Attention Distillation for Detection Transformers: Application to Real-Time Video Object Detection in Ultrasound

Jonathan Rubin[1], Ramon Erkamp[1], Ragha Srinivasa Naidu[1], Anumod Odungatta Thodiyil[2], Alvin Chen[1]

[1]Philips Research North America, Cambridge MA, United States

[2]Philips Innovation Campus, Bangalore, India

ML4H 2021

Introduction

- We introduce a method for efficient knowledge distillation of transformer-based object detectors.
- ***Attention distillation*** makes use of the self-attention matrices generated in the layers of detection transformer (DETR) models.
- Localization information from the attention maps of a large teacher network are distilled into smaller student networks capable of running at much higher speeds.
- We apply the approach to the clinically important problem of detecting medical instruments (e.g. needle insertion procedures) in real-time from ultrasound video sequences, where inference speed is critical on computationally resource-limited hardware.

Data

- Ultrasound video sequences acquired from ~12,200 needle insertions (~2 million individual frames) were used for model training and evaluation.
- Data were collected over two years from *ex vivo* tissues (porcine, bovine, and chicken) as well as human cadavers, and comprised a range of ultrasound transducers, systems, ultrasound imaging settings (gain, depth, and tissue presets), needle types, needle sizes, insertion angles, and bevel orientations.
- A total of 30,770 labeled video clips were used as the training set, and 5,023 labeled clips from independent data collection experiments were used for evaluation.

