【原文】

Continuous Thought Machines

Luke Darlow[1] Ciaran Regan[1,2] Sebastian Risi[1,3] Jeffrey Seely[1] Llion Jones[1]
[1]Sakana AI, Tokyo, Japan
[2]University of Tsukuba, Japan
[3]IT University of Copenhagen, Denmark
{luke, ciaran, sebastianrisi, jeffrey, llion}@sakana.ai

Abstract
Biological brains demonstrate complex neural activity, where neural dynamics are critical to how brains process information. Most artificial neural networks ignore the complexity of individual neurons. We challenge that paradigm by incorporating neuron-level processing and synchronization, as reimposed neural timing as a foundational element. We present the Continuous Thought Machine (CTM), a model designed to leverage neural dynamics as its core representation. The CTM has two innovations: (1) **neuron-level temporal processing**, where each neuron uses unique weight parameters to process incoming histories; and (2) **neural synchronization as a latent representation**. The CTM aims to strike a balance between neural abstraction and biological realism. It operates at a level of abstraction that effectively captures essential **temporal dynamics while remaining computationally tractable**. We demonstrate the CTM's performance and versatility across a range of tasks, including solving 2D mazes, ImageNet-1K classification, parity computation, and more. Beyond displaying rich internal representations and offering a natural avenue for interpretation owing to its internal process, the CTM is able to perform tasks that require complex sequential reasoning. The CTM can also leverage adaptive compute, where it can stop earlier for simpler tasks, or keep computing when faced with more challenging instances. The goal of this work is to share the CTM and its associated innovations, rather than pushing for new state-of-the-art results. To that end, we believe the CTM represents a significant step toward developing more biologically plausible and powerful artificial intelligence systems. We provide an accompanying interactive online demonstration and an extended technical report.

Figure 1: Solving 100 steps from 39 × 39 mazes: (a, b) Observing using attention (no positional encoding (weights shared), imagining route (arrows) from red to green pixels (b)) attending beyond 100 steps; and (c) generalizing to 99 × 99 via sequential re-applications of the same model.
39th Conference on Neural Information Processing Systems (NeurIPS 2025).

(a) Each random-colored subplot is (b) The CTM looks around to build up its prediction, effectively tracing an

a single neuron's activity. intuitive path by synchronizing its neurons to attend dynamically.

【解读】
同学们，大家好！今天我们要一起解读一篇非常有意思的学术论文，它的题目叫《连续思维机器》（Continuous Thought Machines，简称 CTM）。这篇论文来自日本的 Sakana AI 实验室，他们是一群非常有创意的研究者，致力于让 AI 变得更像真正的生物大脑，而不是冷冰冰的计算器。

1. 为什么要发明“连续思维机器”？
大家知道，现在的 AI（比如 ChatGPT）虽然很厉害，但它们的工作方式和我们的大脑其实差别很大。目前的 AI 神经网络通常是“静态”的：你给它一个输入（比如一张图），它经过一层层的计算，直接给你一个输出。这就好比你在看一张静态的照片，像素是固定的。
但是，**生物大脑（比如你的人脑）是动态的**。你的思维像是一条连续流动的河，神经元不仅仅是在传递信号，它们在“时间”上是有节奏、有同步的。这篇论文的核心观点就是：如果我们想让 AI 更聪明，就不能忽略神经元在**时间维度**上的复杂性。

2. CTM 的两大“黑科技”
为了模仿大脑，作者提出了两个核心创新点，我们来通俗地理解一下：

- **神经元级的时间处理（Neuron-level temporal processing）：** 以前的 AI 神经元像是一个只做加法的计算器。而在 CTM 中，每个神经元都有了“记忆”，它们能处理“过去一小段时间”的历史信息。想象一下，原本的 AI 只是盯着眼前的一帧画面，而 CTM 的每个神经元都在看一段小视频，这样它就能理解事情的来龙去脉。
- **神经同步（Neural synchronization）：** 在大脑中，当你想通一个问题时，很多神经元会同时“放电”，这种同步非常关键。CTM 把这种“同步”作为了一种核心的表达方式。这就像一个交响乐团，如果每个人各拉各的，那是噪音；只有大家按同一个节奏同步演奏，才能产生美妙的乐章（也就是智慧）。

3. 它能做什么？（看图说话）
论文中展示了 CTM 如何解决**2D 迷宫**（Figure 1）。这是一个非常经典的测试，因为它需要“推理”。

- 你看图 1 里的那些箭头，那是 AI 在“想象”从起点（红色）到终点（绿色）的路线。它不是一下子猜出来的，而是像人一样，眼睛盯着地图，一步步在脑海里规划路径。
- 更有趣的是“**自适应计算**”（Adaptive Compute）。大家考试时都有这种体会：遇到“1+1=？”这种题，你一秒钟就写出答案；但遇到压轴大题，你得思考很久。目前的 AI 大多很死板，无论题难易，消耗的计算量都一样。但 CTM 可以做到：**简单任务早点停，困难任务多想一会儿**。这就是更像生物的地方！

4. 总结
作者非常谦虚，他们说这篇论文的目的不是为了在跑分榜上拿第一（State-of-the-art），而是为了探索一条新路：**让 AI 的思考过程在时间上连续起来，并且更符合生物学原理**。图 2 展示的

那些彩色斑点，就是 CTM 内部神经元活动的“脑部扫描图”，它们在动态地调整和同步，就像一个正在思考的大脑。

这告诉我们，未来的 AI 不仅仅是堆算力，更是对人类自身智慧原理的深刻模仿。希望这个解读能帮大家理解这篇硬核论文的精髓！你好！我是你的学术导师。今天我们要一起探讨一篇非常前沿的计算机科学论文，主题是关于一种新型的人工智能模型——“连续思维机器”（Continuous Thought Machine，简称CTM）。

现在的AI（比如ChatGPT或者图像识别软件）虽然很强大，但它们的工作方式其实和我们人类的大脑有很大不同。这篇论文提出了一种新的思路，试图让AI更像生物大脑那样运作，特别是引入了“时间”这个概念。

为了让你更好地理解，我将把文档分成两个部分来详细解读。让我们开始第一部分。

【原文】
Figure 2: ImageNet-1K demonstration. (a) Complex neural dynamics whose synchronization are the representation with which the CTM observes and predict. (b) CTM's **attending process**, showing all 1k attention heads (left) and average thereof (middle). Arrows trace the average; highlighting over internal ticks, exemplifying a complex path that emerges without any training signal. We discuss more interesting emergent properties of the CTM in Appendix I. Video demonstrations are here.

1 Introduction

Biological brains exhibit complex time-dependent neural dynamics, but artificial neural networks (NNs) intentionally abstract away the precise timing and interplay of neuron interactions to facilitate large-scale deep learning [11, 2, 3]. While enabling significant advancements over the years, these simplifications deviate from fundamental biological neural computation principles. Emulating the temporal aspects of neural dynamics present in brains remains challenging. Consequently, modern NNs prioritize simplicity and computational efficiency over strict emulation. This abstraction, though task performant, contributes to a gap in our flexible human cognition and current AI capabilities, suggesting missing fundamental components, potentially related to temporal processing [4, 5, 6].

【解读】
同学们，我们先来看这段内容的核心思想：**为什么现在的AI虽然厉害，但在某些方面还是不如人脑灵活？答案可能藏在“时间”里。**

首先，让我们看看Figure 2（图2）的描述，虽然我们看不到图，但文字告诉了我们这台机器（CTM）是如何“看”世界的。

1. **动态与同步**：作者提到了“神经动力学（neural dynamics）”和“同步（synchronization）”。想象一下，你的大脑里有无数个神经元，它们不是像电灯开关那样简单地开和关，而是像一支交响乐团，通过特定的节奏和旋律（同步）来传递信息。CTM正是试图模拟这种复杂的波动，而不是仅仅处理静态的数据。
2. **注意力的轨迹**：描述中提到了“attending process”（关注过程）。当你走进一个房间，你的眼神会扫视，先看门，再看窗，最后看人。CTM也是这样，它在处理图像时，会产生一条复杂的“路径”，就像人的目光移动一样。最神奇的是，这种行为不需要特意去教（without any training signal），它是自然涌现出来的。

接下来是引言（Introduction）的第一段，这里作者进行了一场**生物大脑 vs. 人工神经网络（NNs）**的辩论。

- **生物大脑的魔法**：我们的大脑是高度依赖“时间”的。神经元什么时候放电、频率是多少，这些微小的时间差包含了巨大的信息量。
- **传统AI的取舍**：现在的深度学习模型（也就是大家常听说的神经网络），为了能在计算机上跑得快、算得快，故意**忽略**了这些复杂的生物细节。作者用了一个词叫“Abstract away”（抽象化/剥离）。这就像是为了画一张简单的地图，我们把地形的高低起伏都抹平了，只保留了路线。这样做的好处是计算效率极高，造就了今天AI的繁荣。
- **代价是什么？**：虽然这种简化让AI在下围棋、画画方面表现出色（task performant），但也导致了一个缺憾：AI缺乏人类那种灵活的认知能力。作者认为，现在的AI之所以看起来有点“死板”，或者缺乏某些根本性的智能，很可能就是因为我们把“时间处理（temporal processing）”这个关键组件给丢掉了。

简单来说，这段话告诉我们：现在的AI是“静态快照”式的高手，而生物大脑是“动态视频”式的大师。为了让AI更进一步，我们需要把“时间”这个维度找回来。

【原文】
Despite its outstanding performance, modern AI lacks the flexibility, efficiency, fluidity, generalization capabilities, and common sense of human intelligence, which operates in an world where learning and adaptation are tied to the arrow of time [5, 7, 8, 6]. We argue that incorporating time as part of neural computation is crucial for advancing AI [9, 10]. We introduce the Continuous Thought Machine (CTM), a model explicitly incorporating neural dynamics over time. Our contributions are:

1. The CTM architecture using an **internal dimension** for modeling the temporal evolution of neural activity, **neuron-level models** (NLMs) as a more biologically plausible micro-level abstraction of neurons that unfold neural dynamics, and the use of **neural synchronization** directly as the representation that is implemented via temporal correlations between neuron-level activity (Section 3.4) and observation and prediction, making neural dynamics the core operating principle.
2. An exposition of the capabilities unlocked by the CTM, including strong performance on sequential reasoning tasks (Figure 1), native adaptive compute, time, natural and

interpretable behavior such as ‘looking around’ images before predicting (Figure 2), and learning algorithmic solutions, opening up opportunities to the AI community for new research.

