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

Jonathan Rubin[1] JONATHAN.RUBIN@PHILIPS.COM
Ramon Erkamp[1] RAMON.ERKAMP@PHILIPS.COM
Ragha Srinivasa Naidu[1] RAGHAVENDRA.SRI@PHILIPS.COM
Anumod Odungatta Thodiyil[2] ANUMOD.T@PHILIPS.COM
Alvin Chen[1] ALVIN.CHEN@PHILIPS.COM
[1] *Philips Research North America, Cambridge MA, United States*
[2] *Philips Innovation Campus, Bangalore, India*

Abstract

We introduce a method for efficient knowledge distillation of transformer-based object detectors. The proposed "attention distillation" makes use of the self-attention matrices generated within the layers of the state-of-art detection transformer (DETR) model. 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 further investigate distilling spatio-temporal information captured by 3D detection transformer networks into 2D object detectors that only process single frames. We apply the approach to the clinically important problem of detecting medical instruments in real-time from ultrasound video sequences, where inference speed is critical on computationally resource-limited hardware. We observe that, via attention distillation, student networks are able to approach the detection performance of larger teacher networks, while meeting strict computational requirements. Experiments demonstrate notable gains in accuracy and speed compared to detection transformer models trained without attention distillation.

Keywords: Knowledge Distillation, Self-Attention, Object Detection, Transformers, Medical Ultrasound Imaging

1. Introduction

Transformers and the self-attention mechanism Vaswani et al. (2017) have recently permeated almost every area of modern deep learning. Beginning with their tremendous success in the field of natural language processing Brown et al. (2020); Devlin et al. (2018), transformers have also been applied to image recognition Dosovitskiy et al. (2020) and object detection Carion et al. (2020) tasks, as well as graph processing Choi et al. (2020) and reinforcement learning Chen et al. (2021). One shortcoming of large transformer models is that they require extensive amounts of computation, making them slow to train and use, particularly when inputs into the network are large. Vision and detection transformers that recognize and detect objects in images, respectively, are particularly susceptible to the above computational issues, due to the large input sizes required by these networks. There have been several attempts to improve the computational efficiency of the transformer and self-attention mechanism Touvron et al. (2020); Wang et al. (2020); Xiong et al. (2021).

In this work, we focus on strategies to distill large detection transformers (DETR) Carion et al. (2020) into smaller models for the purpose of reducing model size and speeding up inference. Our approach relies on the observation that *self-attention matrices* offer a natural and elegant means for applying knowledge distillation to object detectors. The original knowledge distillation formulation Hinton et al. (2015) allows smaller student models to generalize from the teacher by taking advantage of **'soft target'** class probabilities, which have higher entropy and information content than 'hard' ground truth targets. Self-attention matrices extracted from a teacher detection transformer allow a corresponding learning mechanism to distill object detectors, i.e. they offer soft probability **'heat maps'** that can be used for *localization* distillation, in addition to 'hard' box labels.

We demonstrate the utility of the approach on the clinically relevant task of detecting medical instru-

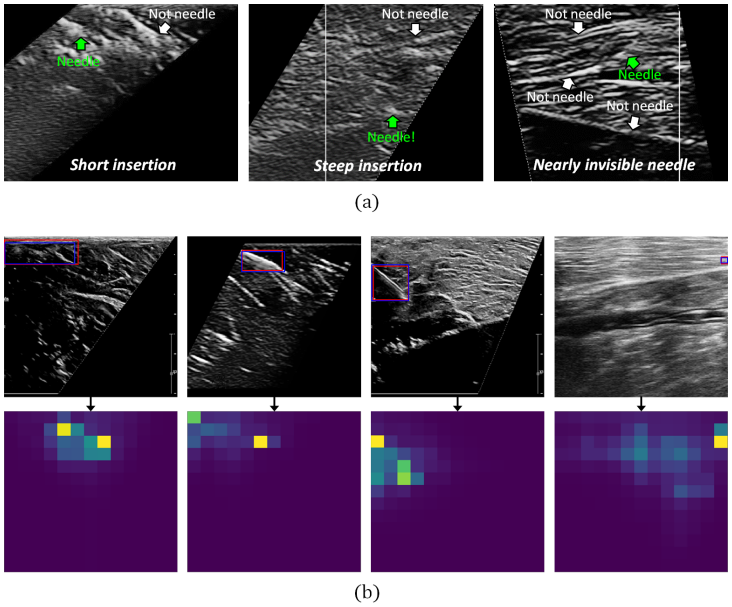

Figure 1: (a) Representative ultrasound frames acquired during live needle insertion procedures on *ex vivo* and human cadaver tissues. The examples highlight the difficulty of correctly identifying the needle in the presence of noise and surrounding tissue structures. (b) Representative self-attention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

ments (needles) in real-time from ultrasound video sequences (Figure 1(a)). Needle detection is a uniquely challenging problem where accuracy, speed, and computational resource-efficiency are critical. We quantify the benefits of attention distillation and show that large teacher networks consisting of several encoder and decoders layers can be distilled to much smaller single encoder/decoder detection transformers. Experimental results suggest that it is possible to *double the number of frames per second processed by the student network* while only taking a minor performance hit compared to using larger teacher networks.