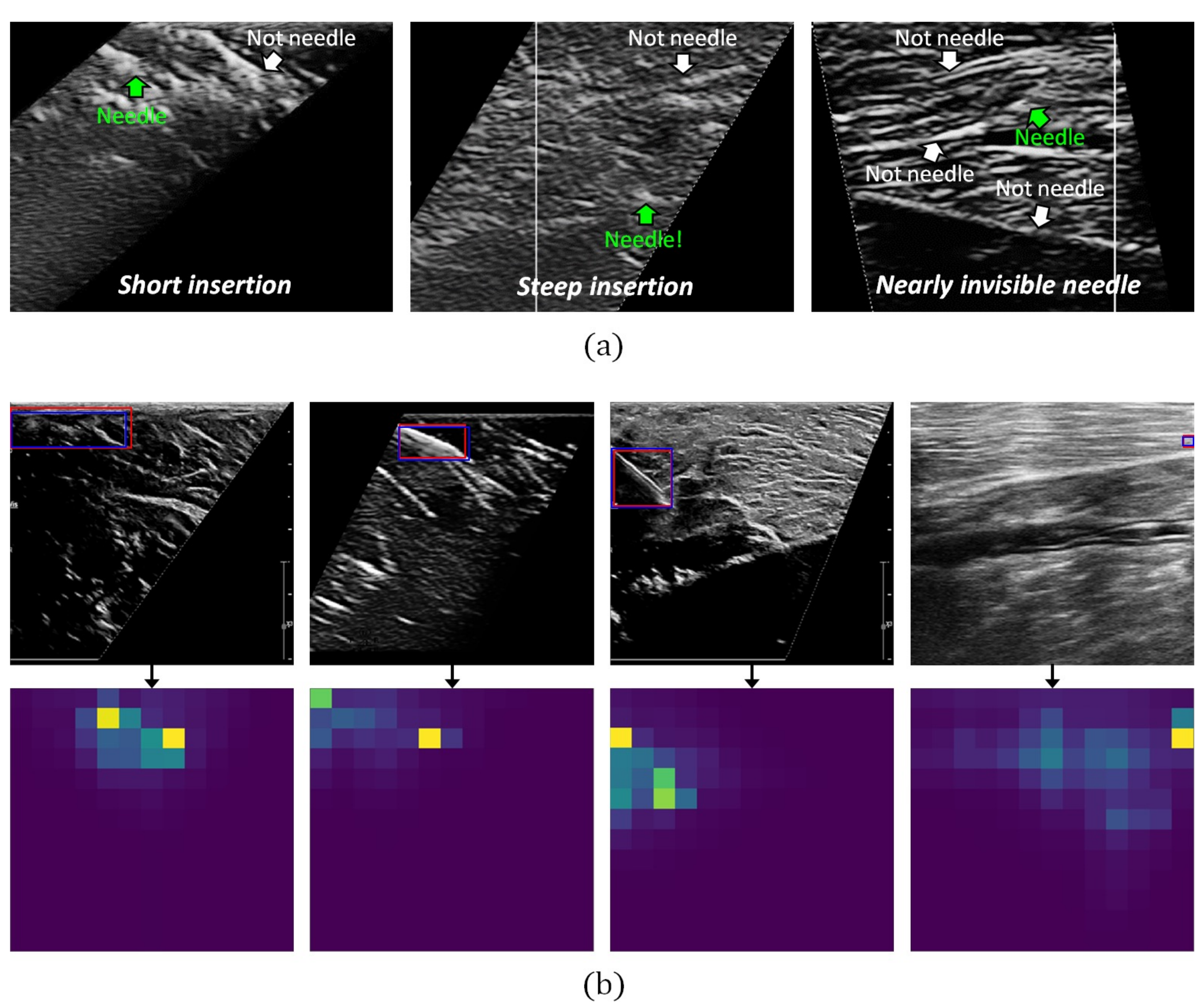

Fig. 1. (a) Representative ultrasound frames acquired during live needle insertion procedures on *ex vivo* and human cadaver tissues. The examples highlight the difficulty of correctly identifying the needle in the presence of noise and surrounding tissues. (b) Representative self-attention maps from the detection transformer. Red boxes are ground-truth and blue boxes are predictions.