【解读】
这一段接着阐述了为什么要创造CTM，以及它到底厉害在哪里。

首先，作者毫不客气地指出了现代AI的短板。尽管AI现在的跑分很高，但在**灵活性（flexibility）、流动性（fluidity）和常识（common sense）**上，依然被人类吊打。为什么？因为人类生活在一个受“时间之箭（arrow of time）”支配的世界里。我们的学习和适应是随着时间流逝连续发生的，而不是像AI训练那样，一次性把几亿张图塞进去算完拉倒。作者坚信：**要想AI进化，必须把“时间”作为计算的一部分。**

于是，主角登场了：**连续思维机器（CTM）**。

作者列出了这篇论文的三大核心贡献（虽然这里只列了两点，但包含了很多技术细节，我来给你们拆解一下）：

第一大贡献：架构创新（怎么造出来的？）

1. **内部维度（Internal dimension）**：现在的AI处理信息通常是一层一层传下去的（空间上的深度）。而CTM在内部引入了“时间轴”，让神经活动可以在时间上演变。就像从看连环画变成了看连续剧。
2. **神经元级模型（NLMs）**：作者没有使用那种极度简化的数学神经元，而是设计了一种更像生物神经元的新模型。这种微观层面的模拟，让由于时间带来的动态变化得以展现。
3. **神经同步（Neural synchronization）**：这是最酷的一点。CTM不是通过单一的数值来表示“这是一只猫”，而是通过神经元活动的**时间相关性（temporal correlations）**。就像摩尔斯电码，信息的意义藏在节奏和同步里，而不仅仅是信号的强弱。这使得“神经动力学”成为了整个系统的核心驱动力。

第二大贡献：能力解锁（能干什么？）

1. **序列推理（Sequential reasoning）**：因为有了时间概念，CTM非常擅长处理需要一步步推导的任务（比如复杂的逻辑题）。
2. **原生自适应计算（Native adaptive compute）**：这是高三物理或数学中常提到的“效率”问题。现在的AI不管问题多简单或多难，消耗的计算量通常是一样的。但CTM像人一样，简单问题想得快（计算少），难问题想得久（计算多）。这种“想多久由题目难度决定”的能力，叫做自适应计算。
3. **可解释的行为（Interpretable behavior）**：正如我们在第一段解读中提到的，CTM在做决定前会“四处看看”（looking around）。这对人类来说非常友好，因为我们可以直观地看到AI关注了图片的哪些部分，而不像传统AI那样是个完全的黑盒子。

总结一下，这段话告诉我们：CTM不仅仅是一个新的算法，它是一次向生物大脑致敬的尝试。它通过模拟神经元随时间的动态变化，试图赋予AI人类般的灵活性和推理过程，让AI不仅能给

出答案，还能像人一样“思考”。你好！我是你的学术导师。很高兴能为你解读这篇关于**CTM（Computational Temporal Model，计算时序模型）**的前沿学术论文片段。

这篇论文探讨的是人工智能如何更像人脑一样，通过“时间”和“同步”来思考，而不仅仅是依靠静态的数据处理。对于高三学生来说，你可以把这理解为从“看照片”进化到了“看电影”，不仅关注画面，更关注画面随时间的变化和节奏。

我们将这段文本分为两个主要部分进行深度剖析。

【原文】
The CTM learns to use neural synchronization as its latent representation, distinguishing it from existing work that explores synchrony as emergent property for post-hoc use [11, 12]. This representation is distinct from the common static, Snapshot, representations used in most modern NNs as it directly encodes the temporal interplay of neural dynamics.

Recurrence & Reasoning. Recurrence is a strong contender for extending model complexity beyond current scaling limitations [13, 14, 15]. We posit that recurrence, while essential, is merely one piece of the puzzle. The temporal dynamics unlocked by recurrence are equally crucial. We demonstrate in this paper that neural dynamics can be leveraged to build a new kind of neural network with surprising capabilities. We show how the CTM navigates complex 2D mazes by forming internal maps (which positional encodings ignore), it can ‘look around’ (without any signal to do so) when classifying images and exhibits native adaptive compute time as a side-effect.

(Section 5), and utilizes its dynamic representations for tasks requiring memory and sequential reasoning (Section 6). These capabilities emerge from the same core architecture applied to different tasks, showcasing its versatility and adaptability. We believe the CTM represents a step towards bridging the gap between powerful modern AI and biological plausibility.

【解读】
这段话极其精彩，它点出了这篇论文的核心创新点：**让AI拥有“时间感”和“节奏感”**。

首先，作者提到了**神经同步（Neural Synchronization）**。想象一下，你们班级合唱比赛，大家并不是每个人只要唱对自己的音就行了，最关键的是要“同步”，大家的节奏要合在一起，这种“合拍”本身就蕴含了巨大的信息。
传统的神经网络（NNs），也就是作者说的“Modern NNs”，大多使用的是**静态快照（Static, Snapshot representations）**。这就像是你复习时只看了一张知识点的截图，虽然能看到内容，但你看不到这些知识点是如何推导出来的过程。而CTM模型不同，它不仅仅学习“是什么”，还学习神经元之间“何时”一起激发，这种**时间上的相互作用（Temporal interplay）**被直接用作了模型的思考方式，而不是像以前的研究那样，只是事后才发现的某种副产品。

接下来，作者提到了**循环与推理（Recurrence & Reasoning）**。
你们在高一信息技术课学过编程中的“循环（Loop）”或“递归”。在AI领域，**循环神经网络（RNN）**曾是王者，因为它能处理序列信息（比如读一段话）。但现在的AI霸主（比如ChatGPT背后的Transformer）主要是靠堆算力、堆参数（Scaling）变强。作者认为，光靠堆参数会遇到瓶颈，要把模型变得更复杂、更聪明，还是得把“循环”请回来。
但是！作者强调，光有循环还不够，**时间动力学（Temporal Dynamics）**才是拼图的关键一块。这就像你解一道很难的数学压轴题，不仅仅是你脑子里在“循环”思考，更重要的是你的思考过程是随着时间流动的，你的思路是动态变化的。

CTM展示了惊人的能力：

1. **走迷宫（Navigates 2D mazes）：** 普通AI走迷宫可能只是记住坐标（就像背地图），但CTM能形成**内部地图（Internal maps）**，这意味着它真的“理解”了空间结构，就像你到了一个陌生城市，脑海里建立起了方位感。
2. **“四处张望”（Look around）：** 在识别图像时，它会像人眼一样主动去扫描图片的重点区域，而且这是它自发的行为，不需要人专门写代码教它。
3. **生物合理性（Biological Plausibility）：** 这是整个AI界的一个终极梦想。现在的AI虽然强，但工作原理和人脑差别很大。CTM试图通过模拟神经元的同步和动态变化，来缩小人工大脑和生物大脑之间的差距。

总的来说，这段话告诉我们：CTM不仅仅是一个静态的计算器，它更像一个有生命、有节奏、能动态思考的生物大脑。

【原文】
The remainder of this paper details work (Section 2), describes the CTM (Section 3), evaluates core capabilities on 2D mazes, ImageNet-1K classification, and parity computation (Sections 4 to 6), summarizes further experiments and applications (Section 7), and discusses findings (Section 8).

2 Related Work

The CTM uses neural timing and synchronization as core computational principles. This positions it relative to, yet distinct from, several lines of research.

Adaptive Computation. Many approaches achieve adaptive computation via implicit mechanisms. Early work on networks [16] use **intermediate classifiers** for early termination. PointerNet [17] and Adaptive Computation Time (ACT) [18] implement learning halting modules governing recurrent steps. More recent methods like AdaTape [19] dynamically extend input sequences, while Sparse Universal Transformers (SUTs) [20] combine recurrent weight sharing with dynamic halting and Mixture-of-Experts. In contrast, the CTM’s adaptive processing emerges from per-tick per-input based on

certainty and loss dynamics; Section 3.5) emerges naturally from its core architecture, driven by the unfolding of its internal neural dynamics without dedicated halting components.

【解读】
这一段主要是论文的“路标”（论文结构）以及**相关工作（Related Work）**的综述。在学术写作中，这一部分至关重要，因为作者必须说明：“别人的研究走到哪一步了？我的研究和他们有什么不同？我厉害在哪里？”

首先，作者快速列出了论文的结构（从第2节到第8节），包括迷宫测试、图像分类（ImageNet-1K是AI界的经典“高考题”）、奇偶校验计算等，这些都是用来验证CTM实力的实验。

然后，重点来了：**自适应计算（Adaptive Computation）**。
这是什么意思呢？试想一下，你们考试做题。

- 遇到一道简单的“1+1=？”，你只需花0.5秒。
- 遇到一道复杂的解析几何大题，你可能要花15分钟。
  这就是“自适应计算”——根据题目的难易程度，动态调整思考的时间和算力。

作者回顾了以前的AI是怎么做这件事的：

1. **早期方法（中间分类器）：** 就像做卷子时，每做几步就有一个老师问你“做完没？确信吗？”，如果你确信了，就让你提前交卷（Early termination）。
2. **ACT (Adaptive Computation Time) 和 PointerNet：** 这些方法专门设计了一个“暂停模块”（Halting module），就像在你脑子里装了一个闹钟，专门负责决定“这道题思考5秒还是10秒”。
3. **AdaTape 和 SUTs：** 这些是更现代的方法，试图通过延长输入序列或者混合专家模型（Mixture-of-Experts）来实现灵活计算。

CTM的降维打击在哪里？
作者用了**“In contrast”（相比之下）**这个词来强调区别。**
以前的方法，无论是装“闹钟”还是设“检查站”，都是人为设计（Engineered）**的额外组件。**
而CTM的自适应能力是涌现（Emerges）**出来的。这词很高级，意思就是“自然而然发生的”。
CTM不需要一个专门的模块来喊“停”。它在思考（神经动力学展开）的过程中，根据自己对答案的确定程度（Certainty）和内部状态，自然地决定何时停止思考。就像你做一道题，当你算出一个靠谱的答案时，你会下意识地停笔，而不需要旁边有人专门按一下秒表告诉你“时间到”。

作者在这里强调：CTM的这种能力是**内生的（Native）**，源于其核心架构，不需要外挂任何“停止组件”。这再次呼应了上一段提到的“生物合理性”——因为我们人类思考时，也是这样自然而然地分配注意力的。你好！我是你的学术导师。很高兴能为你解读这份关于前沿人工智能模型的学术文档。

这段文本探讨的是一个名为 **CTM（Computational Temporal Model，计算时序模型）** 的系统。它涉及两个核心概念：一是如何通过“迭代”和“递归”来进行推理，二是如何从生物大脑的神经动力学中汲取灵感。

为了让你更好地理解，我将这份文档分为两个主要部分进行解读。我们会先看第一部分，关于模型是如何像人类一样“思考”和处理时间的。

【原文】
Iterate and Recurrence Reasoning. The CTM’s internal ticks facilitate iterative refinement, akin to models generating internal computational steps. For instance, Quick-STAF [21] uses hidden rationale generation in language models, and Recurrent Independent Mechanisms (RIMs) [22] employ modular, asynchronous sub-networks for multi-step reasoning. While Recurrent Models of Visual Attention (RAM) [23] leveraged recurrence for sequential processing of visual glimpses, the CTM’s novelty lies in generating internal neural dynamics from neuron-level histories across a decoupled time dimension and then utilizing the emergent **temporal patterns of neural synchronization** as its primary representation. This contrasts with RAM’s focus on perceptual decision-making from external glimpses or models relying solely on a final recurrent state.

【解读】
这一段的主题是**“迭代与递归推理”**。我们先来拆解一下这里的核心逻辑。

首先，想象一下你在做一道很难的高考数学导数题。你读完题目后，不会马上写出答案，而是在脑子里经历一系列的思考步骤：先求导，再找极值点，分类讨论……每一步都在完善你的思路。这段文字开头提到的“CTM’s internal ticks”（CTM的内部时钟/滴答声）指的就是这个过程。**CTM模型拥有“内部计算步骤”**，它不像传统模型那样“输入即输出”，而是允许在内部进行多次“迭代优化”（Iterative Refinement），就像在脑子里多转几个弯，把问题想得更透彻。

文中举了两个例子来类比：

1. **Quick-STAF**：这是一种语言模型，它在给出答案前会生成“隐藏的理由”，类似于打草稿但不给你看。
2. **RIMs（循环独立机制）**：它使用多个独立的模块（像大脑的不同区域）异步地工作，进行多步推理。

接下来，作者引入了一个对比对象——**RAM（视觉注意力循环模型）**。RAM的工作方式像是一个拿着手电筒在黑屋子里找东西的人，它通过一次次地“瞥见”（glimpses）外部图像的不同部分来逐步理解整体。

那么，CTM的创新点（Novelty）在哪里呢？ 这是一个重点考试概念，请注意：
CTM不仅仅是像RAM那样按顺序处理外部信息，它的核心在于**“生成内部神经动力学”**

（Generating internal neural dynamics）。
这里有一个非常酷的概念叫“解耦的时间维度”**（Decoupled Time Dimension）。你可以把它理解为“思考时间”与“物理时间”的分离。比如，现实世界只过了一秒钟（物理时间），但在你的脑海里，神经元可能已经经历了一场复杂的风暴（思考时间）。CTM记录了每个神经元在这个思考过程中的历史变化。

最关键的是，CTM如何表示这些信息？它使用的是**“神经同步的时间模式”**（Temporal patterns of neural synchronization）。**
在生物课上你学过，大脑神经元不仅通过发射信号的强弱来传递信息，还通过“谁和谁同时发射信号”（即同步性）来编码信息。CTM正是模拟了这一点：它不再只盯着最后输出的那个状态（那是传统RNN的做法），也不是只盯着外部的图像碎片（那是RAM的做法），而是去“聆听”内部神经元群体在时间轴上是如何协同共振**的。这种“节奏”和“同步”本身，就是它对知识的表达。

总结来说，这一段在告诉我们：CTM不仅仅是在计算，它在模拟一种具有时间深度的、内在的“思考流”，并把神经元之间的“合唱节奏”作为理解世界的关键。

【原文】
Biologically Inspired Neural Dynamics. There is growing interest in more biologically plausible neural computation [24]. Examples include Liquid Time-Constant Networks (LTC-NNs) [25] with neurons governed by time-varying differential equations and various Spiking Neural Networks (SNN) paradigms that inherently use discrete, timed events, with prior work also exploring synchronization mechanisms [26, 27]. Our model draws inspiration from temporal coding and neural synchrony, but uses: (1) neuron-level models (NLMs) to process histories of continuously-valued pre-activations to produce complex dynamics, and (2) *neural synchronization* as the primary latent representation for observation and output. While inspired by principles like spike-timing and synchrony, CTM abstracts discrete spiking on local temporal integration and population-level synchronization into a tractable, differentiable framework suitable for gradient-based deep learning, rather than replicating detailed Biophysics. This situates the CTM alongside, yet distinct from, extensive work on models such as Liquid State Machines [28], and diverse SNNs that exploit precise spike timing for computation or employ specialized learning rules [29, 30, 31, 32, 33]. These latter models often emphasize event-driven dynamics, explore non-differentiable computation, or focus on online learning. The CTM offers a complementary direction, retaining inspiration from biological timing while ensuring compatibility with established deep learning paradigms.