The contributions of this work are as follows:

1. We introduce attention distillation for the state-of-art detection transformer model by leveraging self-attention matrices as an information-rich feature representation.

2. We formulate a loss that allows a natural transition between traditional localization and class losses and an attention distillation loss.

3. We study attention distillation as a learning mechanism by measuring the dissimilarity (Kullback Leibler divergence) between student and teacher self-attention matrices during training.

4. We investigate the specific contribution of the attention distillation mechanism on model performance, independent of other network design choices. Empirical results indicate consistent improvements in performance and inference speed compared to equivalent models trained without attention distillation.

5. We further extend to 3D-to-2D distillation, where spatio-temporal information learned by a 3D teacher operating on video data is compressed into a 2D student model that only processes single frames.

6. We apply and evaluate detection transformers and attention distillation on a relevant biomedical use case of ultrasound video detection.

Our contributions are aimed at constructing faster and more lightweight image and video object detectors. As in the original DETR model Carion et al. (2020), we use the well-established ResNet-50 network backbone; however, like DETR, our methods

are not limited to specific backbones, and other popular convolutional architectures may be considered based on application-specific requirements.

2. Related Prior Work

2.1. Instrument Detection in Ultrasound

In this study we aimed to demonstrate the real-world utility of attention distillation on the problem of detecting medical instruments, particularly needles, in real-time from medical ultrasound video sequences. This is a problem of strong clinical relevance, as needle insertions are one of the most commonly performed medical procedures in point-of-care, emergency, and interventional medical settings to enable vascular access, regional anesthesia, delivery of fluids and drugs, drainages, and biopsies. Ultrasound-based needle/instrument guidance is often carried out by trained clinical specialists, since the procedure requires expertise and represents a significant injury risk when difficulties occur. Visualization under ultrasound can be challenging due to the presence of confounding tissue structures, speckle noise, and poor instrument visibility (1(a)). **Short insertions**, such as when the instrument is just entering the imaging field, are especially problematic.

Earlier studies have proposed a range of image processing and machine learning approaches to recognize needles and other instruments in ultrasound imagery Beigi et al. (2021). Techniques include use of spatial or phase-based features Draper et al. (2000); Hacihaliloglu et al. (2015) as well as features derived from learning models Hatt et al. (2015); Beigi et al. (2017); Mwikirize et al. (2018); Lee et al. (2020). Some studies have made additional use of temporal information from ultrasound video frames Mwikirize et al. (2019); Beigi et al. (2017). Recently, Rubin et al. (2021) compared video-based deep learning approaches, including two-stream neural networks and 3D spatio-temporal convolution approaches.

Currently, processing speed and computational resource limitations remain major challenges in the efficient deployment of existing detection models for real-time instrument guidance. In order for the overall pipeline (which typically involves image acquisition and beamforming, low-level pre-processing, compression and scan conversion, model inference, post-processing, and visual rendering/display) to be performed in real-time during live scanning, the speed of the model inference step must often exceed 50-100 frames per second to avoid becoming a bottleneck. Increasing the detection power of the models, for example by encoding 3D spatio-temporal information or using larger and more powerful architectures such as DETR Carion et al. (2020), further exacerbates the computational demand.

2.2. Knowledge and Attention Distillation

Knowledge distillation Buciluǎ et al. (2006); Hinton et al. (2015) has been used as a model compression technique within a variety of contexts and applications Gupta et al. (2016); Luo et al. (2018); Romero et al. (2014); Zagoruyko and Komodakis (2017). Several studies have applied knowledge distillation within the context of video representation learning to distill motion *features* from an optical flow network to RGB-only networks Crasto et al. (2019); Stroud et al. (2020). Liu et al. (2020) went one step further and used attention distillation to distill motion knowledge into an RGB network using the soft attention map of a flow-based teacher network. Zagoruyko and Komodakis (2017) investigated attention transfer from a teacher network to smaller student networks using activation-based and gradient-based spatial attention maps. They focused on image classification tasks where the teacher was a deeper residual network and an ℓ_2 attention transfer loss was applied. Knowledge distillation strategies specifically for transformers were investigated in Touvron et al. (2021) where distillation tokens were introduced for teacher-student networks.