Attention Distillation

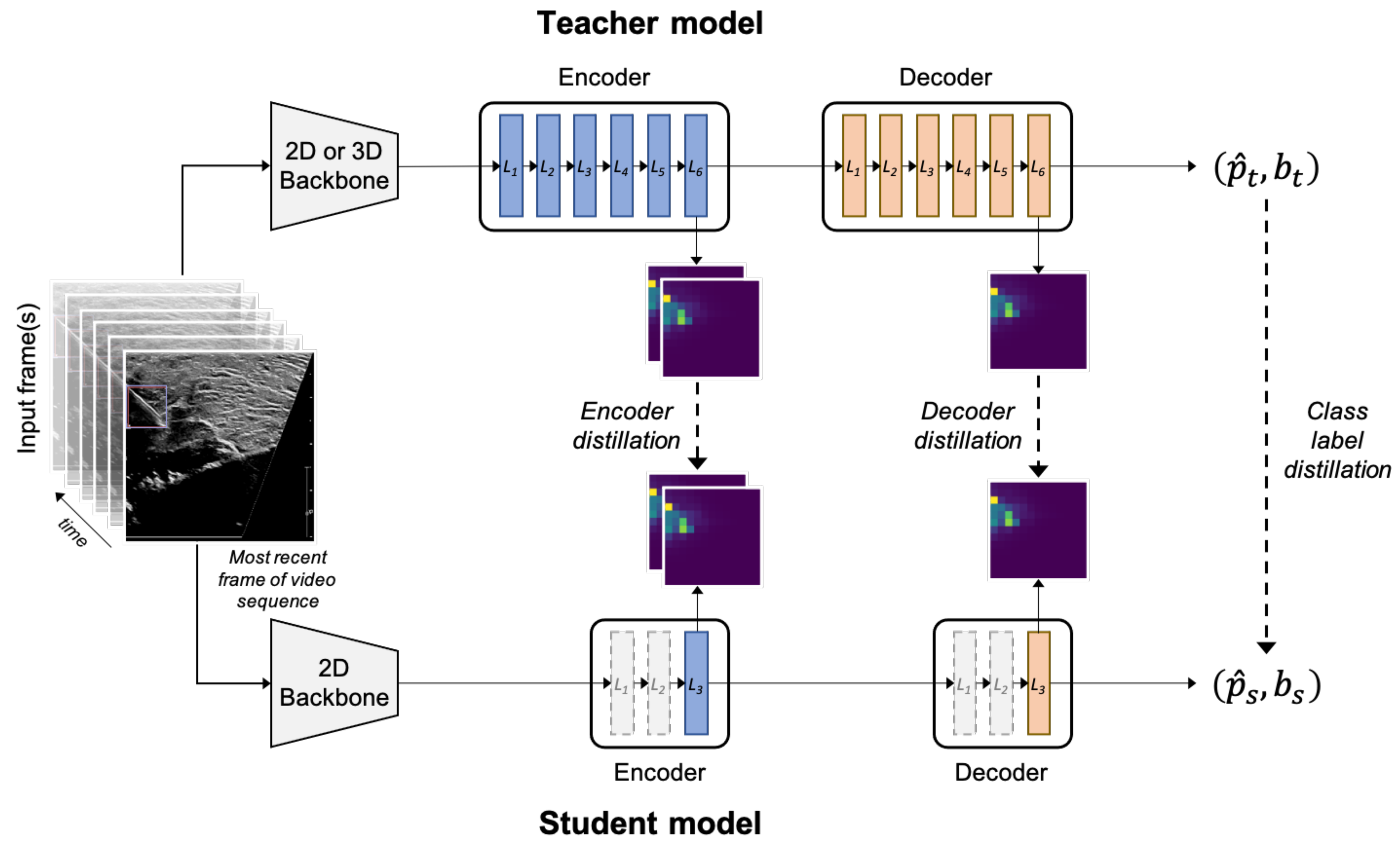

Fig. 2. Overview of attention distillation for detection transformers. Self-attention matrices of large 2D or 3D teacher networks are used to distill localization information to smaller student networks consisting of fewer encoder and decoder layers.

- We apply attention distillation by making use of self-attention matrices generated within the encoder-decoder detection transformer architecture.
- Multi-headed scaled dot-product attention Carion et al. (2020) is applied to learned query, Q, and key, K, matrices.

$$\mathrm{A} = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

- A is the attention matrix and d_k is the size of the multi-headed attention hidden dimension chosen as a hyper-parameter.

$$\mathcal{L}_{distill} = (1-\alpha)\cdot\mathcal{L}_{\mathrm{box}}\left(b_i, \hat{b}_i\right) + \alpha\cdot\left(\mathcal{KL}\left(A_i^s \| A_i^t\right) + T^2\cdot\mathcal{KL}\left(\sigma(\frac{\hat{p}_i^s}{T})\|\sigma(\frac{\hat{p}_i^t}{T})\right)\right)$$

- In the equation above, α is a hyper-parameter that controls mixing between the bounding box loss and the attention distillation loss, where b_i and $\hat{b}_i$ refer to the ground truth and predicted bounding box coordinates, and $\hat{p}_i^s$, $\hat{p}_i^t$ are class prediction probabilities given by the student and teacher networks, respectively.
- The first component of the loss applies knowledge distillation to the self-attention maps created by teacher and student detection transformers.
- The second component of the loss applies knowledge distillation to the class label predictions. T is a temperature hyper-parameter that controls smoothing, as in Hinton et al. (2015), and σ is the softmax operation.

Results

2D-to-2D Attention Distillation for Images

- Our teacher network (DETR-R50-6/6) is a detection transformer with ResNet-50 backbone and six encoder and decoder layers. We trained smaller student networks (DETR-R50-1/1) comprising an identical backbone but consisting of only a single encoder and decoder.

