DONUT-hole: DONUT Sparsification by Harnessing Knowledge and Optimizing Learning Efficiency

Azhar Shaikh, Michael Cochez, Denis Diachkov, Michiel de Rijcke, Sahar Yousefi

{a.shaikh@student.,m.cochez@}vu.nl, {d.diachkov,m.d.rijcke,s.yousefi}@primevision.com

NEURAL INFORMATION PROCESSING SYSTEMS

primevision smart vision, smart flow

VU UNIVERSITY AMSTERDAM

Abstract

The DONUT-hole model enhances the foundational DONUT's OCR and VSU capabilities within a transformer framework, significantly improving deployability with a 54% reduction in model density via knowledge distillation and pruning. It retains robust performance and closely mirrors DONUT's internal representations, indicated by a CKA score of 0.79, affirming its proficiency in extracting key document information, crucial for logistics operations.

Proposed DONUT Model Configurations

DONUT-base-11M: is employed as the **teacher network** in the distillation and pruning experiments and has the same pretrained weights and architecture as the original DONUT model.

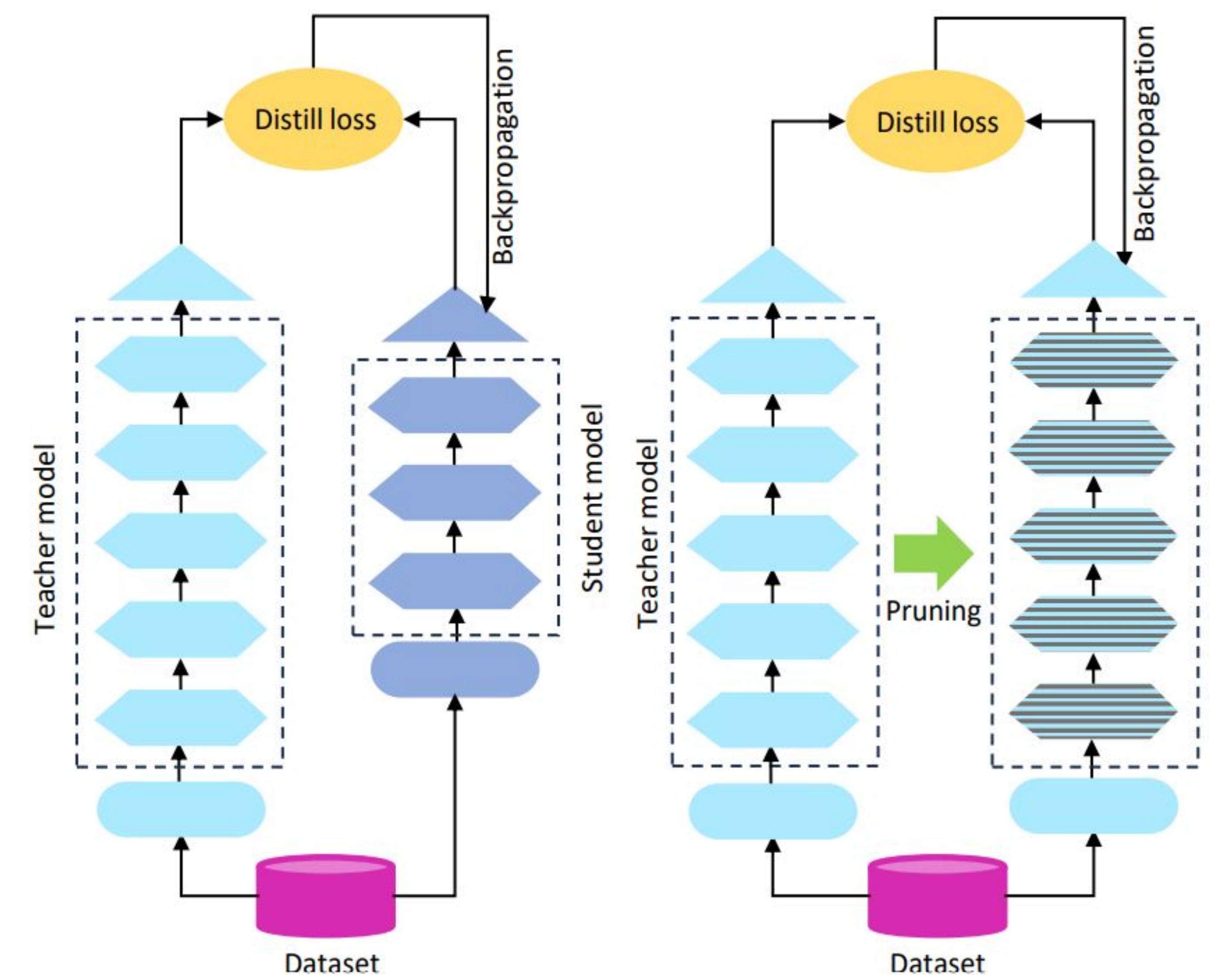
DONUT-base-0.5M: is a scaled down version of the original DONUT model pre trained on 500k SynthDog-En images. Due to the unavailability of ground truth data DONUT was originally trained on.