2.3. Distilling Object Detectors

While the above works focused primarily on action recognition and related tasks, there have also been prior attempts to distill object detectors. Chen et al. (2017) introduced a number of concepts for distilling detectors, including a weighted cross-entropy loss to address class imbalance, a teacher bounded loss to handle the regression component, and adaptation layers to learn from intermediate teacher distributions. Wang et al. (2019) investigated localization distillation by selecting anchor boxes that are near ground truth bounding boxes and forcing a student network to imitate a larger Faster R-CNN Ren et al. (2015) detector. They compared their approach to hint learning Romero et al. (2014). More recently, Guo et al. (2021) showed improved object detection results for distilled Faster R-CNN and RetinaNet Lin et al. (2017) models using *decoupled features*. All of

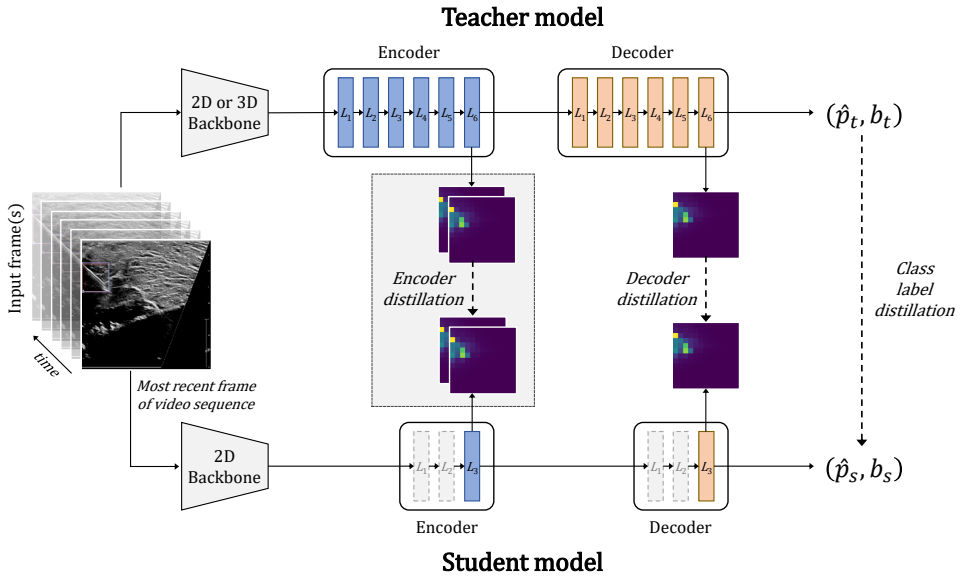

Figure 2: Overview of attention distillation for detection transformers. Self-attention matrices of large 2D or 3D teacher networks are used to distill localization information to smaller student networks consisting of fewer encoder and decoder layers. Attention distillation can take place within the encoder or decoder layers (in this work, experiments focused on *encoder*-based distillation). Class label distillation can also optionally be applied.

the above works relied on single and two-stage detectors and hence did not make use of the self-attention matrices - which are naturally produced by detection transformers - for localization distillation.

3. Detection Transformers

We utilize DETR Carion et al. (2020) as a state-of-art object detection architecture. The DETR architecture demonstrates several benefits over traditional single and two-stage object detectors.

- First, the DETR architecture is conceptually simple. It consists of a backbone convolutional neural network (e.g. ResNet50) that downsamples an input image to produce a tensor of activations that are then processed by an encoder-decoder transformer architecture that directly predicts a set of bounding boxes. There is no need for anchor boxes or non-maximum suppression. Instead, the architecture relies on bipartite matching and imposes a parameter, N, that limits the maximum number of objects that can be detected in an image.
- Second, DETR has shown improvements in accuracy and speed compared to two-stage detectors, such as Faster R-CNN, on common object detection benchmark datasets Carion et al. (2020).
- Finally, the self-attention maps produced by detection transformers provide a high-content and visually explainable representation of object locations and appearances. For biomedical applications, these maps can increase the transparency of automated medical imaging systems, helping end-users to understand the salient features used by the model for prediction and providing a mechanism for human clinical review.