【解读】
这一段探讨了CTM的另一个支柱：**“受生物学启发的神经动力学”**。也就是说，这个模型试图模仿真正的人脑运作方式，但又做了一些聪明的改良。

1. 背景：让AI更像人脑
现在的学术界很流行让计算机模仿生物大脑（Biologically plausible neural computation）。作者提到了两个“仿生”前辈：

- **LTC-NNs（液体时间常数网络）**：你可以把神经元想象成一个个漏水的水桶，水位变化由微分方程控制，非常动态。
- **SNNs（脉冲神经网络）**：这是最像人脑的模型。人脑神经元不是一直通电的，而是“biu-biu-biu”地发射脉冲（Spikes）。这是一个离散的、基于事件的过程。

2. CTM的独特配方
CTM虽然借鉴了上述思想（如时间编码和神经同步），但它做了一个至关重要的**取舍**。
真正的生物大脑非常复杂，且脉冲信号是断断续续的（离散的）。在数学上，处理“断断续续”的函数非常麻烦，因为它们**不可导**。而在高三数学导数章节，你知道，如果一个函数不可导，我们就很难求极值。在AI训练中，如果不可导，我们就无法使用最主流的“梯度下降法”来训练模型。

因此，CTM做了一个聪明的抽象：

- 它保留了**神经元级别的历史处理能力**（NLMs），让每个神经元都有“记忆”。
- 它保留了**神经同步**（Neural Synchronization）作为核心表达方式。
- 但是！它抛弃了那种难以计算的“离散脉冲”，将其转化为了一个**“可处理的、可微的框架”**（Tractable, differentiable framework）。

3. 为什么这很重要?
这就像是设计飞机。我们从鸟类那里学到了空气动力学（仿生），但我们不会真的给飞机粘上羽毛去模仿鸟的每一个细胞（那是过度复制生物物理学）。

- **LSMs（液体状态机）和传统SNNs**：它们就像是试图完全复制鸟的羽毛和肌肉，往往强调“事件驱动”或“非可微计算”，这导致它们很难用现代强大的深度学习工具（如PyTorch, TensorFlow）进行训练，或者需要专门的、冷门的学习规则。
- **CTM**：它走了一条“折中互补”的路线。它保留了生物学的灵魂（时间性、同步性），但披上了现代数学的外衣（可微性、连续值），从而保证了它能与现有的深度学习范式（Compatibility with established deep learning paradigms）无缝兼容。

总结：
CTM试图在“生物真实性”和“工程实用性”之间找到最佳平衡点。它告诉我们：不需要完全照搬大脑的每一个生物细节，只要抓住“时间同步”这个核心逻辑，并用微积分能处理的数学语言把它描述出来，就能创造出既聪明又好训练的新一代AI模型。你好！我是你的学术导师。这一部分的内容非常核心，它不仅综述了前人的研究，还正式推出了这种名为“连续思维机器”（CTM）的新架构。

由于提供的Markdown文本总长度适中，且逻辑紧密相连（第一段为第二段的方法提供了理论背景），我将把这部分内容作为一个完整的整体来进行深度解读。这能帮助你更好地理解“同步”这个概念是如何从理论演变为CTM的核心组件的。

【原文】

Synchronization. Reichert & Serre [11] proposed a model where synchronization emerges from in-teractions among complex-valued neurons, serving as a gating mechanism that modulates information flow and enables post-hoc grouping of neurons for tasks like object segmentation. Unlike CTM, however, their model does not use synchrony as a learned latent representation during computation. Other approaches in complex-valued neural networks [12] employ synchronization from a control-theoretic perspective, aiming to stabilize or coordinate networks via externally enforced synchrony. In contrast, CTM integrates synchronization intrinsically, optimizing neural phase relationships during training to encode task-relevant representations. This integration in CTM is a computationally grounded model of synchrony, fundamentally distinct from prior works that treat synchrony as a control objective.

3 Method

Figure 3: CTM architecture overview. Key components include: ① Synapse model generating pre-activations from prior post-activations $z^{'}$ and attention output $o^{'}$. ② History of pre-activations $A^{'}$. ③ Neuron-level models (NLMs) processing $A^{'}$ to modulate ④ post-activations $z^{''}$. ⑤ History of post-activations $Z^{''}$. ⑥ Neural synchronization matrix $S^{'}$ computed from $Z^{''}$. ⑦ Selection pairs from $S^{'}$ form ④ latent representations used for ④ outputs $y^{'}$ and attention queries $q^{'}$. ④ Attention output $o^{'}$ is concatenated with $Z^{''}$ for the next internal tick. Owing to the inherent difficulty in visualizing its dynamics, time based architecture, we include the supplementary video 'arch_mp4' (hosted here) that visualizes functional data flow.

The Continuous Thought Machine (CTM) is a neural network architecture that explicitly incorporates neural dynamics as a core component. Figure 3 (① → ④) and pseudocode in Listing 1 illustrate the CTM's flow. The CTM differs from other recurrent architectures [34, 35, 36, 18, 37] in two ways: (1) it applies neuron-level models (NLMs), each with private weights, to histories of pre-activations to produce complex neuron-level activity (Section 3) and (2) it uses neural synchronization directly as the latent representation for modulating data and producing outputs (Section 3.4).

3.1 Continuous Thought: The Internal Sequence Dimension

The CTM uses an internal dimension $\ell \in \{1, \cdots, \tau'\}$, decoupled from data dimensions. This timeline of internal ticks [34, 35, 36, 37] enables iterative refinement of representations, even for static data. Unlike conventional recurrent models that process data in fixed sequences, the CTM adopts a self-generated timeline of 'thought steps' that unfolds neural dynamics for downstream use.

【解读】

同学们，这一大段内容是我们理解“连续思维机器”（CTM）这篇论文的“发动机”。它首先回顾了科学界对“神经同步”的看法，然后揭示了CTM是如何颠覆性地利用这一点的。让我们像拆解一台精密仪器一样来分析它。

1. 什么是“同步”（Synchronization）？它为什么重要?
想象一下交响乐团。如果每个乐手各吹各的，那是噪音；如果大家按照同一个节奏演奏，那就成了音乐。在神经科学中，“同步”就是指不同的神经元在同一时间“放电”或活跃。

- **前人的做法（Reichert & Serre等）：** 以前的研究认为，同步就像是一个**“开关”或“门控机制”**。比如，当我们看一张图片时，负责识别“狗”的那些神经元会一起同步放电，告诉大脑“这是一组”。但这只是一种事后发生的现象，或者是一种维持秩序的手段（控制理论视角），就像指挥家强行要求大家节奏一致以免乱套。
- **CTM的突破：** CTM说：“不，同步不仅仅是开关，同步本身就是**语言**。”在CTM中，神经网络在训练过程中会主动去学习和优化神经元之间的相位关系（Phase Relationships）。这种同步关系直接编码了任务所需的信息。这不仅是为了“稳”，而是为了“懂”。这被称为“内在集成”（Integrates synchronization intrinsically），是CTM与前人工作的根本区别。

2. 走进CTM的“核心引擎”（Method & Figure 3）
接下来作者展示了CTM的架构图（Figure 3）。别被那些希腊字母和箭头吓倒，我们来看它的逻辑流：

- **复杂的神经元（NLMs）：** CTM不像传统的神经网络那样使用简单的加减乘除神经元。它引入了**“神经元级模型”（NLMs）**。每个神经元都有自己的“私人权重”和“历史记忆”（History of pre-activations A'）。你可以把它想象成：以前的神经元是只会传话的士兵，而CTM的神经元是每一个都带着笔记本、能独立思考的战术专家。
- **同步即表达（Latent Representation）：** 图中的第⑥步计算了“神经同步矩阵 S'”。这正是上面理论的应用——机器通过观察神经元之间是否步调一致，来形成对当前事物的理解（潜在表示）。

3. 最酷的概念：内部思维维度（The Internal Sequence Dimension）
这是高三物理或数学中很少见到的概念，但在人工智能哲学中非常迷人。

- **传统的RNN（循环神经网络）：** 处理数据的节奏是被动的。读入第一帧画面，处理一次；读入第二帧，处理一次。就像你在考场上，老师念一题你做一题，没时间多想。
- **CTM的“思维时间”：** CTM引入了一个**与数据无关的内部时间轴**（Internal ticks）。哪怕你给它一张静止的图片（Static data），它也可以在内部“思考”好几步（Self-generated timeline）。
  - **类比：** 就像你在美术馆看一幅画。画是静止的（数据不变），但你的思维在动：“这是什么？哦，是星空。为什么这里颜色这么深？那个漩涡代表什么？”
  - 这个过程就是“Iterative refinement”（迭代优化）。CTM赋予了机器“驻足思考”的能力，让它在输出结果之前，先在内部把神经动力学“展开”来推演一番。

总结一下： 这段话的核心在于确立CTM的江湖地位——它不是为了同步而同步，而是把“神经元之间的共鸣”当作思考的语言，并且它拥有独立的“思考时间”，不再单纯被外部数据的输入速度推着走。这让它更像一个真实的大脑。你好！很高兴能为你解读这份关于前沿神经网络架构的学术文档。这段内容描述了一种模仿生物大脑运作机制的独特神经网络设计，涉及到记忆、神经元之间的独立计算以及如同脑电波般的“同步”机制。

我们将把文档分为两个主要部分来详细剖析。第一部分关注这个“大脑”内部是如何思考和处理信息的（突触与神经元模型）；第二部分关注它是如何通过“同步”来产生输出并感知外部世界的。

让我们开始第一部分的解读。

【原文】

3.2 Recurrent Weights: Synapses

A ① synapse model, f_{syn}, interconnects neurons in a shared D-dimensional latent space, $z' \in \mathrm{R}^D$. We found a U-NET-esque [38] MLP (details in Appendix C.1) performs best, suggesting benefit from deeper and more flexible synaptic computation. It produces *preactivations*, a':

$$a'^t = f_{syn}(\text{concat}(z'^t, o'^t)) \in \mathrm{R}^D$$

where o' is attention output (Section 3.4). The M most recent pre-activations form a ② history A':

$$A'^t = [a'^{(t-M+1)}, \cdots, a'^t] \in \mathrm{R}^{D \times M}$$

Initial pre-activation history and $z'^{t=1}$ are learnable parameters. We found that setting $M \approx 10 - 100$ was effective during our initial exploration.