DONUT-small: a lighter version of DONUT made by replacing Swin-B encoder with Swin-T encoder and replacing 4 layer MBART decoder with 2 layer BART encoder. Additionally an adaptor layer (a small neural network) is used to align text and visual token dimensions for effective cross-modal fusion.

DONUT-small-distilled: Knowledge distillation is employed with DONUT-small as the student and DONUT-base as the teacher to produce this model.

DONUT-base-pruned: Magnitude pruning is employed on the teacher DONUT-base-11M model. Model is pruned to ~50% spasticity to produce this model.

DONUT-hole: The prune-then-distill paradigm produces this model by distilling knowledge from the DONUT-base model to the DONUT-base-pruned model.



Left: Knowledge Distillation

Right: Prune-then-Distill

Metrics

Tree Edit Distance (TED) Accuracy: Measures the similarity between the predicted and the actual structure of the document trees. Higher is better.

Field F1 Accuracy: Assesses precision and recall in field-level predictions, crucial for information extraction accuracy.

Centered Kernel Alignment(CKA) is a method used to compare representations based on comparing representational similarity matrices.

$$X \in \mathbb{R}^{m\times d1} \quad Y \in \mathbb{R}^{m\times d2} \quad \mathrm{HSIC}_0(K,L) = \frac{\mathrm{vec}(K_0)\cdot \mathrm{vec}(L_0)}{(m-1)^2}$$

$$K = XX^T \quad L = YY^T$$

$$K_0 = HKH \quad L_0 = HLH \quad \mathrm{CKA}(K,L) = \frac{\mathrm{HSIC}_0(K,L)}{\sqrt{\mathrm{HSIC}_0(K,K)\cdot \mathrm{HSIC}_0(L,L)}}$$

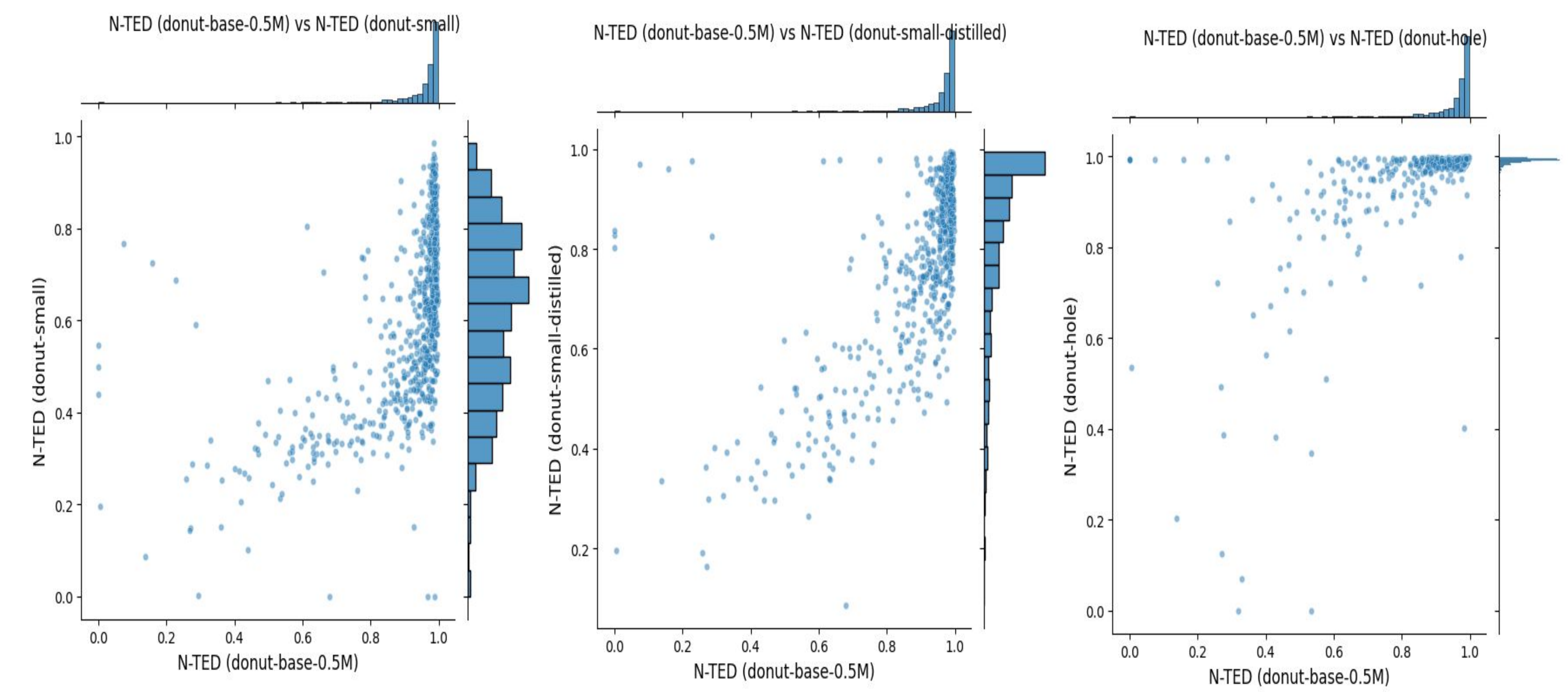
Results

Model	#Non-embedding Params	TED Accuracy	F1 Accuracy
donut-base-0.5M	140M	0.76	0.61
donut-small	37M	0.55	0.37
donut-small(with distillation)	37M	0.61	0.41
donut-base-pruned	37M	0.0	0.0
donut-base-pruned(with distillation)	37M	0.85	0.75