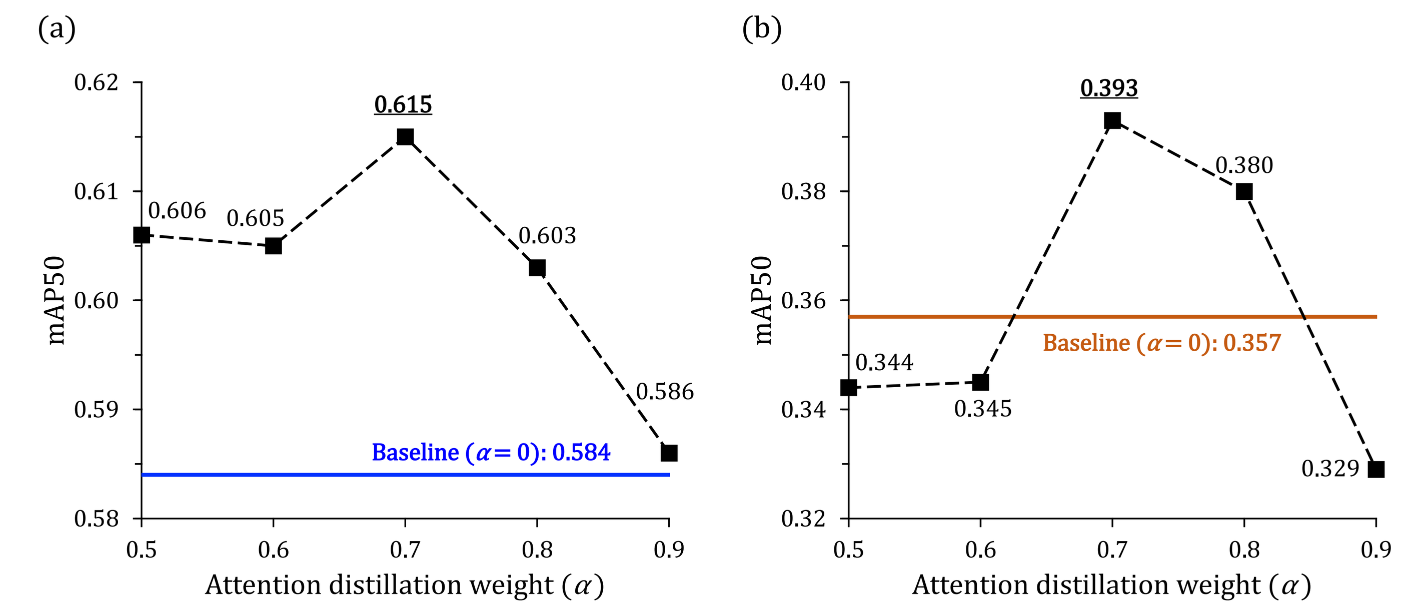

Fig. 3. Results for a sweep over α values from 0.5 – 0.9 using a DETR-R50-1/1 student network with attention distillation applied at the final encoder layer. Baseline at α = 0 (no distillation) also shown.

(a) mAP on test dataset (5,023 video clips).

(b) mAP_{short} on a challenging dataset of short needle insertions.

3D-to-2D Attention Distillation for Videos

- We apply attention distillation to compress a 3D detection transformer, which infers on a video sequence, into a 2D student model that processes single frames independently.