3.3 Privately-Parameterized Neuron-Level Models (NLMs)

Each neuron $d \in \{1, \cdots, D\}$ has a ③ privately parameterized NLM, g_d (depth 1 MLP of width D_{hidden}), processing its M-dimensional pre-activation history A'^t to produce ④ post-activations:

$$z''^t_d = g_d(A'^t_d)$$

The full set of post-activations z''^t is ④ concatenated with attention output, o', and fed into the synapse model f_{syn} for the next internal tick, $t+1$. See Listing 2 for pseudo-code.

【解读】
这一大段主要介绍了一个非常有趣的神经网络内部构造。为了让你更好地理解，我们可以把这个系统想象成一个**“高效率的学生研讨小组”**。

1. 突触模型（Synapse Model）：研讨组的“交流中心”
首先，文档提到了“突触模型” f_{syn}。在生物学中，突触是神经元之间传递信号的接点。在这里，作者把 f_{syn} 设计成一个复杂的函数（使用了类似 U-Net 结构的 MLP，一种多层感知机）。你可以把它想象成研讨小组的**“公共讨论区”**。

- **输入是什么？** 它接收两类信息：一是小组内部当前的思维状态（z'^t，即神经元在共享空间的状态）；二是外部传来的新知识（o'^t，这是后面会讲到的注意力输出）。
- **做什么？** 它把内部想法和外部知识“拼接”（concat）在一起，然后进行深度的加工和混合。
- **输出是什么？** 产生了“预激活值”（pre-activations, a'）。这就像是大家讨论后产生的一个初步想法或草稿。

2. 历史记录（History A'）：不仅看当下，还要看过去
这部分非常关键。系统不会只根据“现在这一秒”的信息做决定，它会回看过去 M 个时刻的记录。
公式 $A'^t = [a'^{(t-M+1)}, \cdots, a'^t]$ 告诉我们，系统维护了一个长度为 M 的“短期记忆窗口”。

- **类比：** 就像你在做阅读理解时，不会只盯着当前这个字看，而是会联系前一句话或前一段的内容（比如过去 10 到 100 个时刻的信息）。这赋予了模型处理时间序列和上下文的能力。

3. 私有参数化的神经元模型（NLMs）：每个人都有独特的思考方式
这是该模型最独特的地方之一！
在普通的神经网络（如 CNN 或 Transformer）中，大部分神经元共享同一套权重参数（比如卷积核是一样的）。但在 Section 3.3 中，作者提出了“私有参数化”（Privately-Parameterized）。

- **这意味着什么？** 每一个神经元 d 都有它自己专属的小模型 g_d。
- **类比：** 回到我们的研讨小组。普通的神经网络像是一群只会照本宣科的克隆人，大家都用同一本手册处理工作。而这里的 NLM 就像是小组里的每个学生都有自己独特的专业背景和性格。学生 A（神经元 1）可能擅长处理数学，学生 B（神经元 2）擅长处理语言。
- **工作流程：** 每个神经元 d 拿着属于自己的那份历史记录 $A_d^{'t}$，用自己独特的脑子 g_d 进行思考，最后得出一个“后激活值”（post-activation, $z_d^{''t}$）。

4. 闭环（Loop）
最后一段描述了循环是如何完成的：所有神经元思考出来的结果（$z^{''t}$）被收集起来，再次和外部信息 $o^{'}$ 结合，扔回给“突触模型” f_{syn}，从而开启下一个时刻 $t+1$ 的计算。这就形成了一个不断循环、自我更新的动态系统，很像我们大脑中不断的意识流。

【原文】

3.4 Neural Synchronization: Modulating Data and Outputs

Synchronization is inspired by biological brains [39]. The CTM modulates data via the **synchroniza- tion of neural activity**[1]. We first collect post activations into ⑤ $Z^{''t}$ (non-fixed length history):

$$Z^{''t} = [z^{''1}, \cdots, z^{''t}] \in \mathrm{R}^{D \times t}$$

We define neural synchronization as defined as the ⑥ inner product of the histories of each neuron:

$$S^{'t} = Z^{''t} \cdot Z^{''t} \in \mathrm{R}^{D \times D}$$

3.4.1 Neuron Pairing: A Sub-sampling Approach

Since $S^{'}$ scales with $O(D^2)$, it can grow very large. We sample (i, j) neurons at the start of training by randomly selecting D_{out} and D_{action} pairs for two synchronization representations, $S^{out} \in \mathrm{R}^{D_{out}}$ and $S^{action} \in \mathrm{R}^{D_{action}}$. These are projected by W_{out} and W_{in} for outputs $y^{'}$ and attention queries $q^{'}$:

$$y^{'t} = W_{out} \cdot S^{out}$$

$$q^{'t} = W_{in} \cdot S^{action}$$

We use standard cross attention [40] for $o^{'}$:

$$o'^t = \text{Attention}(q'^t, K', \text{FeatureExtractor}(data))$$

where a FeatureExtractor (e.g., ResNet [41]) provides keys/values, $o' \in \mathrm{R}^{D_{input}}$ is then concatenated with z''^{t+1}. This process, including learnable temporal scaling, is shown in Listing 3.

【解读】
这一部分将前面的内部思考机制与外界联系起来，解释了系统如何通过“同步”来感知世界并做出反应。这是一个非常仿生（模仿生物学）的设计。

1. 神经同步（Neural Synchronization）：寻找共鸣
作者引入了一个核心概念：**同步**。在神经科学中有一句名言：“Fire together, wire together”（一起激发的神经元会连在一起）。这里，作者想通过数学方式捕捉神经元之间的这种“共鸣”。

- **公式解读：** Z''^t 是所有神经元从开始到现在的完整历史记录。$S'^t = Z''^t \cdot Z''^t$ 是一个内积运算。
- **这是什么意思？** 在向量代数中，内积（点积）衡量两个向量的相似度或相关性。计算 $Z'' \cdot Z''$ 实际上是在计算**每一对神经元在整个时间轴上的活动相关性**。
- **类比：** 想象一个管弦乐团。S' 就像是一个复杂的统计表，记录了“小提琴手”和“大提琴手”在整场演奏中，有多少次是同时拉响的。这种“同步率”包含了巨大的信息量，代表了系统当前的整体状态。

2. 神经元配对与采样（Neuron Pairing）：解决计算量爆炸的问题
这里遇到了一个数学上的挑战。如果你有 D 个神经元，要计算两两之间的同步关系，你需要计算 D^2 个数值（即 $D \times D$ 的矩阵）。

- **问题：** 如果 D 很大（比如 1000 个神经元），D^2 就是 100 万，计算量太大了，这叫 $O(D^2)$ 复杂度。
- **解决方案：** 这里的“Sub-sampling Approach”采用了**随机采样**策略。就像搞民意调查不需要问遍全国 14 亿人，只需要随机抽取几千个样本就能大概知道结果。
- 作者在训练开始前，随机挑选了一些神经元对子（pairs），只计算这些特定对子的同步情况，形成了两个较小的向量 S^{out}（用于输出）和 S^{action}（用于行动/查询）。这极大地节省了算力。

3. 输出与注意力（Outputs & Attention）：看与做
最后，系统如何利用这些同步信息呢？

- **产生动作 (y')：** 通过矩阵 W_{out} 将同步状态 S^{out} 投影出来，这就是模型的最终输出（比如分类结果或预测值）。
- **产生查询 (q')：** 通过矩阵 W_{in} 将同步状态 S^{action} 转化为“查询向量”（Query）。这就像是系统根据当前的思考状态，决定“我现在应该去关注什么信息”。

- **交叉注意力 (Cross Attention)：** 这是一个标准的 Transformer 组件。系统拿着“查询向量” q'，去外部数据（由 FeatureExtractor 提取，比如用 ResNet 处理过的图片特征）中寻找匹配的信息。
- **结果闭环：** 注意力机制的结果 o'（即系统“看”到的重点信息）会被送回下一轮循环，与内部状态拼接，继续下一秒的思考。

总结：
这整个系统就像一个模拟的大脑：

1. **内部循环：** 神经元们（NLMs）拥有各自的性格，通过突触（Synapse）不断交流，并保留短期记忆。
2. **宏观同步：** 系统通过观察神经元之间长期的协同共振（Synchronization）来理解当前的处境。
3. **交互：** 根据这种共振状态，它决定输出什么结果，以及下一步该去关注外部世界的哪一部分信息。你好！很高兴能以导师的身份为你解读这份关于**计算时间模型（CTM）**的学术文档。这部分内容涉及神经网络如何处理“时间”和“记忆”，以及如何像人类一样根据问题的难易程度“动态”地思考。

这对于高三学生来说是非常棒的思维拓展，因为它结合了你们数学课本里的**指数函数**、**统计学**概念，以及生物学中的**神经元**原理。

我们将文档切分为三个部分来详细解读。

【原文】

Scaling Temporal Dependency.

To modulate the influence of past activity on S', we introduce learnable exponential decay factors $r_{ij} \geqslant 0$ for each neuron pair ij. The rescaling vector over ℓ is:

$$R'_{ij} = [\exp(-r_{ij}(\ell - 1)), \cdots, \exp(-r_{ij}(\ell - t))] \in \mathrm{R}^t$$

The rescaled synchronization is (see Appendix H for efficient recursive computation):

$$S''_{ij} = \frac{\sum_{\ell=1}^{t}(Z''_{\ell,i} \cdot Z''_{\ell,j} \cdot R'_{\ell,ij})}{\sqrt{\sum_{\ell=1}^{t}(R'_{\ell,ij})^2}}$$

Higher r_{ij} bias towards recent ticks ($r_{ij} = 0$ means no decay). Learnable decay rates r_{ij} allow the CTM to modulate synchronization across multiple time scales[2]. Details on

neuron-pair sub-sampling strategies, including recovering snapshot dependencies, are in Appendix C.2.

【解读】
这段话的核心是在讨论**“遗忘”与“记忆”的数学机制**。

试想一下，你在做一道复杂的物理大题。有些信息（比如题目给出的初速度）是你一开始就要记住且一直有用的；而有些信息（比如中间步骤算出的临时变量）可能过一会儿就不重要了。大脑在处理信息时，对不同记忆的“保质期”是不同的。

在这个模型（CTM）中，作者引入了一个概念：**时间依赖性的缩放（Scaling Temporal Dependency）**。

1. **指数衰减因子（Exponential Decay Factors）r_{ij}：**
   大家在高一数学学过指数函数 $y = a^x$。这里使用的是 e^{-rt} 的形式。

   - r_{ij} 是一个**可学习的参数**（learnable parameter），意味着神经网络会在训练过程中自己摸索出这个值应该是多少。
   - 公式 R'_{ij} 展示了一个随时间 ℓ 变化的向量。
   - **关键点**：如果 r_{ij} 很大，指数函数 $e^{-r\cdots}$ 的值会迅速接近 0。这意味着模型会迅速“遗忘”过去的信息，只关注最近发生的（bias towards recent ticks）。
   - 反之，如果 $r_{ij} = 0$，那么 $e^0 = 1$，意味着没有任何衰减，过去的记忆和现在的记忆一样重要，这代表了“长期记忆”。

2. **重新缩放的同步性（Rescaled Synchronization）S''_{ij}：**
   那个看起来很复杂的 S''_{ij} 公式，其实本质上是一个**加权平均**。

   - 分子部分是神经元活动 Z 的乘积，乘以了我们刚才说的时间权重 R。
   - 分母部分是为了归一化（Normalization），确保数值不会无限变大。
   - 这就像你们算加权平均分一样，期末考权重高，平时作业权重低。这里是“离现在越近的时间点，通常权重越高（除非 $r = 0$）”。

3. **多时间尺度（Multiple Time Scales）：**
   作者提到，通过学习不同的 r_{ij}，模型可以同时处理快慢不同的节奏。这就像你能一边哼着快节奏的歌（短时间尺度），一边在大脑里规划下周的复习计划（长时间尺度）。

总结来说，这一段是给神经网络装上了一个“可调节的记忆过滤器”，让它能够自主决定哪些信息该立马忘掉，哪些信息该铭记在心。

【原文】

Loss Function: Optimizing Across Internal Ticks

The CTM produces outputs $y^{'} \in \mathrm{R}^{C}$ (e.g., class probabilities) at each internal tick t. We compute a loss $\mathrm{L}^{t} = \mathrm{CrossEntropy}(y^{'t}, y_{true})$ and certainty C^{t} (1-normalized entropy) per tick. For each forward pass we select to ticks:

1. The point of minimum loss: $t_{L} = \mathrm{argmin}_{t}(\mathrm{L}^{t})$, to select the 'best' prediction; and

[1]We found that 'snapshot' representations were too constraining: projecting from $Z^{'}$ strongly ties it to the downstream task and thereby limits the types of dynamics it can produce, whereas synchronization decouples it. For full details we have found that the CTM learns to average this for ImageNet (Section 5) but more so for 2D mazes (Section 4), suggesting task-dependent temporal sensitivities

The final loss for optimizing θ_{syn} and $\theta_{d-1:D}$ is:

$$L = \frac{(L_1 + L_2)}{2}$$

Since t_1 and t_2 are dynamically defined per data point, the CTM can attribute variable compute (internal ticks) to different data points as needed without explicit restrictions on which tick should be used in the loss function. This effectively implements native adaptive computation [18] as opposed to a post-hoc addition. We give pseudocode in Listing 2.

