5
MambaVision-B	145	964	52.8	71.3	57.2	45.7	68.7	49.4

Table 2 – Object detection and instance segmentation benchmarks using Cascade Mask R-CNN [13] on MS COCO dataset [19]. All models are trained by using a 3× schedule and a crop resolution of 1280 × 800.

זיהוי אובייקטים וסגמנטציה

לצורך השוואה בין המודלים בביצוע NAMES זיהוי אובייקטים, סגמנטציה של מופע וסגמנטציה סמנטיות, וכדי להציג את היעילוות של המודל היברידי ב嚷ון NAMES, אימנו מודלים בגדים דומים, מסוגים שונים ועל פי תנאי אימון שווים. גם בשלוש המשימות האלו המודל היברידי מציג תוצאות עדיפות באופן מובהק, כפי שניתן לראות בטבלה 10 ובטבלה 11.

אימון בקנה מידה גדול על ImageNet-21K

ראשונה עברו מודל מבוסס מבנה כלשהו, בוצעו עבור מאמר זה אימונים של מודלים מבוססי מבנה על DATASET גודל ImageNet-21K, ועם מודלים גדולים יותר באופן משמעותי משנoso בעבר. הניסוי מראה כי המודל הקטן B- MambaVision בעל M739.6M פרטיטרים השתפר ממוначי Top1-accuracy ב-0.7% עבור מודלים שאומנו ברזולוציה של 224. בנוסף נמצאו כי אימון מוקדים – Fine-tuning של המודל הגדול יותר L- Vision- MambaVision משפר את הדיקוק מ-85% ל-86.1% – 86.6% ברזולוציה של 224. ווריאנט גודל יותר של המודל L3- Vision- MambaVision, שגודלו 739.6M פרטיטרים מציג תוצאות טובות אף יותר של 87.3% עבור רזולוציה של 224 – 88.1% עבור רזולוציה של 512. תוצאות אלו מראות את האפשרות לאמן בקנה מידה גדול מודלים מבוססי MambaVision על דאטה סטים שונים וגדולים במיוחד, רזולוציות שונות ועל מודלים בגדים משתנים. הצלחה זו היא קריטית בכך שמודלים מבוססי מבנה יהיו רלוונטיים עבור משימות בעולם האמיתי, הדורשות אימון על DATASETים גדולים במיוחד.

עיצוב בлок ה-MambaVision

כותבי המאמר ביצעו ניתוחים במבנה בлок ה-MambaVision בצד לעצב את המבנה בצורה מיטבית עבור משימות ראייה ממוחשבת. בתחילת, לא הוציאו קובולוציה בענף השינוי של הבלוק (בדומה לבlok המבנה המקורי) וכן השאירו בענף המרכזי את הקובולוציה הסיביתית כמו המקורי. אולם, בצורה זו הבלוק הציג תוצאות פחות טובות מהמתחרים. לאחר מכן הוחלפה הקובולוציה הסיביתית בקובולוציה רגילה והוצאות השתפרו. לאחר מכן הוציאו את הקובולוציה גם לענף השינוי וקיבלו תוצאות טובות יותר. לאחר מכן, החליפו את מגנון השער בין שני הענפים בפעולת שרור. שינוי זה הציג שיפור משמעותי נוסף ביחס לכל המשימות שנבדקו, מה שומוכח את ההיפותזה שרשרו שני הענפים מוגבר את ההבנה של ההקשר הגלובלי. טבלה 12 מצינית את התוצאות של השלבים השונים במבנה.

טבלה 13: השוואת דפוסים היברידיים [12].

Table 5 – Ablation study of on the effectiveness of different hybrid integration patterns. S and M denote self-attention and MambaVision token mixer blocks, respectively.

Model	Pattern	Params (M)	Top-1
Random	-	31.8	81.3
First $N/2$ layers	SSSSMMMM	31.8	81.5
Mixed layers-1	SMSMSMSM	31.8	81.4
Mixed layers-2	MSMSMSMS	31.8	81.6
Last $N/4$ layers	MMMMMMSS	31.8	81.9
Last $N/2$ layers	MMMMSSSS	31.8	82.3

טבלה 12: ניתוח השינויים בבלוק הממבה [12].

