TSUBAME 共同利用　令和 6 年度　産業利用　成果報告書

利用課題名　マルチモーダル基盤モデルとモデル間連携技術の研究開発
英文：Research and development of inter-model adaptation technology and multimodal foundation models

シーン　誠

Sakana AI
https://sakana.ai/

邦文抄録（300 字程度）

大規模言語モデルとマルチモーダル基盤モデルは高い性能を示す一方で、そのサイズが実用展開の障壁となっている。本研究では、異なるモデル間の効率的な知識連携を実現する「Temporally Adaptive Interpolated Distillation（TAID）」を提案した。TAID は生徒モデルの初期分布から教師モデルの分布へと徐々に移行する時間依存的な中間分布を導入することで、容量ギャップ問題とモード平均化/崩壊問題を同時に解決する。実験の結果、TAID は言語モデルとマルチモーダルモデルの両方で既存手法を上回る性能を示し、その有効性が国際会議 ICLR での Spotlight 採択によっても認められた。

英文抄録（100 words 程度）

While large language models and multimodal foundation models demonstrate remarkable capabilities, their size poses significant deployment challenges. We introduce Temporally Adaptive Interpolated Distillation (TAID), a novel inter-model adaptation technique that dynamically bridges teacher and student models through time-dependent intermediate distributions. TAID gradually transitions from the student's initial distribution to the teacher's distribution, effectively addressing capacity gap and balancing mode-averaging/collapse issues. Our experiments demonstrate TAID's superior performance in developing both language and multimodal models, showcased through our state-of-the-art TAID-LLM-1.5B and TAID-VLM-2B, advancing efficient AI deployment.

Keywords: モデル間連携技術、マルチモーダル基盤モデル、知識蒸留、時間適応型学習、効率的 AI

背景と目的

大規模言語モデル（LLM）は様々な分野で革新的な能力を示しているが、そのサイズが増大するにつれて、リソース制約のある環境への展開が大きな課題となっている。例えば、エッジデバイスでの実行や、リアルタイムアプリケーションでの低遅延応答、省エネルギー運用などが困難になっている。
このような課題に対応するためには、異なるモデル間の効率的な連携技術が重要となる。知識蒸留（Knowledge Distillation）はその代表的なモデル間連携技術であり、大型の教師モデルから小型の生徒モデルへと知識を移転することで、小型でありながら高性能なモデルを作成する有望なアプローチである。しかし、従来のモデル間連携手法には二つの根本的な課題がある。一つは教師モデルと生徒モデル間の大きな容量ギャップで、もう一つはモード平均化（生徒モデルが教師の全モードを過度に平滑化）とモード崩壊（生徒モデルが特定のモードのみに集中）のバランスの問題である。
本プロジェクトでは、これらの問題を解決するための新しいモデル間連携技術「TAID: Temporally Adaptive Interpolated Distillation」の技術実験を実施し、大規模言語モデルおよびマルチモーダル基盤モデルの効率的な開発を実現することを目的とした。これにより、小型で高性能な言語モデルおよびマルチモーダルモデルの開発を可能にし、リソース制約のある環境での AI 技術の普及に貢献する。

概要

TAID の核心は、従来の知識蒸留法と比較して根本的にアプローチが異なる点にある。下図に示すように、従来の知識蒸留（左図）では生徒モデルを固定された教師分布に直接最適化するのに対し、TAID（右図）では時間に依存した中間分布を通じて段階的な最適化を行う。

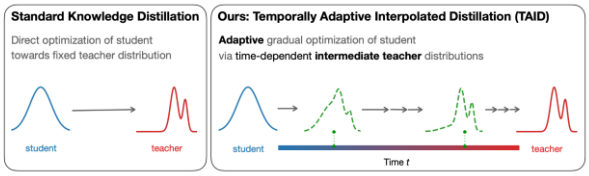


Figure 1: **Comparison of standard KD and TAID. (Left)** Standard KD methods typically employ direct optimization towards a fixed teacher distribution. **(Right)** TAID creates a dynamic bridge through adaptive, time-dependent *intermediate teacher* distributions (green dashed lines), enabling gradual optimization of the student. This approach facilitates a flexible transition from the student's initial distribution towards the teacher's distribution over time, effectively addressing the capacity gap and balancing knowledge transfer across varying model sizes.