【解读】
这一段非常精彩，它解释了模型是如何**“像人类一样思考”**并进行自我优化的。

1. **内部时刻（Internal Ticks）与持续输出**：
   传统的神经网络通常是“输入图片 -> 瞬间 -> 输出结果”。但这个CTM模型不一样，它有一个内部的时间轴（Ticks）。就像你在考场上解题，第1秒你可能没思路，第10秒有个模糊的想法，第30秒确信了答案。模型在每一个时刻 t 都会输出一个预测 $y^{'}$。

2. **损失函数（Loss Function）**：
   这是机器学习的核心，相当于“评分标准”。

   - L^{t} 使用了**交叉熵（CrossEntropy）**，这是高三统计学中尚未深入涉及但AI中极基础的概念，你可以简单理解为衡量“预测答案”和“标准答案”之间差距的指标。
   - 模型会在所有时刻中寻找**损失最小**的那个时刻 t_{L}，也就是它表现最好的那一瞬间。

3. **脚注中的洞察（关于 Snapshot vs. Synchronization）**：
   夹在中间的那段脚注（[1]）其实道出了设计的哲学。作者发现，如果只是简单地对神经元的状态拍个“快照”（snapshot），会限制模型的灵活性。而使用“同步性”（synchronization）能让模型从具体的任务中解耦出来。

- 类比：如果你背书只是死记硬背（Snapshot），换个题型你就不会了。如果你掌握了知识点之间的逻辑联系（Synchronization），你就能举一反三。作者发现在走迷宫（2D mazes）这类复杂任务中，这种灵活性尤为重要。

4. **原生自适应计算（Native Adaptive Computation）**：
   这是本段的**高光时刻**。

   - 原文提到 t_1 和 t_2 是针对每个数据点**动态定义**的。
   - 这意味着：对于简单的图片，模型可能在第2个tick就给出了最佳答案，计算就结束了；对于复杂的迷宫，它可能要算到第100个tick。
   - **核心意义**：模型不需要人为规定“必须思考5秒钟”。它实现了**“按需分配算力”**。这就像考试时，简单的填空题你花30秒，压轴大题你花20分钟。这种机制不是后来补丁加上去的（post-hoc），而是模型天生自带的（native）。

总结：这段描述了模型如何通过评估自己每一刻的表现来学习，并且它学会了对于不同的难题投入不同时长的“思考时间”。

【原文】

Experimental Evaluation

The following sections present a focused evaluation of the CTM on tasks that highlight its core principles: neuron-level temporal processing and neural synchronization as a direct latent representation. We aim to demonstrate how neural dynamics enable the CTM to implement complex reasoning or adaptive processing, while yielding interpretable strategies. We prioritize in-depth in three key experiments: 2D maze navigation, ImageNet-1K classification, and parity computation. We also summarize and highlight additional experiments demonstrating the CTM’ s broader capabilities.

4 2D Mazes: Complex Sequential Reasoning and Internal World Models

【解读】
这一段是实验部分的开篇，也就是**“实战检验”**环节。在提出了前面那些高大上的数学模型（衰减因子、自适应计算）之后，必须通过实验来证明它真的有效。

1. **实验目的**：
   作者想要验证CTM的两个核心原则：

   - **神经元级别的时间处理**：微观层面，神经元随时间的变化是否有意义。

- **神经同步作为潜在表示**：宏观层面，神经元之间的协同工作是否代表了某种知识。
- 更重要的是，他们希望看到**“可解释的策略”（Interpretable strategies）**。现在的很多AI像个黑盒子，我们不知道它怎么想的。作者希望CTM能让我们看到它的“思考过程”。

2. **三大关键实验**：
   作者精心挑选了三个任务，分别对应三种不同的能力，就像高三的三门主课一样：

   - **2D Maze Navigation（二维迷宫导航）**：对应**逻辑推理和规划**。走迷宫需要记住走过的路，规划未来的路，这需要强大的序列推理能力和建立“内部世界模型”（Internal World Models）。这就是为什么第4节的标题特别强调了“复杂序列推理”。
   - **ImageNet-1K Classification（图像分类）**：对应**感知能力**。这是AI界的“高考”，看模型能不能准确识别出图片里是猫还是狗。
   - **Parity Computation（奇偶性计算）**：对应**基础运算逻辑**。这是一个经典的计算问题，测试模型处理抽象逻辑的能力。

3. **为什么强调“内部世界模型”？**
   在第4节的标题中，提到了“Internal World Models”。这对于理解现代AI非常重要。这意味着模型不仅仅是在做“输入A -> 输出B”的简单反射，而是在它的“脑海”里构建了一个迷宫的地图。它可以像人类一样，在脑子里预演怎么走，而不是只会乱撞。

总结：这一部分预告了模型将在视觉、逻辑和规划三个维度接受挑战，特别是通过迷宫任务来展示它像人类一样构建心理地图和进行长时间推理的能力。你好！很高兴能为你解读这份关于人工智能前沿研究的文档。这是一篇关于深度学习模型如何像人类一样思考、规划和解决空间问题的学术论文片段。

为了让你更好地理解，我将这段复杂的学术文本分成了两个主要部分。我们会先看实验的**背景与设置**，然后再看令人兴奋的**结果与分析**。

让我们开始第一部分。

【原文】
In this section we analyze the CTM’s capacity for sequential reasoning, planning, and spatial understanding using a challenging phrasing of the 2D maze navigation task. Solving mazes can be easy with the right inductive bias. For example, matching the output dimensions to the input space; a model can perform blurry classification at each location. Such a setup is amenable to machines by design, as they can learn iterative algorithmic solutions [37, 42], but this is not how humans solve mazes.

Setup. The setup of our maze task deviates from the norm, specifically to necessitate the formation of an internal world model [43] by (1) requiring a direct sequence-of-actions output and (2) disallowing positional embeddings in the visual input. This requires a model to build its own spatial representation via introspection (see Appendix D.6 for

further discussion). We compare the CTM against LSTM and feed-forward (FF) baselines. For the results that follow, we trained a CTM, LSTMs ($L = 1$ and 3 layers), and a FF baseline to predict up to 100 steps down the path of 39 × 39 mazes, where predictions took the form of a sequence of 'up', 'down', 'left', 'right', and 'wait' moves, and 'waiting' meant to pass until the next state. For the CTMs and LSTM baselines, we used 75 internal ticks. The LSTM usually means memory using 50 ticks to predict superior performances so we report those. In each case, we used a automatic curriculum approach with increasing maze difficulty (see details in Appendix D.3). Appendices D.2 and D.4 detail hyper-parameters for the CTM and baselines.

【解读】
这一段非常有意思，它实际上是在给AI设一个“局”，来看看它到底是不是真聪明。

首先，你要明白，让AI走迷宫其实可以很简单。如果在设计AI时加入了很强的“归纳偏置”（Inductive Bias）——这就像是给学生考试前划重点，或者给AI开了“天眼”——比如让AI把迷宫看作一张图，直接在图上画线（像素分类），那AI只要算出每一步的最短路径就行了。这种方法虽然快，但它是纯粹的计算，不具备真正的“思考能力”，也不是人类走迷宫的方式。人类走迷宫时，是置身其中的，我们需要规划：“前面左转，然后再右转”。

所以，研究者设计了一个**地狱难度的迷宫任务**，目的是为了逼迫这个名为**CTM**（我们可以暂且理解为一种模仿大脑认知机制的新型AI模型）去建立一个**“内部世界模型”**（Internal World Model）。

怎么个“地狱难度”法呢？主要有两个限制：

1. **必须输出动作序列**：AI不能只画一条线，它必须像玩游戏一样，一步步说出“上、下、左、右”或者“等待”。这对逻辑推理和规划能力要求极高。
2. **禁用位置编码（Positional Embeddings）**：这一点非常关键。在高三物理或数学中，我们习惯建立坐标系（x, y）。通常的AI也会偷偷利用坐标信息（比如“我现在在坐标(5, 5)的位置”）。但在这里，研究者把这个“GPS”关掉了。AI只能看到迷宫的墙壁，却不知道自己在绝对坐标系的哪里。这就要求AI必须通过**“内省”**（Introspection），自己在脑子里构建一张地图，记住“我刚走了两步，现在应该在路口”。

为了证明CTM的厉害，研究者找来了两个“陪练”作为基准线（Baselines）：

- **LSTM**（长短期记忆网络）：这是以前处理序列问题（比如翻译句子）的王者，有单层和三层的版本。
- **FF**（前馈神经网络）：最基础的神经网络，有点像单纯的条件反射。

比赛规则是：在39x39的迷宫中，预测接下来最多100步怎么走。为了公平，所有模型都采用了一种**“自动课程学习”**（Automatic Curriculum Approach）的方法。这就像你们复习备考一样，先做简单的基础题，等掌握了再慢慢增加难度，做复杂的综合题，而不是上来就扔给你一道压轴题。

这段话的核心在于：研究者故意剥夺了AI的辅助工具（GPS坐标），强迫它像人一样，通过记忆和推理在脑海中构建迷宫的结构。

【原文】

4.1 Results

The CTM significantly outperforms the baselines in solving these mazes, demonstrating superior trainability and generalization to longer paths (Figure 4). The FF model and LSTMs struggled to learn effectively or overfit (see Appendix D.5), whereas the CTM achieved high accuracy. This suggests that the CTM’ s architecture, particularly its use of neural dynamics and synchronization, is well suited for tasks requiring robust internal state maintenance and planning.

