Novelty Assessment Report

Paper: Boomerang Distillation Enables Zero-Shot Model Size Interpolation
PDF URL: https://openreview.net/pdf?id=4ZU8v4s3IR
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-07

Abstract

Large language models (LLMs) are typically deployed under diverse memory and compute constraints. Existing approaches build model families by training each size independently, which is prohibitively expensive and provides only coarse-grained size options. In this work, we identify a novel phenomenon that we call boomerang distillation: starting from a large base model (the teacher), one first distills down to a small student and then progressively reconstructs intermediate-sized models by re-incorporating blocks of teacher layers into the student without any additional training. This process produces zero-shot interpolated models of many intermediate sizes whose performance scales smoothly between the student and teacher, often matching or surpassing pretrained or distilled models of the same size. We further analyze when this type of interpolation succeeds, showing that alignment between teacher and student through pruning and distillation is essential. Boomerang distillation thus provides a simple and efficient way to generate fine-grained model families, dramatically reducing training cost while enabling flexible adaptation across deployment environments.

Core Task Landscape

This paper addresses: **Zero-Shot Model Size Interpolation Through Knowledge Distillation**

A total of **24 papers** were analyzed and organized into a taxonomy with **13 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core Knowledge Distillation Mechanisms and Training Paradigms**
- **Capacity Gap and Model Size Adaptation**
- **Domain-Specific Distillation Applications**
- **Complementary Compression and Optimization Techniques**

Complete Taxonomy Tree

- Zero-Shot Model Size Interpolation Through Knowledge Distillation Survey Taxonomy
- Core Knowledge Distillation Mechanisms and Training Paradigms
  - Interactive and Adaptive Distillation Frameworks (2 papers)
  - [2] TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models (Shing, 2025) View paper
  - [23] Interactive Knowledge Distillation (Shipeng Fu, 2020) View paper
  - Intermediate Representation and Feature-Based Distillation (4 papers)
  - [1] Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes (Fujii, 2023) View paper
  - [4] Distilling Reasoning Capabilities into Smaller Language Models (Kumar Shridhar, 2023) View paper
  - [15] EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval (Kim, 2023) View paper
  - [20] Knowledge Distillation with Feature Maps for Image Classification (Weiâ□□Chun Chen, 2018) View paper
  - Sequence-Level and Output Distribution Matching (2 papers)
  - [6] Understanding knowledge distillation in non-autoregressive machine translation (Zhou, 2019) View paper
  - [24] Sequence-Level Knowledge Distillation (Yoon Kim, 2016) View paper
  - Regularization and Robustness Enhancement in Distillation (2 papers)
  - [7] Mixkd: Towards efficient distillation of large-scale language models (Liang, 2020) View paper
  - [8] Enhancing data-free adversarial distillation with activation regularization and virtual interpolation (Qu Xiaoyang, 2021) View paper
- Capacity Gap and Model Size Adaptation
  - Zero-Shot Model Size Interpolation ★ (1 papers)
  - [0] Boomerang Distillation Enables Zero-Shot Model Size Interpolation (Anon et al., 2026) View paper
  - Capacity Gap Mitigation Strategies (2 papers)
  - [11] Lifting the Curse of Capacity Gap in Distilling Language Models (Song Dawei, 2023) View paper
  - [12] Bridging the Capacity Gap for Online Knowledge Distillation (MaoRong Wang, 2023) View paper
  - Dynamic Capacity and Progressive Compression (1 papers)
  - [18] Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval (Yi Xie, 2023) View paper
- Domain-Specific Distillation Applications
  - Audio and Multimodal Distillation (2 papers)
  - [5] Distilling a speech and music encoder with task arithmetic (Lin Yi-cheng, 2025) View paper
  - [17] Application of Knowledge Distillation to Multi-task Speech Representation Learning (Mine Kerpicci, 2022) View paper
  - Vision and Video Processing Distillation (3 papers)

- [9] A Motion Distillation Framework for Video Frame Interpolation (Shili Zhou, 2023) View paper
- [10] ST-MFNET Mini: Knowledge Distillation-Driven Frame Interpolation (Crispian Morris, 2023) View paper
- [16] Plant Disease Detection Using Dynamic Knowledge Distillation and Attention Mechanism (Mohammad Ghasemi Arian, 2025) View paper
- Specialized Domain Adaptation (3 papers)
- [13] Interpolative Distillation for Unifying Biased and Debiased Recommendation (Sihao Ding, 2022) View paper
- [19] Adaptation of Embedding Models to Financial Filings via LLM Distillation (Eliot Brenner, 2025) View paper
- [22] Distilling mathematical reasoning capabilities into Small Language Models. (Xunyu Zhu, 2024) View paper