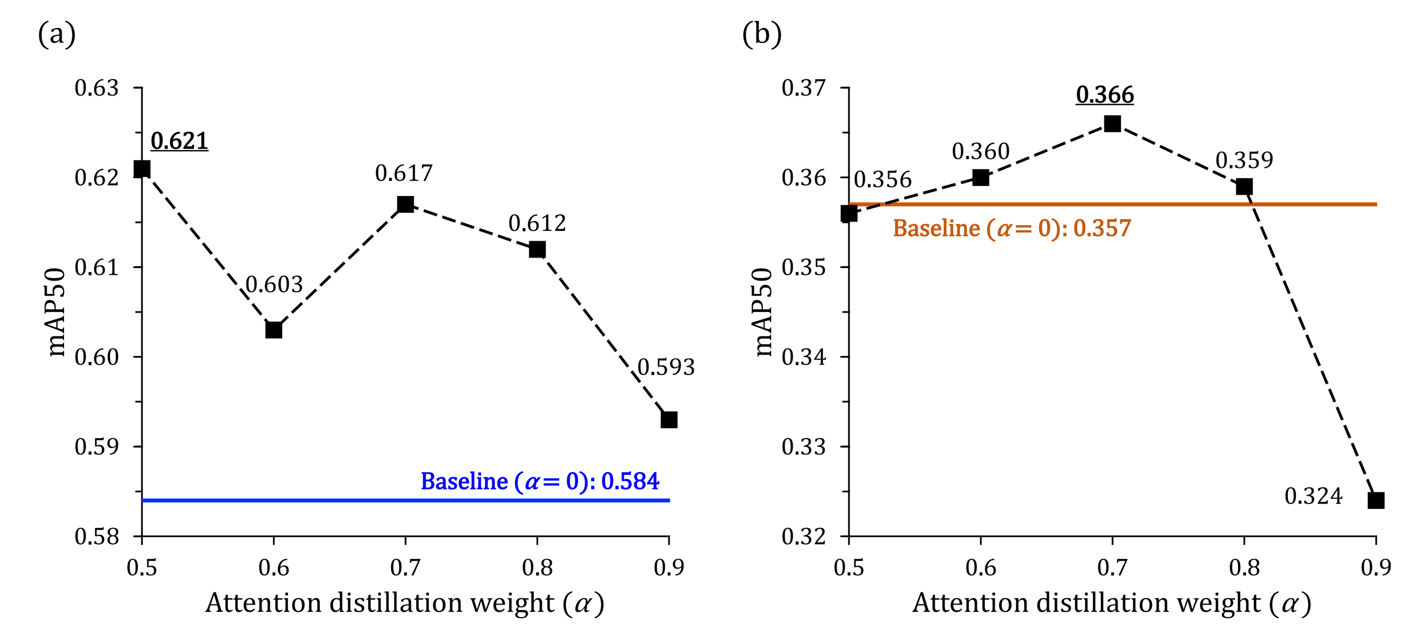

Fig. 4. 3D attention distillation results for a sweep over α values from 0.5 – 0.9, again with attention distillation applied at the final encoder layer. Baseline at α = 0 (no distillation) also shown.

(a) mAP on test dataset (5,023 video clips)..

(b) mAP_{short} on a challenging dataset of short needle insertions.

Model	Parameters	GMac	FPS	mAP^{50}	$\mathrm{mAP}^{50}_{short}$
Faster R-CNN	41,299,161	134.1	19	0.773	0.681
DETR-R50-1/1 (baseline, α=0)	27,007,174	14.4	53	0.584	0.357
DETR-R50-1/1 (student)	27,007,174	14.4	53	0.615	0.393
DETR-R50-2/2 (student)	29,900,998	14.6	43	**0.643**	0.437
DETR-R50-3/3 (student)	32,794,822	14.8	38	0.633	**0.445**
DETR-R50-6/6 (teacher)	41,476,294	15.3	26	0.655	0.467

Model	Parameters	GMac	FPS	mAP^{50}	$\mathrm{mAP}^{50}_{short}$
DETR-R50-1/1 (baseline, α=0)	27,007,174	14.4	53	0.584	0.357
DETR-R50-1/1 (student)	27,007,174	14.4	53	0.617	0.366
DETR-R50-2/2 (student)	29,900,998	14.6	43	0.639	0.425
DETR-R50-3/3 (student)	32,794,822	14.8	38	**0.669**	**0.450**
3D-DETR-R50-6/6 (teacher)	41,477,149	15.7	21	0.784	0.595

Tables 1 & 2. Model sizes, inference speeds, and mAP of attention distilled student DETR models compared to a large 2D teacher model **(upper)** and 3D teacher model **(lower)**. DETR-R50-*n*/*n* refers to the model type, where *n*/*n* indicates the number of encoder and decoder layers. All student models were trained with α =0.7. For comparison, a baseline model trained without attention distillation (α = 0) is also shown, as well as comparison to a Faster R-CNN model. Reported inference speeds (FPS) are based on model inference on a P100 GPU.

References

Carion et al. (2020). "End-to-end object detection with transformers". in European Conference on Computer Vision, pages 213–229. Springer, 2020

Hinton et al. (2015). "Distilling the knowledge in a neural network". In NIPS Deep Learning and Representation Learning Workshop, 2015. URL http://arxiv.org/abs/ 1503.02531.

PHILIPS