4.2 Demonstrations and Generalization

Qualitative analysis shows the CTM methodically tracing paths (Figures 1a and 1b; supplementary video *mazes.mp4*), exhibiting emergent behaviors such as continuing to explore paths beyond its training horizon. This suggests the CTM learns a general procedure rather than simple memorizing. Furthermore, the CTM, trained on 39 × 39 mazes, generalizes effectively to longer paths and larger 99 × 99 mazes (Figure 1c) by applying its learned policy, as shown in Figure 1c (see supplementary video: *maze_large_1.mp4* to *maze_large_4.mp4* for examples). Crucially, this CTM is not using any positional embedding, meaning that in order for it to follow a path through the maze it must craft the

【解读】
这一部分揭晓了比赛结果，CTM可以说是“完胜”。

4.1 结果分析：
CTM的表现显著优于那两个“陪练”（LSTM和FF）。

- **陪练的问题**：FF模型和LSTM模型要么根本学不会（struggled to learn），要么就是**“过拟合”**（overfit）。作为高三学生，你们对“过拟合”应该不陌生——这就好比一个学生背下了《五年高考三年模拟》里所有题目的答案，但并没有理解解题思路。一旦题目数字稍微变一点，他还是不会做。LSTM在这里表现不佳，说明它很难长时间维持这种复杂的空间记忆。
- **CTM的优势**：CTM不仅准确率高，而且它的成功归功于**“神经动力学和同步”**（neural dynamics and synchronization）。你可以把这想象成大脑神经元之间的协同共

振。它不是死记硬背，而是像人脑一样，通过神经活动的节奏来维持一种“思维状态”，这让它非常擅长做长期规划和保持内部状态。

4.2 演示与泛化能力（这是最精彩的部分）：
这一节展示了CTM到底有多像人。

1. **涌现行为（Emergent Behaviors）**：研究者观察到，CTM在走迷宫时表现得非常有条理（methodically tracing paths）。最神奇的是，它展现出了一种“探索”的本能，甚至走出了训练范围之外的路径。这说明它学会的不是“这道题的答案”，而是一种“通用的解题方法”（general procedure）。就像你学会了勾股定理，不管是多大的三角形你都能算，而不是只背下了边长3、4、5的三角形。
2. **强大的泛化能力（Generalization）**：这是AI领域最看重的指标。CTM是在 **39x39** 的小迷宫上训练的，但当研究者把它扔进 **99x99** 的超大迷宫（比训练时大得多、路径长得多）时，它依然能从容应对！这就好比你在学校的小操场学会了开车，结果把你放到复杂的城市立交桥上，你依然开得很稳。
3. **核心结论**：最后那句未完的话强调了一个关键点——别忘了，这个CTM是没有使用“位置编码”（GPS）的。这意味着，它能在99x99的大迷宫里不迷路，完全是因为它学会了如何在脑子里**“手工打造”**（craft）位置信息。它通过观察环境和记忆自己的动作，构建出了属于自己的坐标系。

总结来说，这段文本展示了一个非常接近人类思维方式的AI模型。它不是靠死记硬背数据，而是通过理解规则、构建心理模型来解决从未见过的难题。这对于人工智能从“弱人工智能”（只会做特定任务）向“通用人工智能”（像人一样思考）迈进，是非常重要的一步。你好！我是你们的学术导师。很高兴能带大家一起研读这篇关于 **CTM（Continuous Thought Model，连续思维模型）** 的前沿学术论文片段。

这篇文档探讨的是一种让人工智能像人类一样拥有“思维过程”的新尝试。为了方便大家理解，我们将原来的文本拆分为几个核心部分，我会像带大家做精读理解一样，把其中的专业术语转化为我们高中熟悉的知识体系。

咱们开始吧！

【原文】
Figure 1: CTM versus baselines on 2D mazes. The CTM demonstrates superior trainability compared to baselines yielding higher accuracy for longer paths. Using iterative re-applications, we show in (b) that the CTM can generalise to longer paths and bigger mazes. See Appendix D.5 for loss curves.

cross-attention query by 'imagining' the future state of the maze; a process known as 'episodic future thinking' [44] in humans. Appendix I discusses some of the emergent properties we observed.

5 ImageNet-1K Classification: Adaptive Processing and Emergent Dynamics

We evaluate the CTM on ImageNet-1K to understand its internal processing dynamics when trained to solve a standard classification task. We are not yet aiming for state-of-the-art accuracy (with 50 internal ticks and a ResNet-152 backbone: 72.47% top-1, 89.89% top-5 on uncropped data). Since the CTM uses new neural computation principles, it would require a thorough hyperparameter search to find the optimal settings, and this is outside the scope of this work. Instead, we focus on *how* the CTM leverages neural dynamics (setup details in Appendix E.1) as a new mechanism for reasoning.

【解读】
同学们，这一段落包含了两个截然不同的实验场景：一个是走迷宫（Maze），一个是看图识物（ImageNet分类）。

首先，让我们看看“走迷宫”。大家想象一下，当你置身于一个复杂的迷宫中，你会怎么做？你可能会在脑海里先“预演”一下：“如果我往左走，前面好像是死胡同；往右走，似乎能通向出口。” 文档中提到的 **CTM** 模型也是这么做的。它通过一种叫做“交叉注意力查询（cross-attention query）”的机制，去**“想象”**迷宫未来的状态。这在心理学和神经科学上有一个很酷的名字，叫**“情景未来思维（episodic future thinking）”**。就像人类在做决定前会在脑中模拟未来情景一样，这个AI学会了不再盲目碰撞，而是先“想”后“动”。这种能力让它在处理长路径和更大迷宫时，比传统的基础模型（Baselines）表现得更好，准确率更高。

接下来，文章的话题转到了大家可能听说过的 **ImageNet-1K**。这是一个包含1000个类别、上百万张图片的巨大数据库，被称为AI界的“高考”。通常，大家在ImageNet上比拼的是谁的准确率更高（即State-of-the-art accuracy，SOTA）。但在这里，作者非常坦诚地告诉我们：他们这次的目标**不是**为了刷分拿第一。

这就好比一位短跑教练研发了一种全新的跑步姿势，他现在关注的不是马上破世界纪录，而是研究这种新姿势下，运动员的肌肉是如何发力的（即内部处理动力学，Internal processing dynamics）。

文中提到了一些技术参数，比如“ResNet-152 backbone”。大家可以把ResNet-152想象成人的“视网膜”或“眼睛”，它是负责提取图像特征的基础网络；而CTM则是负责处理这些信息的“大脑”。虽然目前的准确率（Top-1 72.47%）还不是业界最高，但这是因为新模型采用了全新的神经计算原理，还没来得及做精细的参数调优（Hyperparameter search）。作者强调，目前的重点是研究CTM是**如何**利用神经动力学来进行**推理（Reasoning）**的。这是一种从“结果导向”到“过程导向”的研究思维转变，对于理解AI的本质非常重要。

【原文】

5.1 Adaptive Computation and Calibration

Figure 5: ImageNet-1K results: (a) Native adaptive compute potential based on a 0.8 certainty threshold, showing performance expected at each internal tick. (b) Excellent model calibration when averaging probabilities up to each tick shown. See Appendix E.3 for further analysis.

The CTM exhibits adaptive computation (Figure 5a), the synchronization of which is the representation with which it observes data and forms predictions. We show in Figure 2b how the CTM learns to 'hover' around an image in order to gather information and make a prediction. It does this entirely without prompting or any guide, implementing computationally beneficial adaptive compute of the CTM. In this scenario, the CTM will make a prediction and stop if it is certain (above 0.8 confidence) after a fewer number of internal ticks of compute. If the CTM is less certain, it performs additional internal ticks, yielding greater accuracy in the final prediction (Figure 5a) and excellent calibration of probabilities, meaning that a prediction with 0.75 confidence is correct 75% of the time (Figure 5b). The CTM also demonstrates a consequent excellent calibration of probability based on its iterative refinement process (Appendix E.3).

【解读】
这一段非常精彩，它揭示了CTM模型最核心的两个类人特质：**自适应计算（Adaptive Computation）和校准（Calibration）**。

咱们先说**“自适应计算”**。现在的很多AI模型，无论处理什么问题，用的“脑力”都是一样的。这就好比不管是做“1+1=？”这道题，还是做“解析几何压轴题”，你都必须强制自己思考整整5分钟才能写答案。这显然不合理，对吧？

CTM模型就聪明多了，它学会了“看菜下碟”。文中提到的“Internal ticks”可以理解为AI思考的“时间步”或“念头”。

1. **简单的图：** 如果CTM看了一眼（经过少数几个ticks），确信度（Confidence）超过了0.8（即80%把握），它就会立刻停止思考，给出答案。这不仅省时，还省算力。
2. **复杂的图：** 如果它觉得这图模棱两可，不太确定，它就会自动进行更多的“ticks”，也就是多想一会儿，直到它看懂为止。

文中用了一个很生动的词叫 **"Hover"（盘旋/徘徊）**。想象一下你在看一幅复杂的抽象画，你的视线会在画面上游移，收集信息，直到你理解它。CTM也是这样，它不需要人类教它（without prompting），自己就学会了在图像周围“盘旋”以收集信息。

接下来是**“校准（Calibration）”。**这个概念在AI和统计学中非常重要，它代表了模型的**“诚实程度”或“自知之明”**。

- 如果一个学生每道题都说“我100%确定做对了”，结果只考了50分，那他的“校准”就很差（过度自信）。

- 如果CTM说“我有75%的把握这张图是猫”，而我们在大量类似的预测中统计发现，它确实有75%的时候是对的，那么我们就说它的**概率校准极佳（Excellent calibration）**。

这说明CTM不仅能根据题目难易程度调整思考时间（迭代优化过程），而且它对自己的判断能力有非常精准的认知。这种机制让AI变得更像一个成熟的思考者，而不是一个只会死记硬背的机器。

【原文】

5.2 Reasoning sequentially about static images

The CTM exhibits diverse temporal dynamics (Figure 2a), the synchronization of which is the representation with which it observes data and forms predictions. We show in Figure 2b how the CTM learns to 'hover' around an image in order to gather information and make a prediction. It does this entirely without prompting or any guide, implementing computationally beneficial adaptive

【解读】
虽然这段原文最后中断了，但结合标题**“关于静态图像的顺序推理（Reasoning sequentially about static images）”**和前面的内容，我们依然可以深入解读其中的奥妙。

这里探讨的一个核心矛盾是：**图像是“静态”的，但思维是“动态”的。**

通常我们认为，看一张照片是“一瞬间”的事。但在神经科学中，人类视觉处理其实是一个序列过程（Sequential Process）。当你看到一张“操场上有人打篮球”的照片时，你的大脑并不是在一个毫秒内同时也同等清晰地处理了所有像素。相反，你可能会先注意到篮球，然后是投篮的人，再是防守的人，最后是背景。这有一个**时间顺序（Temporal dynamics）**。

CTM模型试图模仿这种机制。

1. **多样化的时间动态（Diverse temporal dynamics）：** 这意味着模型在处理不同图片时，其内部状态的变化节奏是丰富多样的，而不是千篇一律的机械反应。
2. **同步（Synchronization）：** 这里指的是模型内部的计算节奏与它观察数据、形成预测的过程是同步协调的。

再次提到的 **"Hover"（盘旋）** 在这里有了更深的含义。对于一张静止的图片（比如一只躲在草丛里的豹子），CTM通过在图片特征上“盘旋”，实际上是在**用时间换取空间**上的理解深度。它通过一步步的顺序推理，逐渐拼凑出完整的图像信息。

这给我们的启示是：CTM不再把图像识别看作是一个简单的输入输出函数（Input-Output Mapping），而是看作一个**随时间演变的动力学过程**。这种处理方式让AI在面对模糊、遮挡或者复杂的视觉场景时，能够像人类一样，通过“多看几眼”、“多想一会儿”来获得更准确的结

论。这种**“计算有益的自适应性（computationally beneficial adaptive...）”**正是下一代AI试图突破的关键方向。【原文】
compute in an intuitive fashion. . This internal process can even manifest emergent phenomena like low-frequency traveling waves [45] across UMAP-projected neuron activations (see supplementary video *umap.mp4*). Unpacking every increasing facet of these attention map progressions is simply infeasible for a static form; we encourage viewing supplementary video *attention.mp4* for demonstrations of the CTM 'gazing' in a manner not quite entirely unlike how humans might look around images. Appendix E.4 has further demos and UMAP visualizations. These observations underscore that the CTM solves classification by leveraging an internal, dynamic reasoning process, a departure from typical feed-forward approaches.

6 Parity: Learning Sequential Algorithms and Interpretable Strategies

To test the CTM’ s ability to learn algorithmic procedures and develop interpretable strategies, we use a cumulative parity task: given a 64-length binary sequence, predict the parity at each position (Figure 6a). Unlike prior work focusing on final parity [18], our setup requires the model to output sequences at each internal tick, enabling us to examine how the full output evolves across ticks and throughout training. Setup details are in Appendix F.1.