- Complementary Compression and Optimization Techniques
  - Pruning and Sparsification with Distillation (1 papers)
  - [3] Enhanced Sparsification via Stimulative Training (Tang Shengji, 2024) View paper
  - Data Distillation and Synthetic Dataset Generation (1 papers)
  - [21] Better Data Distillation by Condensing the Interpolated Graphs (Yang Sun, 2023) View paper
  - Reinforcement Learning Enhanced Distillation (1 papers)
  - [14] Distribution Matching Distillation Meets Reinforcement Learning (Dengyang Jiang, 2025) View paper

Narrative

Core task: Zero-shot model size interpolation through knowledge distillation. The field of knowledge distillation has evolved into a rich landscape organized around several complementary themes. At the highest level, researchers explore core distillation mechanisms and training paradigms that define how knowledge transfers from teacher to student models, including foundational techniques like Sequence-Level Distillation[24] and interactive approaches such as Interactive Knowledge Distillation[23]. A second major branch addresses the capacity gap and model size adaptation, investigating how to bridge performance differences when student models vary significantly in scale—exemplified by works like Bridging Capacity Gap[12] and Interpolative Distillation[13]. Domain-specific applications form another branch, tailoring distillation to specialized tasks ranging from speech and music encoding (e.g., Speech Music Encoder[5]) to plant disease detection and financial document processing. Finally, complementary compression and optimization techniques integrate distillation with pruning, quantization, and other efficiency methods, as seen in Enhanced Sparsification[3] and related efforts.

Within the capacity gap and model size adaptation branch, a particularly active line of work focuses on enabling flexible deployment across diverse hardware constraints without retraining multiple models. Boomerang Distillation[0] sits squarely in this area, proposing zero-shot interpolation to generate student models of arbitrary sizes on the fly. This contrasts with earlier methods like Interpolative Distillation[13] and Capacity Dynamic Distillation[18], which typically require explicit training for each target size or rely on predefined capacity schedules. Meanwhile, approaches such as Bridging Capacity Gap[12] and Lifting Capacity Gap[11] emphasize architectural or training adjustments to narrow performance drops when scaling down, but do not directly address the zero-shot interpolation scenario. The central tension across these works revolves around balancing deployment flexibility, training efficiency, and the preservation of teacher-level performance, with Boomerang Distillation[0] offering a novel pathway by decoupling model size selection from the distillation training phase.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf focuses on achieving continuous model size variation post-training through interpolation techniques, particularly layer reincorporation, without requiring retraining. The sibling subtopics address related but distinct challenges: one targets bridging large capacity gaps between teacher-student pairs, while the other explores dynamic or progressive capacity adjustment during the distillation process itself. All three share the common theme of managing model capacity in knowledge distillation contexts.

Similarities: - All three categories operate within the knowledge distillation framework and address model capacity considerations - Each approach aims to improve flexibility or efficiency in creating models of varying sizes - All methods seek to avoid or minimize the computational cost of full retraining for different model sizes

Differences: - Zero-Shot Model Size Interpolation enables post-hoc continuous size variation without retraining, while siblings focus on training-time strategies - Capacity Gap Mitigation targets the specific problem of large teacher-student disparities, whereas the original leaf addresses arbitrary size selection - Dynamic Capacity approaches modify capacity during training through progressive or editable mechanisms, while the original leaf achieves variation after training completion - The original leaf emphasizes layer reincorporation for interpolation, while siblings use capacity bridging techniques or progressive compression strategies

Suggested Search Directions: - Investigate whether zero-shot interpolation techniques could be combined with capacity gap mitigation to handle extreme size variations - Explore if progressive compression methods could generate intermediate checkpoints suitable for post-training interpolation - Examine whether layer reincorporation strategies could inform dynamic capacity adjustment during training

Sibling Subtopics

- **Capacity Gap Mitigation Strategies** (leaves: 1, papers: 2)
- Scope: Methods explicitly designed to bridge large capacity differences between teacher and student models.
- Exclude: Interpolation methods and dynamic capacity approaches belong in sibling categories.
- **Dynamic Capacity and Progressive Compression** (leaves: 1, papers: 1)
- Scope: Approaches using editable model capacity or progressive reduction strategies during distillation.
- Exclude: Static capacity methods and fixed-size distillation belong in sibling categories.