この中間分布は、生徒モデルの初期分布（図 1 右側の左端）から教師モデルの分布（図 1 右側の右端）へと時間の経過とともに徐々に変化する。緑の破線で示される中間分布は、補間パラメータ t によって制御され、次式で定義される：

Definition 3.1 (TAID Interpolated Distribution)**.** *For any input sequence $y^{<s} \in \mathcal{Y}^{s-1}$ and any output token $y_s \in \mathcal{Y}$, the TAID interpolated distribution p_t is defined as:*

$$p_t(y_s|y^{<s}) := \text{softmax}\left((1-t)\cdot \text{logit}_{q_\theta'}(y_s|y^{<s}) + t \cdot \text{logit}_p(y_s|y^{<s})\right) \quad (1)$$

where $t \in [0,1]$ is a time-dependent interpolation parameter, $\text{logit}_{q_\theta'}$ represents a detached version of the student logits (i.e., treated as a constant without being backpropagated), and logit_p represents the teacher logits.

TAID の最適化過程では、生徒モデルは常にこの中間分布を目標として学習を進める。学習初期には中間分布が生徒モデル自身に近いため学習が容易であり、徐々に教師モデルの特性を取り入れることができる。この段階的アプローチにより、容量ギャップによる学習の困難さを軽減し、同時にモード平均化とモード崩壊のバランスも取ることができる。

さらに、TAID では補間パラメータ t を生徒モデルの学習進度に応じて適応的に更新する機構を導入している。これにより、学習過程を通じて一貫した難易度の学習タスクを維持し、効率的かつ安定した知識移転を実現している。

このような高度なモデル間連携技術により、単一モダリティの言語モデルだけでなく、複数のモダリティを扱うマルチモーダル基盤モデルの効率的な開発も可能となる。後述するように、TAID を活用してテキストのみを扱う言語モデル（TAID-LLM-1.5B）と、画像とテキストを統合的に処理するマルチモーダル基盤モデル（TAID-VLM-2B）の両方を開発することに成功した。

結果および考察

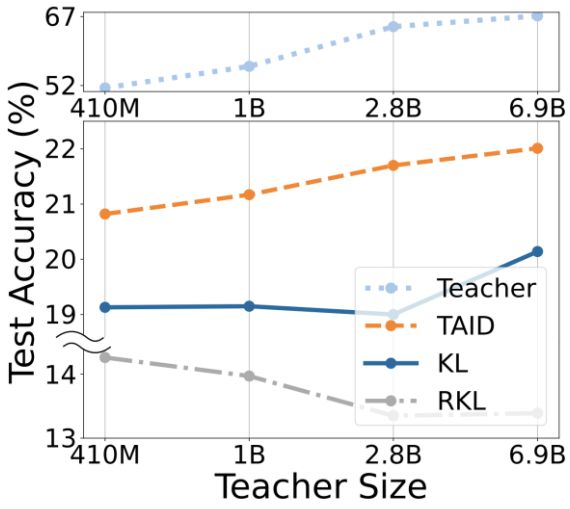
TAID の有効性を検証するため、まず UltraChat 200k データセットを用いた指示チューニング実験を行った。下表に示すように、様々な教師-生徒モデルペアでの MT-Bench スコアによる評価において、TAID は既存の知識蒸留手法を一貫して上回るパフォーマンスを示した。

Table 1: **Evaluating distillation methods for LLM instruction tuning.** The MT-Bench scores after training are listed, where higher scores indicate better conversational performance. For each of the three teacher-student pairs, different distillation algorithms, including the proposed TAID method, are compared. The highest score in each column is highlighted in **bold**.