【解读】
同学们，在这一部分，我们要先给前面关于图像识别的讨论做一个精彩的收尾，然后立刻进入一个全新的、更考验逻辑思维的挑战领域。

首先，让我们回顾一下刚才提到的“直觉式计算”。作者在这里描述了一种非常迷人的现象：当我们观察CTM（连续思维模型）内部神经元的活动时，竟然发现了一种“涌现现象”（Emergent Phenomena）。什么是涌现？简单来说，就是无数个简单的个体协同工作时，突然产生了一种宏大的、未被预设的复杂行为。作者提到，如果你把这些神经元的活动通过UMAP（一种把高维复杂数据“拍扁”成二维地图以便观察的技术）可视化出来，你会看到类似“低频行波”的东西。想象一下，大脑在思考时，脑电波像水面的涟漪一样扩散，CTM在这个瞬间竟然展现出了类似生物大脑的波动特征，这简直是人工智能拥有“动态思维”的物理证据。

更有趣的是，作者用了一个很生动的词——“凝视”（Gazing）。传统的AI看图通常是一口吞，瞬间处理整张图。但CTM不一样，它处理图片的过程像极了人类：先看这里，再看那里，眼神在图像上游走。这种动态的注意力转移过程，证明了它不是在死记硬背，而是在进行即时的、动态的推理。这就好比你在做一道几何题，你的眼睛会在辅助线、角度和已知条件之间来回扫描，这就是一种“动态推理过程”，完全不同于传统那种输入立刻得输出的“前馈”（Feed-forward）反射式AI。

紧接着，文章话锋一转，进入了第6章——“奇偶校验任务”（Parity Task）。如果说刚才看图是在考“美术”，现在就是要考“数学逻辑”了。

为什么要考这个？因为作者想知道，这个模型是不是真的能学会一种“算法”，也就是解决问题的固定步骤。这里的任务是“累积奇偶校验”：给你一串由0和1组成的长度为64的序列，你要从头走到尾，每走一步都要回答“到现在为止，1的个数是奇数还是偶数？”。

这听起来很简单，对吧？但对AI来说其实很难。因为这要求模型必须拥有极好的“短期记忆”和“逻辑连续性”。如果你忘了前面数了几个1，后面就全错了。而且，作者设计的这个实验比以前的研究更难，以前只要求最后给个总结果（是不是奇数），现在要求模型在每一个“时间滴答”（Tick，可以理解为思考的最小时间单位）里都要输出当前的序列状态。这就好比老师不光看你卷子最后的答案，还要盯着你的草稿纸，看你每一步是怎么推导出来的，以此来观察模型的思维是如何随着时间推移而进化的。

【原文】

6.1 Results and Learned Strategies

Figure 6: CTM performance on the parity task: (a) example; (b) training accuracy comparisons; (c) impact of internal ticks on accuracy; and (d) an example showing how this CTM uses at least one attention head to scan the input sequence from start to end. Error bars (b, shaded) represent 1 standard deviation over seeded runs. Appendix F.2 discusses the implications of seed variations.

The CTM’ s accuracy improves with more internal ticks, significantly outperforming parameter-matched LSTMs, which struggled with parity task performance (Figure 6b). LSTMs with 75 and 100 ticks could achieve perfect accuracy if some seeded runs (Figure 6d shows how the attention shifts over the input data, and Figure 7 shows a specific demonstration of 4 attention heads), revealing a distinct and interpretable strategy. Which specific style of solution depends on the configuration and seed, so we show other examples and analyses in Appendix F.2). Crucially, this experiment demonstrates that the CTM can learn to form and follow an internal strategy for an algorithmic task.

Figure 7: Determining parity: (a, b, c) are the trajectories of the argmax of attention for 4 heads and the corresponding prediction at different internal ticks, and (d) is the target (perfectly predicted here). See supplementary material 'parity.mp4' for video format.

7 Other Experiments and Analyses

We also evaluated the CTM in a number of other settings in order to probe its functionality and versatility. Owing to space constraints, we provide the details of these additional

experiments in the appendices (referenced below). In summary, these additional experiments investigated:

【解读】
这段内容非常核心，它展示了CTM在逻辑任务上的“考试成绩”以及它是如何“解题”的。

首先看成绩单（6.1 Results）。结论非常令人振奋：CTM给出的“思考时间”越长（即internal ticks越多），它的准确率就越高。这就好比考试时，我多给你几分钟思考，你就能把题做对。与之形成鲜明对比的是LSTM（长短期记忆网络），这是以前处理序列任务的霸主模型。但在参数量相同的情况下，LSTM在这个任务上表现得很挣扎，可以说被CTM完爆。这说明CTM的架构在处理这种需要严密逻辑步骤的任务上，具有代际优势。

最精彩的部分在于“可解释性策略”（Interpretable Strategies）。大家知道，现在的AI常被诟病为“黑箱”，我们知道它给出了答案，但不知道它为什么这么想。但在这里，研究人员通过图6d和图7看到了惊人的一幕：

CTM的一个“注意力头”（Attention Head，可以理解为模型的“眼睛”或“聚光灯”），竟然在输入序列上从头到尾进行了一次完美的扫描！

想象一下，你让AI数一串珠子里的红珠子是奇数还是偶数。普通的AI可能就是盯着一堆数据瞎猜。但CTM被观察到，它的“注意力焦点”随着时间推移，像咱们人类的手指指读一样，从序列的第一个数字移到了最后一个数字。图7展示了这一过程的轨迹：注意力头按部就班地移动，预测结果也随之精准变化。

这意味着什么？这意味着CTM不仅仅是在拟合数据，它是真的“学会”了算法。它自己发明了一种策略：“我要从左到右一个一个看，每看到一个1就切换一下奇偶状态”。这种能形成并遵循内部策略的能力，是通向通用人工智能（AGI）的重要一步，因为它证明了机器拥有了类似人类的“程序性思维”。

最后，第7章的开头简要提到，除了这个奇偶校验，研究团队还把CTM扔到了许多其他场景里去“通过压力测试”，以验证它的多才多艺（Versatility）。虽然因为篇幅限制（论文只有那么多页），详细内容被放到了附录里，但这暗示了CTM不仅仅是个偏科生，它在多种任务上都有潜力。

总结一下，这一大段告诉我们：CTM不仅视力好（能看图），逻辑还好（能做算法题），而且最重要的是，它的思考过程是透明的、有迹可循的，像人类一样按步骤解决问题。这对于我们理解AI到底在想什么，提供了绝佳的窗口。【原文】

CIFAR-10 Classification Compared to Human

CIFAR-10 (Appendix G.1): The CTM, feed-forward, and LSTM baselines were trained on CIFAR-10, with results compared against human data for difficulty and uncertainty. The CTM demonstrated good model calibration and alignment with humans.

CIFAR-100 Ablation Studies

CIFAR-100 (Appendix G.2): We investigated the impact of model width and the number of internal ticks. We found that the diversity of neural activity are functions of these. Wider models tended to exhibit more varied neural dynamics. Using more internal ticks allowed the CTM to engage in extended processing, sometimes revealing distinct computational phases.

Neuron-Level Models and Synchronization Ablations

Neuron-Level Models and Synchronization Ablations (Appendix G.3): We compared the CTM to parameter-matched variants without NLMs and without synchronization, as well as an LST with synchronization. The results show that the combination of neuron-level models and synchronization as a representation is key to the success of the CTM.

Sorting Real Numbers

Sorting Real Numbers (Appendix G.4): The CTM was tasked with sorting sequences of 30 real numbers, outputting sorted indices sequentially using a Connectionist Temporal Classification (CTC) loss [46]. This experiment showed that the CTM could learn an algorithmic sorting procedure and exhibited adaptive computation by varying its internal processing duration ('wait count') based on characteristics of the input sequence, such that the difference between successive values.

Q&A MNIST

Q&A MNIST (Appendix G.5): In this task, the CTM processed sequences of MNIST digits followed by index and operator embeddings to perform multi-step modular arithmetic. This investigation highlighted the CTM’ s capacity for memory and retrieval, using its synchronization mechanism to recall digit information beyond the immediate attention window of individual neuron-level models, and to generalize to longer computational sequences than seen during training.

Reinforcement Learning

Reinforcement Learning (Appendix G.6): The CTM was adapted for reinforcement learning in several partially observable Markov decision processes (POMDPs), including classic control (CartPole, Acrobot) and grid-world navigation (Navigational Four Rooms). This highlighted the CTM’ s applicability to sequential decision-making in continuous interaction settings, where it achieved performance comparable to LSTM baselines while developing richer internal state dynamics.

8 Discussion and Conclusion

【解读】
各位同学，大家好！今天我们要解读的这段文档，是对一个名为 **CTM**（Communicating Transformer Model，我们可以暂时把它想象成一种新型的“人工智能大脑”）进行全面体检的报告。这部分内容位于论文的附录中，展示了研究人员如何通过一系列不同类型的“考试”来验证这个模型的真正实力。这就像是高三模拟考，不能只考语文，还要考数学、理综，甚至还要测体育，以此来全面评估一个学生的综合素质。

首先是**视觉识别能力的考试**（CIFAR-10 和 CIFAR-100）。
研究人员让 CTM 去识别图片（CIFAR-10 数据集）。这里有一个非常有意思的指标叫“模型校准（Model Calibration）”。这是什么意思呢？就好比你做选择题，不仅要选对，还要看你对答案的“确信程度”是否合理。如果一个学渣瞎蒙对了答案，但他自己觉得是瞎蒙的，这叫有自知之明；如果他明明错了却坚信自己是对的，这就叫“校准”很差。结果显示，CTM 不仅识别得准，而且它的自信程度与人类非常接近，这说明它具有很好的类人认知特性。
而在更难的 CIFAR-100 测试中，研究人员进行了**消融实验（Ablation Studies）**。这就像控制变量法，他们调整了模型的“宽度”（脑容量）和“内部时钟滴答数”（思考时间）。结果发现，给模型更多的时间去“思考”（更多的 internal ticks），它就能进行更深度的处理，甚至展现出明显的思维阶段。这就像我们在解难题时，多想一会儿，思路会更清晰。

其次是**核心组件的必要性测试**。
在“神经元级模型与同步消融”这一节中，研究人员试图回答一个问题：CTM 这么强，是因为它独特的结构，还是单纯因为参数多？于是他们设计了“阉割版”的模型——去掉了 NLM（神经元级模型）或去掉了同步机制。结果证明，只有当这两者结合时，模型才能成功。这就像组装一台电脑，光有顶级 CPU 不行，还得有好的主板把它们连起来（同步机制），缺一不可。

接下来是**逻辑与算法能力的考试**。
研究人员让 CTM 去**排序实数**。这可不是简单的死记硬背，而是要学会“冒泡排序”或者“快速排序”这样的算法逻辑。这里最精彩的一点是 CTM 展现了**自适应计算（Adaptive Computation）**。就像你们做数学题，简单的题一眼看出答案，难的题要演算半天。CTM 也是如此，当需要排序的数字靠得很近、很难分辨时，它会自动增加“等待计数（wait count）”，也就是多花点时间思考再输出。这种根据难度动态调整算力的能力，是非常高级的智能体现。

紧接着是**记忆与多步推理的考试**（Q&A MNIST）。
这个任务要求模型看一串手写数字图片，然后根据指令做数学运算。这考验的是**记忆检索**能力。模型不仅要认出当前的数字，还得记起刚才看过的数字（哪怕它已经不在眼前的关注窗口里了）。CTM 利用它的同步机制成功地把之前的记忆“捞”了出来，甚至还能处理比训练时更长的算式，展现了强大的举一反三能力。

最后是**体育课——强化学习（Reinforcement Learning）**。
研究人员把 CTM 放到游戏环境（如 CartPole 平衡杆、迷宫导航）中，看它能否像玩游戏一样做出连续决策。这里提到的 **POMDP**（部分可观测马尔可夫决策过程）是一个专业术语，其实很好理解：这就好比你在玩《王者荣耀》或者《吃鸡》，你看不到全图（部分可观测），只能根

据眼前的信息和记忆来判断下一步往哪走。CTM 在这种动态环境下表现得和经典的 LSTM 模型一样好，而且其内部的思维状态更加丰富。

总结一下：这段文档通过从静态图像识别、动态算法逻辑、长期记忆检索到连续决策控制等多个维度的测试，有力地证明了 CTM 架构的通用性和先进性。它不只是一个只会认图的“偏科生”，而是一个能根据问题难度调整思考时间、拥有良好记忆机制的全能型“学霸”。你好！很高兴能以学术导师的身份，为高三的同学们解读这篇关于人工智能前沿研究的文档。这是一篇关于“连续思维机器”（Continuous Thought Machine, CTM）的论文片段。

在高三物理或生物课上，你们可能学过神经元是如何通过电信号交流的，或者在数学课上接触过算法逻辑。今天我们要看的这篇文档，就是试图把生物大脑的运作机制（特别是时间上的动态变化）引入到计算机的深度学习中。

为了让大家更好地理解，我将文档分成了两个部分进行详细解读。

【原文】

The Continuous Thought Machine (CTM) represents a new perspective, where the temporal dynamics of neural activity are central to artificial cognition. Its core innovations—neuron-level models and synchronization as a latent representation—effectively enable it to both unfold and leverage neural dynamics to solve problems. We showed in this work that such an approach is not only feasible but also leads to unique computational capabilities and emergent properties.