Contributions Analysis

Overall novelty summary. The paper introduces boomerang distillation, a method for zero-shot model size interpolation that progressively reconstructs intermediate-sized models by re-incorporating teacher layers into a distilled student without retraining. Within the taxonomy, it occupies the 'Zero-Shot Model Size Interpolation' leaf under 'Capacity Gap and Model Size Adaptation'. Notably, this leaf contains only the original paper itself, indicating a sparse research direction. The broader parent category includes three leaves addressing capacity gaps, dynamic compression, and interpolation strategies, suggesting the paper targets an underexplored niche within a moderately active research area.

The taxonomy reveals that neighboring work primarily focuses on capacity gap mitigation through architectural adjustments or training modifications (e.g., Bridging Capacity Gap, Lifting Capacity Gap) and dynamic capacity approaches requiring progressive training schedules (e.g., Capacity Dynamic Distillation). The original paper diverges by decoupling size selection from training, enabling post-hoc interpolation. Related branches address core distillation mechanisms (feature-based, sequence-level matching) and domain-specific applications, but these do not directly tackle zero-shot size variation. The taxonomy's scope and exclude notes clarify that methods requiring retraining for each size belong in sibling categories, positioning this work as methodologically distinct.

Among twenty-one candidates examined via semantic search and citation expansion, none clearly refute the three core contributions. The boomerang distillation phenomenon itself was examined against ten candidates with zero refutable overlaps. The claim that interpolated models match or surpass standard distilled models drew six candidates, again with no refutations. The analysis of enabling conditions examined five candidates without finding prior work that anticipates this specific alignment-based interpolation mechanism. This limited search scope suggests the contributions appear novel within the examined literature, though the small candidate pool and sparse taxonomy leaf indicate the field context remains relatively unexplored.

Given the restricted search scale and the paper's position in a singleton taxonomy leaf, the work appears to introduce a genuinely new interpolation paradigm within the examined scope. However, the analysis covers top-K semantic matches and immediate citations, not an exhaustive survey of all distillation or model compression literature. The absence of sibling papers in the same leaf and the limited candidate pool mean the novelty assessment reflects current indexed work rather than a comprehensive field review.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Boomerang distillation phenomenon for zero-shot model size interpolation

Description: The authors identify and introduce boomerang distillation, a novel phenomenon where a small distilled student model can be progressively patched with teacher layers to create intermediate-sized models without additional training. This process produces models whose size and performance smoothly interpolate between the student and teacher.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Categories of response-based, feature-based, and relation-based knowledge distillation

URL: View paper

Brief Assessment

Categories Knowledge Distillation[29] is a survey paper categorizing existing knowledge distillation methods (response-based, feature-based, relation-based) but does not describe creating intermediate-sized models by patching student with teacher layers without additional training.

---

2. Efficient Technical Term Translation: A Knowledge Distillation Approach for Parenthetical Terminology Translation

URL: View paper

Brief Assessment

Technical Term Translation[32] focuses on knowledge distillation for terminology translation tasks, not on creating intermediate-sized models through layer patching. The candidate addresses a completely different problem domain (translation) with different objectives than the original paper's model size interpolation.

---

3. DANet: Multi-scale UAV target detection with dynamic feature perception and scale-aware knowledge distillation

URL: View paper

Brief Assessment

DANet[34] focuses on knowledge distillation for multi-scale UAV target detection in infrared imagery, not on creating intermediate-sized models through progressive layer patching. The distillation approach is task-specific (UAV detection) rather than a general model compression technique for size interpolation.

---

4. VLsI: Verbalized Layers-to-Interactions from Large to Small Vision Language Models

URL: View paper

Brief Assessment

VLsI[28] focuses on layer-wise distillation for vision-language models using verbalizers to align reasoning processes, not on creating intermediate-sized models by patching student with teacher layers without additional training.

---

5. Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert

URL: View paper

Brief Assessment

Distilhubert[26] focuses on distilling speech representations from HuBERT models using multi-task learning, not on creating intermediate-sized models by patching student with teacher layers for zero-shot interpolation.

---

6. On-Device Large Language Models: A Survey of Model Compression and System Optimization

URL: View paper

Brief Assessment

On-Device LLMs[27] is a survey paper that reviews existing compression techniques including quantization, pruning, and knowledge distillation. It does not present a novel method for creating intermediate-sized models through layer patching or demonstrate zero-shot model size interpolation as described in the original contribution.

---

7. Student network learning via evolutionary knowledge distillation

URL: View paper

Brief Assessment

Evolutionary Knowledge Distillation[31] focuses on training teacher and student networks simultaneously with guided modules for knowledge transfer, not on creating intermediate-sized models by patching student with teacher layers without additional training.