Method \ Teacher / Student	Phi-3-mini (3.8B) TinyLlama (1.1B)	Llama-2 (6.7B) TinyLlama (1.1B)	StableLM Zephyr (2.8B) Pythia (0.4B)
SFT	2.00	3.94	2.57
KL (Hinton et al., 2015)	2.71	3.99	2.74
RKL (Wen et al., 2023; Gu et al., 2024)	3.48	3.92	2.53
TVD (Wen et al., 2023)	3.27	3.64	2.57
Adaptive KL (Wu et al., 2024)	3.27	3.77	2.64
GKD (Agarwal et al., 2024)	2.24	3.82	2.59
DistiLLM (Ko et al., 2024)	3.23	3.97	2.97
CTKD (Li et al., 2023b)	1.78	2.84	1.39
DKD (Zhao et al., 2022)	2.70	4.14	2.90
(Ours) TAID w/o adaptive update	3.44	4.18	2.88
(Ours) TAID	**4.05**	**4.27**	**3.05**

特筆すべきは、TAID が学生生成出力（SGO）を必要とせずにこの性能を達成している点である。これにより、SGO ベースの手法と比較して訓練時間が約 2〜10 倍短縮された。また、適応型更新機構を導入することで、適応型更新なしの TAID と比較して 2.2%から 17.7%の性能向上が確認された。

また、TAID の容量ギャップへの耐性を検証するため、固定サイズの生徒モデル（70M）に対して異なるサイズの教師モデル（410M から 6.9B）を用いた実験を行った。下図は、教師モデルのサイズと生徒モデルの性能の関係を示している。



TAID は教師モデルのサイズが増加するにつれて単調に性能が向上しており、容量ギャップの問題を効果的に解決していることが示されている。一方、KL や RKL といった従来手法では教師モデルのサイズ増加に伴う一貫した性能向上が見られず、容量ギャップの問題を抱えていることが確認された。

研究の実用的インパクトを示すため、TAID を用いて二つの最先端モデルを開発した。TAID-LLM-1.5B は 2B 未満のパラメータを持つモデルで最高のスコアを達成し、TAID-VLM-2B は 4B までの視覚言語モデルでトップパフォーマンスを示した。特に TAID-VLM-2B の開発成功は、本研究の主目的であるマルチモーダル基盤モデ

ルの効率的な開発が達成されたことを示している。これらの結果は、TAID というモデル間連携技術が単一モダリティだけでなくマルチモーダルモデルの開発においても有効であることを実証している。

まとめ、今後の課題

本研究では、マルチモーダル基盤モデルとモデル間連携技術の研究開発を目指し、効率的なモデル間知識移転手法「TAID」を提案した。TAID は時間依存的な中間分布を導入することで、異なるモデル間の容量ギャップを橋渡しし、モード平均化とモード崩壊のバランスを取る。広範な実験により、TAID は指示チューニングと事前学習の両シナリオにおいて既存のモデル間連携手法よりも優れたパフォーマンスを示し、その学術的重要性は ICLR 2025 での Spotlight 採択によっても認められた。
実用面では、TAID-LLM-1.5B と TAID-VLM-2B の開発により、リソース制約のある環境での高性能言語モデルおよびマルチモーダル基盤モデルの実装可能性が示された。特に視覚とテキストを統合処理する TAID-VLM-2B の成功は、本研究の主目的であるマルチモーダル基盤モデル開発の達成を示す重要な成果である。
今後の課題としては、以下の点が挙げられる：

1. 他の距離メトリクスへの TAID の拡張：現在の KL ダイバージェンス以外のメトリクスでの有効性検証
2. 非線形補間の探索：より複雑な補間スキームによる性能向上の可能性
3. マルチモーダル知識蒸留への適用：視覚-言語タスクなど複数モダリティにまたがる知識移転
4. マルチテーチャー蒸留への拡張：複数の教師モデルからの知識統合手法の開発

これらの課題に取り組むことで、TAID はより広範な AI アプリケーションにおいて効率的なモデル開発を可能にし、AI 技術のアクセシビリティと実用性をさらに向上させることが期待される。