Our experiments demonstrate that the CTM can effectively solve challenging tasks. We trained a CTM to observe, plan, and implement routes through 2D mazes using a setup that necessitated the

formation of an internal world model. On ImageNet, the CTM exhibited native adaptive computation, naturally tailoring its processing time to input difficulty, and achieved strong calibration—a desirable property often requiring specialized techniques. On algorithmic tasks like parity checking, the CTM developed interpretable, emergent problem-solving strategies. Notably, the core architecture remained consistent across tasks, highlighting its robustness.

【解读】

同学们，这段文字主要介绍了“连续思维机器”（CTM）的核心理念以及它在实际测试中的惊人表现。让我们拆解一下其中的奥秘。

首先，文章开篇提出了一个颠覆性的观点：**时间动态（Temporal Dynamics）是核心**。目前的很多AI模型（比如你们听说的图像识别软件），它们看一张图就像“啪”地拍了一张快照，是

静态的处理。而CTM认为，真正的“思考”是一个随时间流动的过程，就像你们做数学压轴题时，大脑里的神经元是持续不断地在活动、在交流的。CTM的两大创新点在于：一是**神经元级模型**（Neuron-level models），它不像传统AI那样把神经元简化成一个冷冰冰的数学函数，而是试图模拟真实生物神经元的复杂性；二是把**同步（Synchronization）**作为一种信息的表达方式。想象一下，如果大脑里的一群神经元同时“喊话”，这种“节奏的共鸣”本身就传递了重要的信息。

接下来，作者通过几个精彩的实验来证明这个理论不仅仅是纸上谈兵。

第一个实验是**走迷宫**。这不仅仅是让AI记住路，CTM展示了它能“观察、计划并执行”。最关键的是，文中提到它形成了一个“内部世界模型”（Internal World Model）。这就好比当你走进一个陌生的校园，你闭上眼睛能在脑海里构建出教学楼和操场的相对位置。CTM也能在它的“脑”中构建这样的地图，并在行动前先在脑子里“彩排”路线。

第二个实验涉及**ImageNet**（一个巨大的图像识别数据库）。这里有一个非常像人类的特性叫“自适应计算”（Adaptive Computation）。试想一下，当你在考试中遇到一道送分题，你会秒答；而遇到一道难题，你会停下来思考很久。CTM也是如此！它能根据图片的难易程度，自动调整“思考”的时间长短。这是一般AI做不到的，通常的AI处理一张白纸和处理一张复杂的风景画用的计算量是一样的。此外，它还具有很强的“校准能力”（Calibration），意思是它知道自己有多大把握——它不会明明不懂却装懂，这在AI安全领域是非常宝贵的品质。

第三个实验是**算法任务**（如奇偶校验）。CTM展现出了“可解释性”和“涌现性”。简单说，它不是一个黑箱，我们能看懂它是怎么解决问题的，而且它像进化一样，自己“悟”出了一些解决策略。

最后，作者强调了这个架构的**鲁棒性（Robustness）**。无论是看图、走迷宫还是算数学，CTM用的都是同一套核心架构，没有这就好比一个学霸，不仅物理好，语文和体育也很好，说明他的底层学习能力（核心架构）非常强大。

【原文】

The CTM’s NLMs are inspired by the complexity of biological neurons, but are implemented with a level of abstraction appropriate for modern deep learning. The direct use of neural synchronization as a representation is, to our knowledge, a novel approach at this scale. Such a design, such as a high-cardinality representational space and the potential to capture the temporal aspects of ‘thought’. While traditional deep learning has abstracted away neural timing for computational efficiency, the CTM shows that reintroducing such dynamics in a structured way can unlock new functionalities.

Limitations.

Limitations. The CTM uses an internal sequence, meaning training times are extended. NLMs also increase parameter counts compared to standard activation functions, but also provide a new avenue for scaling. The experiments in this paper are preliminary and not intended to be state-of-the-art models tailored for performance. Therefore, a limitation of this paper is its relatively limited depth of comparison since we favored breadth to investigate the CTM’ s overall functionality.

Future Work.

Future Work. We plan to apply the CTM to language modeling, self-supervised video understanding, life-long learning, biologically-inspired memory and plasticity, multi-modal systems, and more. We believe that, conceptually, synchronization representations have high widespread potential.

References

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning. MIT press Cambridge, 2016.

[3] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

【解读】

在这一部分，我们深入探讨了CTM的设计哲学、它目前面临的局限性，以及科学家们对未来的展望。

首先，作者谈到了**生物学灵感与工程实现的平衡**。虽然CTM的灵感来自于复杂的人脑神经元，但它并不是要1:1地复制大脑（那太复杂了，目前的计算机跑不动的）。作者做了一种“抽象化”（Abstraction），即保留了生物神经元最核心的“时间”和“复杂性”特征，但又把它简化到适合现代深度学习框架运行的程度。这里提到的“神经同步作为表征”是一个非常新颖的概念。传统的深度学习为了追求计算速度（Efficiency），把“时间”这个维度给扔掉了，就像把一部电影压缩成了一张海报。虽然处理快了，但也丢失了情节的流动。CTM证明了，重新把“时间”和“动态”引入AI，虽然麻烦一点，但能解锁全新的功能，捕捉到类似人类“思维”的过程。

然而，科学研究必须是诚实的，作者非常坦率地列出了**局限性（Limitations）**。

1. **训练时间长**：因为CTM要模拟随时间变化的序列（Internal sequence），就像读完一本书比看一眼封面要花更多时间一样，训练这种模型非常耗时。
2. **参数量大**：更复杂的神经元模型（NLM）意味着更多的参数，这不仅占用内存，也增加了计算负担。
3. **初步实验**：作者强调，目前的实验更像是“概念验证”。他们现在的目标不是去打破世界纪录（State-of-the-art），而是为了测试这个新机器在各种不同任务上的通用性（Breadth over depth）。就好比发明第一架飞机时，莱特兄弟关心的是“它能不能飞”，而不是“它能不能飞得比波音747还快”。

关于**未来工作（Future Work）**，前景非常广阔。研究团队计划把CTM应用到更多领域：

- **语言建模**：现在的ChatGPT已经很强了，但如果是基于“连续思维”的语言模型，会不会有更深刻的理解力？
- **视频理解**：视频本身就是时间的艺术，CTM的时间动态特性在这里将大放异彩。
- **终身学习（Life-long learning）**：这是一个非常酷的概念。现在的AI学了新知识容易忘旧知识，而人类可以活到老学到老。CTM试图模拟这种记忆的可塑性。

最后的参考文献（References）列出了深度学习领域的几座大山，包括图灵奖得主LeCun, Bengio和Hinton的经典之作。这告诉我们，任何前沿的创新都是站在巨人的肩膀上完成的。

总的来说，这篇文档向我们展示了一种**回归生物本源**的AI设计思路。它告诉我们，也许让计算机变得更聪明的秘诀，不是单纯地堆算力，而是去模仿大自然历经亿万年进化出来的最精密的仪器——我们的大脑。【原文】

[4] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[5] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[6] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[7] Neil C Thompson, Kristian Greeneewald, Keecheon Lee, Gabriel F Manso, et al. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, 10, 2020.

[8] Patrick Hohler and Thomas Lukasiewicz. Analogy in reasoning with deep neural networks. *Journal of Artificial Intelligence Research*, 68:503–540, 2020.

[9] Peter Cariani and Janet M Baker. Time is of the essence: neural codes, synchronies, oscillations, architectures. *Frontiers in Computational Neuroscience*, 16:898829, 2022.

[10] Wolfgang Maass. On the relevance of time in neural computation and learning. *Theoretical Computer Science*, 261(1):157–178, 2001.

[11] David P Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. *arXiv preprint arXiv:1312.6115*, 2013.

[12] ChiYan Lee, Hideyuki Hasegawa, and Shangce Gao. Complex-valued neural networks: A comprehensive survey. *IEEE/CAA Journal of Automatica Sinica*, 8(8):1406–1426, 2022.

[13] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver general perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021.

【解读】
各位同学，如果你看过漫威电影，你会知道电影结束后的演职员表（Credits）虽然枯燥，但却藏着很多彩蛋和关键信息。学术论文也是如此，这部分叫“参考文献”（References），它是一篇论文的“家谱”和“致敬名单”。通过这份清单，我们可以窥探作者的学术品味、研究立场以及这篇论文试图解决的核心问题。让我们像侦探一样，把这组看似枯燥的引用文献拆解开来，看看它们背后隐藏了哪些激动人心的科学故事。

首先，我们要关注**第[4]、[5]、[6]、[7]号文献**，这几篇是人工智能领域的“反思与批判”之作。
你们可能听说过现在的AI（比如ChatGPT）非常强大，但在2017-2020年左右，学术界有一股很强的声音在反思：深度学习（Deep Learning）是不是走进了死胡同？

- **文献[4]** 由MIT的大牛Josh Tenenbaum等人撰写，他们提出要制造像人类一样思考的机器。人类学习不需要看几百万张猫的照片才能认出猫，我们有直觉、有推理能力。这篇论文呼吁AI要向人类认知科学取经。
- **文献[5]** 的作者Gary Marcus是AI界的著名“反对派”，他总是犀利地指出：现在的深度学习更像是“死记硬背”的笨学生，缺乏真正的理解。
- **文献[6]** 的作者François Chollet是著名深度学习框架Keras的作者（你们将来上大学学编程大概率会用到Keras）。他在这篇论文中探讨了“到底什么是智力”，认为真正的智能是**泛化能力**（举一反三的能力），而不是单纯靠刷题（海量数据训练）得来的高分。
- **文献[7]** 更是直接指出了物理限制：单纯靠增加算力和数据量，这种“大力出奇迹”的方法快要撞上天花板了，成本太高，我们必须寻找更巧妙的算法。
  总结来说，这组文献表明，本文的作者可能在试图解决传统深度学习“笨重、缺乏推理能力、甚至不可持续”的问题。

接下来，看**第[9]、[10]、[11]号文献**，关键词变成了“Time”（时间）、“Synchrony”（同步）、“Oscillations”（振荡）。
这就涉及到了生物学和神经科学的知识。同学们，你们的大脑神经元并不是像计算机那样按照时钟节拍一步步运算的，而是通过脉冲（Spike）和特定的节奏（脑电波）来传递信息。

- **文献[10]** Wolfgang Maass是脉冲神经网络（SNN）的泰斗级人物，他早在2001年就强调了“时间”在计算中的重要性。这暗示本文的作者可能在研究一种更接近生物大脑运作机制的“类脑计算”模型。
- **文献[11]** 提到了“Complex-valued”（复数值）。大家高三数学应该刚学过复数（$a + bi$），可能会觉得除了考试没啥用。但在信号处理中，复数是描述“波”和“相位”的神器！这说明作者可能想利用复数来模拟神经元的同步放电现象。

最后，**文献[12]** 是关于复数神经网络的综述，进一步印证了作者的技术路线——**用复数数学工具来构建更强大的神经网络**。而**文献[13]** 提到了“Perceiver”，这是Google DeepMind开发的一种通用感知模型，能同时处理声音、图像和视频。这代表了该领域最前沿的架构设计。

综上所述，通过解读这份参考文献，我们可以给这篇论文画个像： 这是一篇试图挑战传统深度学习局限性的硬核论文。作者不满足于现有的AI只会“大力出奇迹”，而是试图从**人类认知科学**（文献4-6）中寻找灵感，引入**时间维度和生物神经机制**（文献9-11），并利用**复数数学**（文献11-12）作为工具，去构建一个更智能、更像人脑的新型AI系统。这不仅是计算机科学，更是数学、生物学和哲学的交叉路口！