---

8. LAD: Layer-Wise Adaptive Distillation for BERT Model Compression

URL: View paper

Brief Assessment

LAD[30] focuses on task-specific distillation with adaptive layer-wise knowledge transfer for BERT compression, not on creating intermediate-sized models by patching student with teacher layers without additional training.

---

9. Robustdistiller: Compressing Universal Speech Representations for Enhanced Environment Robustness

URL: View paper

Brief Assessment

Robustdistiller[33] focuses on compressing speech representations through knowledge distillation with data augmentation and multi-task denoising, not on creating intermediate-sized models by patching student layers with teacher layers for zero-shot interpolation.

---

10. TransKD: Transformer knowledge distillation for efficient semantic segmentation

URL: View paper

Brief Assessment

TransKD[25] focuses on knowledge distillation for semantic segmentation using transformer-specific patch embeddings and feature maps. It does not address zero-shot model size interpolation by patching student models with teacher layers, which is the core novelty of boomerang distillation.

Contribution 2: Demonstration that interpolated models match or surpass standard distilled models

Description: The authors demonstrate through experiments that models created via boomerang distillation achieve comparable or superior performance to models trained with standard knowledge distillation at the same size, despite requiring no additional training for the intermediate sizes.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Bi-Temporal Feature Relational Distillation for On-Board Lightweight Change Detection in Remote Sensing Imagery

URL: View paper

Brief Assessment

Bi-Temporal Feature Distillation[35] focuses on knowledge distillation for remote sensing change detection tasks, not on model size interpolation or comparing interpolated models to standard distilled models in language model families.

---

2. Mixkd: Towards efficient distillation of large-scale language models

URL: View paper

Brief Assessment

Mixkd[7] focuses on knowledge distillation with mixup data augmentation during training, not on zero-shot model size interpolation. The candidate requires training student models, whereas the original contribution creates interpolated models without additional training by patching teacher layers.

---

3. 4D trajectory lightweight prediction algorithm based on knowledge distillation technique

URL: View paper

Brief Assessment

4D Trajectory Prediction[37] applies knowledge distillation to trajectory prediction in air-traffic management, not to creating interpolated model families of different sizes. The candidate focuses on teacher-student distillation for a single lightweight model, whereas the original work demonstrates zero-shot model size interpolation.

---

4. TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models

URL: View paper

Brief Assessment

TAID[2] focuses on temporally adaptive interpolation during knowledge distillation training, creating a dynamic intermediate teacher distribution. The original paper's contribution concerns zero-shot model size interpolation after training by patching student models with teacher layers, which is a fundamentally different approach to achieving interpolated models.

---

5. Diversity-rewarded CFG distillation

URL: View paper

Brief Assessment

Diversity-rewarded CFG[36] focuses on distilling classifier-free guidance for music generation and uses model merging to trade off quality-diversity. This is technically distinct from boomerang distillation's zero-shot layer interpolation approach for creating model families of different sizes.

---

6. Distilling a speech and music encoder with task arithmetic

URL: View paper

Brief Assessment

Speech Music Encoder[5] focuses on task arithmetic for merging speech and music SSL models through distillation, not on general model size interpolation via boomerang distillation. The domains and technical approaches differ fundamentally.

Contribution 3: Analysis of conditions enabling boomerang distillation

Description: The authors conduct extensive experiments and ablations to characterize when boomerang distillation succeeds, showing that student initialization from teacher weights and training with alignment losses (such as cosine distance) are essential conditions, and that the approach consistently outperforms layer pruning methods.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Epsd: Early pruning with self-distillation for efficient model compression

URL: View paper

Brief Assessment

Epsd[39] focuses on early pruning combined with self-distillation for model compression, not on analyzing conditions for boomerang distillation. The candidate addresses different technical objectives: identifying distillable weights during pruning rather than characterizing when student-teacher interpolation succeeds.

---

2. Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification

URL: View paper

Brief Assessment

Face Classification Distillation[42] focuses on knowledge distillation for face recognition tasks (classification, alignment, verification) using different network architectures and task-specific targets. It does not address the conditions for model size interpolation or layer-based student initialization that are central to boomerang distillation.

---

3. End-to-end model compression via pruning and knowledge distillation for lightweight image super resolution

URL: View paper

Brief Assessment

Lightweight Image Super-Resolution[38] focuses on image super-resolution model compression through pruning and knowledge distillation, not on language model distillation conditions or student initialization strategies for zero-shot model size interpolation.