Table 4 – Systematic design of MambaVision token mixer. w/o and concat refer to "without" and concatenation. Conv1 and conv2 denote the conv operations in the SSM and additional symmetric branch as shown in Fig. 3. COCO experiments are performed using Mask-RCNN [13] head and $\times 1$ LR schedule.

	ImageNet top-1	COCO		ADE20k mIoU
		AP ^{box}	AP ^{mask}	
causal conv1 - w/o conv2	80.5	44.8	40.4	44.2
conv1 - w/o conv2	80.9	45.0	40.8	44.7
conv1 - conv2 - w/o concat	81.3	45.3	41.0	45.7
conv1 - conv2 - concat	82.3	46.4	41.8	46.0

דפוס היברידי

במחקר זה נעשו ניסויים במספר אפשרויות שלילוב בין הבלוקים של ארכיטקטורת הממבה לבין הטרנספורמר. כל הניסויים בוצעו על אותן ארכיטקטורות ועם אותו אתחול לפרמטרים כך שהשוני הוא רק בשכבות 3 ו- 4 של המודל (ראה איור 18). כאשר נעשה שילוב אקריאו התוצאות היו לא טובות (81.3%) במנוחי-Top-1 Accuracy. לאחר מכן נעשו שילובים שונים כך שהמחצית הראשונה של הבלוקים היא Self-attention והשנייה ממבה, שילוב בלוקים לשירותן פעם כאשר בלוק הטרנספורמר ראשון ופעם כאשר בלוק הממבה ראשון. לאחר מכן נבדקו התוצאות כאשר הבלוקים הראשונים הם בלוקים של ממבה ורק הרביע האחרון של הבלוקים הם בלוקים של טרנספורמר. לבסוף, נמצא כי התוצאות הטובות ביותר היו כאשר הממחצית הראשונה של הבלוקים הייתה בלוקים של ממבה והמחצית השנייה של בלוקים של טרנספורמר. תוצאות הניסוי מפורטות בטבלה 13.

סיכום

במאמר זה נסקרו הביצועים של המודל ההיברידי MambaVision המשלב בלוקים של Mamba המותאמים לראייה ממוחשבת עם בלוקים של Self-Attention. ראשית בוצעו ניסויים שהראו כי ניתן לאמן מודלי ממבה-טרנספורמר היברידיים בגודלים שונים ועל דאטא-סטים גדולים ובמגון רזרולציות ולהניב תוצאות טובות על מגוון משימות. נמצא, כי המודל ההיברידי מציג תוצאות עדיפות על מתחורי בשלל המשימות והפרמטרים שנמדדו. ממצאים אלו מראים כי ביסוס מודלי ראייה ממוחשבת על מבנה היברידי זה הוא בעל פוטנציאל לעקוף את ביצועי המודלים הטהורים מבוססי ממבה, טרנספורמר או קונבולוציה, הן במנוחה דיק ו הן במנוחה מהירות ו שימוש יעיל בכוח חישובי.

סיכום

עבודה זו עוסקת בניתוח עמוק של שתי ארכיטקטורות מרכזיות בימידת מכונה – Mamba ו-Mamba Transformer – ביחס לישומיה המעשיים בתחום הראייה הממוחשבת. החלק הראשון (פרק 1–3) סוקר את ה-Mamba, מהצגתו הראשונית בשנת 2017, דרך התאמתו לעיבוד תמונת מודל (ViT) ו-Vision Transformer, ועד לשימוש קליני ככלים רפואיים המשמשים סגןטייה של צילומים רפואיים. עם כל עוצמתו, מודל זה סובב מגבלות מוגבלות כמו עלות חישובית גבוהה ותלות במנגנון חישוב ריבועי שמצווץ את טווח ההקשר.

החלק השני (פרק 4–6) מוקדש לארכיטקטורת Mamba, המהווה גישה שונה לימיידת רצפים ומיצעה יתרון משמעותי מבחןית יעילות חישובית הודות לסייעות לניאריט. לצד הסקירה התיאורטית, נבחנים גם שימושים קונקרטיים שלה בעולם הראייה הממוחשבת, הן בדמויות Vision Mamba והן בגרסתה מותאמת לسانטייה של הדמויות רפואיות תלת-ממדיות. עם זאת, בשל היotta ארכיטקטורה חדשה יחסית, Mamba עדין חסורה את התשתיות והבשлот של Transformer, ותוכנותיה – כמו עיבוד סדרתי או-סימטריה – עלולות להקשות על אימון ואופטימיזציה במרקם מסוימים.