Results of the downstream KIE task on the cord-v2 dataset

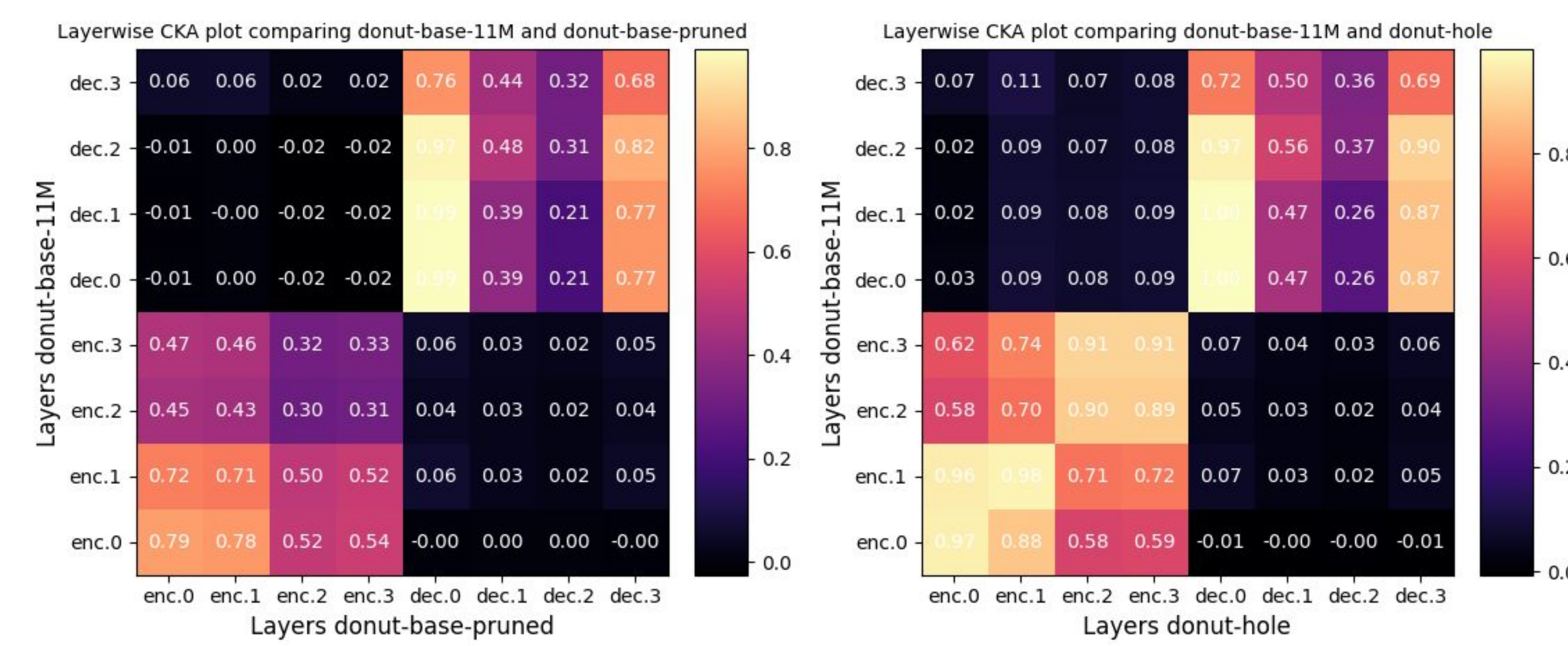
Model	#Non-embedding Params	TED Accuracy	F1 Accuracy
donut-base-0.5M	140M	0.65	0.50
donut-small	37M	0.44	0.26
donut-small(with distillation)	37M	0.50	0.35
donut-base-pruned	37M	0.24	0.048
donut-base-pruned(with distillation)	37M	0.73	0.57

Results of the downstream KIE task on the parcel reader dataset



(a) DONUT-small (b)DONUT-small-distilled (c) DONUT-hole

Scatter plot of N-TED values of the proposed DONUT model configurations vs DONUT-base-0.5M on the SynthDog-EN test set on the upstream reading task



(a) DONUT-pruned (b) DONUT-hole

Visualizing Layerwise CKA Representational Similarity Index Heatmaps comparing representations of the trained models and DONUT-base-11M

Conclusions

- Prune-then-distill is a simple yet effective paradigm reducing the DONUT model's size by 54% while retaining its performance efficacy.
- Distillation shows promising results in boosting model performance and bringing model representation closer to the teacher.