---

4. From Variance-Reduced Initialization to Knowledge Distillation-Inspired Pruning at Initialization: Embedding Efficiency Right from the Onset of Neural Network â¦

URL: View paper

Brief Assessment

Variance-Reduced Initialization[40] focuses on initialization techniques and pruning-at-initialization methods, not on analyzing conditions for model distillation with student initialization and alignment losses versus pruning methods.

---

5. Exploring Pruning-based Efficient Object Tracking via Hybrid Knowledge Distillation

URL: View paper

Brief Assessment

Pruning-based Object Tracking[41] focuses on efficient object tracking via pruning and knowledge distillation for visual tracking models, not on language model distillation with student initialization and alignment losses for model size interpolation.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Boomerang Distillation Enables Zero-Shot Model Size Interpolation View paper
- [1] Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes View paper
- [2] TAID: Temporally Adaptive Interpolated Distillation for Efficient Knowledge Transfer in Language Models View paper
- [3] Enhanced Sparsification via Stimulative Training View paper
- [4] Distilling Reasoning Capabilities into Smaller Language Models View paper
- [5] Distilling a speech and music encoder with task arithmetic View paper
- [6] Understanding knowledge distillation in non-autoregressive machine translation View paper
- [7] Mixkd: Towards efficient distillation of large-scale language models View paper
- [8] Enhancing data-free adversarial distillation with activation regularization and virtual interpolation View paper
- [9] A Motion Distillation Framework for Video Frame Interpolation View paper
- [10] ST-MFNET Mini: Knowledge Distillation-Driven Frame Interpolation View paper
- [11] Lifting the Curse of Capacity Gap in Distilling Language Models View paper
- [12] Bridging the Capacity Gap for Online Knowledge Distillation View paper
- [13] Interpolative Distillation for Unifying Biased and Debiased Recommendation View paper
- [14] Distribution Matching Distillation Meets Reinforcement Learning View paper
- [15] EmbedDistill: A Geometric Knowledge Distillation for Information Retrieval View paper
- [16] Plant Disease Detection Using Dynamic Knowledge Distillation and Attention Mechanism View paper
- [17] Application of Knowledge Distillation to Multi-task Speech Representation Learning View paper
- [18] Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval View paper
- [19] Adaptation of Embedding Models to Financial Filings via LLM Distillation View paper
- [20] Knowledge Distillation with Feature Maps for Image Classification View paper
- [21] Better Data Distillation by Condensing the Interpolated Graphs View paper
- [22] Distilling mathematical reasoning capabilities into Small Language Models. View paper
- [23] Interactive Knowledge Distillation View paper
- [24] Sequence-Level Knowledge Distillation View paper
- [25] TransKD: Transformer knowledge distillation for efficient semantic segmentation View paper
- [26] Distilhubert: Speech Representation Learning by Layer-Wise Distillation of Hidden-Unit Bert View paper
- [27] On-Device Large Language Models: A Survey of Model Compression and System Optimization View paper
- [28] VLsI: Verbalized Layers-to-Interactions from Large to Small Vision Language Models View paper
- [29] Categories of response-based, feature-based, and relation-based knowledge distillation View paper
- [30] LAD: Layer-Wise Adaptive Distillation for BERT Model Compression View paper
- [31] Student network learning via evolutionary knowledge distillation View paper
- [32] Efficient Technical Term Translation: A Knowledge Distillation Approach for Parenthetical Terminology Translation View paper
- [33] Robustdistiller: Compressing Universal Speech Representations for Enhanced Environment Robustness View paper
- [34] DANet: Multi-scale UAV target detection with dynamic feature perception and scale-aware knowledge distillation View paper
- [35] Bi-Temporal Feature Relational Distillation for On-Board Lightweight Change Detection in Remote Sensing Imagery View paper
- [36] Diversity-rewarded CFG distillation View paper
- [37] 4D trajectory lightweight prediction algorithm based on knowledge distillation technique View paper
- [38] End-to-end model compression via pruning and knowledge distillation for lightweight image super resolution View paper
- [39] Epsd: Early pruning with self-distillation for efficient model compression View paper
- [40] From Variance-Reduced Initialization to Knowledge Distillation-Inspired Pruning at Initialization: Embedding Efficiency Right from the Onset of Neural Network â¦ View paper
- [41] Exploring Pruning-based Efficient Object Tracking via Hybrid Knowledge Distillation View paper
- [42] Model Distillation with Knowledge Transfer from Face Classification to Alignment and Verification View paper