החלק השלישי (פרק 7) מציג את MambaVision – ארכיטקטורה היברידית המשלבת את היתרון של Mamba Transformer ושל Mamba כדי לעמוד שדרה חזק ויעיל למודלים בתחום הראייה הממוחשבת. במהלך המחבר נבחנו שילובים שונים בין שתי הגישות, והמודל שהתקבל בהנהן מיכולת התמקדות וסינון והן מיכולת לכידת הקשרים רחבים. בנוסף, זהו הניסיון הראשון לאמן מודל מבוסס Mamba בקנה מידה רחב על ImageNet-21K, תוך שימוש במודלים גדולים במיוחד – ציון דרך בדרכה של הארכיטקטורה אל עבר ישומים בעולם האמיתי.

מעבר להשוואה הישירה בין הארכיטקטורות, העבודה ממחישה כיצד מרכיבים חדשניים כמו *Self-Attention* [1], *Fine-Tuning* [5] ו-*Mamba* [6] ו-*FUE* [11] ו-*Downsampling* [12], יכולם לשמש כפתרונות יצירתיים וgamifies לאתגרים מגוונים בתחום הראייה הממוחשבת ולימיידת המכונה בכלל.

לסיכום, העבודה מציעה מבט רחב ועמוק על שתי גישות מובילות בתחום, תוך הדגשת החוזקות והחולשות של כל אחת והפוטנציאל המשולב בינהן. הבנה של הארכיטקטורות, היישומים והטכניקות שנדרנו כאן מהוות תשתיית לחישבה ביקורתית, חדשנית וסילבילית – יסודות חיוניים להמשך ההתפתחות של תחום הבינה המלאכותית.

ביבליוגרפיה

- [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [2] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. In *International Conference on Learning Representations (ICLR)*.
- [3] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *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, Volume 1 (Long and Short Papers)*.
- [4] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment anything*. *arXiv preprint arXiv:2304.02643*.
- [5] Ma, J., He, Y., Li, F., Han, L., You, C., & Wang, B. (2024). *Segment anything in medical images*. *Nature Communications*, 15, 654.
- [6] Gu, A., & Dao, T. (2023). *Mamba: Linear-time sequence modeling with selective state spaces*. *arXiv preprint arXiv:2312.00752*.
- [7] Gu, A., Dao, T., Ermon, S., Rudra, A., & Ré, C. (2020). *HiPPO: Recurrent memory with optimal polynomial projections*. *arXiv preprint arXiv:2008.07669*.
- [8] Gu, A., Goel, K., & Ré, C. (2022). *Efficiently modeling long sequences with structured state spaces*. In *International Conference on Learning Representations (ICLR)*.
- [9] Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H., Schärli, N., & Zhou, D. (2023). *Large language models can be easily distracted by irrelevant context*. In *International Conference on Machine Learning (ICML)*. PMLR, 31210–31227.
- [10] Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). *Vision Mamba: Efficient visual representation learning with bidirectional state space model*. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, Vienna, Austria. PMLR 235.
- [11] Xing, Z., et al. (2024). *Segmamba: Long-range sequential modeling Mamba for 3D medical image segmentation*. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer Nature Switzerland.
- [12] Hatamizadeh, A., & Kautz, J. (2024). *MambaVision: A hybrid Mamba-Transformer vision backbone*. *arXiv preprint arXiv:2407.08083*.
- [13] Verma, P., & Berger, J. (2021). *Audio transformers: Transformer architectures for large-scale audio understanding. Adieu convolutions*. *arXiv preprint arXiv:2105.00335*.
- [14] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). *Masked autoencoders are scalable vision learners*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

- [15] Smith, J. T. H., Warrington, A., & Linderman, S. W. (2023). *Simplified state space layers for sequence modeling*. In *International Conference on Learning Representations (ICLR)*.