Figure 1(b) shows examples of self-attention maps for four representative, randomly selected ultrasound frames containing needle objects. For our purposes, the bipartite matching required by the DETR architecture is trivial, as there is either no object ($\emptyset$) or at most only one object to detect within an ultrasound frame, hence we fix this limit to be $N = 1$. For a single instance, y_i, ground truth class labels and bounding box information is denoted by $y_i = (c_i, b_i)$, where c_i is either $\emptyset$ or the target class label, i.e. nee-

dle, and $b_i \in [0,1]^4$ is a vector that defines the standardized $center_x$, $center_y$, $width$ and $height$ for the ground truth bounding box. The probability of predicting class $c_i \in \{\emptyset, 1\}$, where 1 is the object class, is given by $\hat{p}_i(c_i)$ and $\hat{b}_i$ is the predicted bounding box. The bounding box loss (as defined in Carion et al. (2020)) is a linear combination of ℓ_1 loss and the scale-invariant generalized IoU loss Rezatofighi et al. (2019), given by Eq. (1), where λ_{iou}, λ_{L1} are hyper-parameters that control mixing between the loss terms:

$$\mathcal{L}_{\text{box}}\left(b_i, \hat{b}_i\right) = \lambda_{\text{iou}}\, \mathcal{L}_{\text{iou}}\, \left(b_i, \hat{b}_i\right) + \lambda_{L1} \left\| b_i - \hat{b}_i \right\|_1 \tag{1}$$

Eq. (1) combines the two losses, as ℓ_1 alone will result in different scales for small and large boxes.

4. Attention Distillation

We apply attention distillation by making use of self-attention matrices generated within the encoder-decoder detection transformer architecture. Figure 2 shows an overview of the approach. Recall, that a backbone convolutional neural network (e.g. ResNet50) processes an input image and learns a downsampled feature representation, $f \in \mathbb{R}^{C \times H \times W}$. The number of channels in the learned representation is first reduced using 1x1 convolution and then the H and W dimensions flattened to give the sequence $(x_1, \ldots, x_n)$, where $x_i \in \mathbb{R}^{HW}$ is fed to the detection transformer encoder, along with positional encodings Carion et al. (2020).

Multi-headed scaled dot-product attention Vaswani et al. (2017) is applied to learned query and key matrices (Q and K, respectively) by multiplying each x_i in the sequence by learned network weight matrices, W^Q and W^K.

$$\text{A} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2}$$

In Eq. (2), A is the attention matrix and d_k is the size of the multi-headed attention hidden dimension chosen as a hyper-parameter. The idea behind attention distillation is to force the self-attention matrix of a small student network, A^s, to be similar to that of a larger teacher network, A^t. We use the Kullback Leibler divergence between student and teacher self-attention matrices to accomplish this, as illustrated in Eq. (3).

$$\begin{aligned}\mathcal{L}_{distill} =& (1-\alpha)\cdot \mathcal{L}_{\text{box}}\left(b_i, \hat{b}_i\right) + \\ & \alpha \cdot \left(\mathcal{KL}\left(A_i^s \| A_i^t\right) + T^2 \cdot \mathcal{KL}\left(\sigma(\frac{\hat{p}_i^s}{T}) \| \sigma(\frac{\hat{p}_i^t}{T})\right)\right)\end{aligned} \tag{3}$$

In Eq. (3), α is a hyper-parameter that controls mixing between the bounding box loss, $\mathcal{L}_{\text{box}}$, and the attention distillation loss. Recall that 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 attention distillation loss, $\mathcal{KL}\left(A_i^s \| A_i^t\right)$, applies knowledge distillation to the self-attention maps created by teacher and student detection transformers. It attempts to match the distribution of the attention maps between teacher and student networks. The attention maps can come from either the encoder or decoder. The second component of the attention distillation loss optionally applies knowledge distillation to the class label predictions, $T^2 \cdot \mathcal{KL}\left(\sigma(\frac{\hat{p}_i^s}{T}) \| \sigma(\frac{\hat{p}_i^t}{T})\right)$, where T is a temperature hyper-parameter that controls smoothing, as in Hinton et al. (2015) and σ is the softmax operation. The overall loss, $\mathcal{L}_{distill}$, is applied to mini-batches of data samples.

Attention distillation can also be used to distill a 3D detector, designed to process a temporal sequence of multiple frames, into a 2D student model that processes only a single frame. The additional size and complexity of the 3D detectors, and their reliance on 3D convolution operations, leads to increased processing times compared to 2D counterparts. 3D-to-2D distillation allows a 2D student model to ingest temporal information from a 3D teacher, while maintaining low computational complexity.

5. Results

5.1. 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 a period of 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 (e.g. gain, depth, and tissue-specific presets), needle types, needle sizes, insertion angles, and bevel orientations.

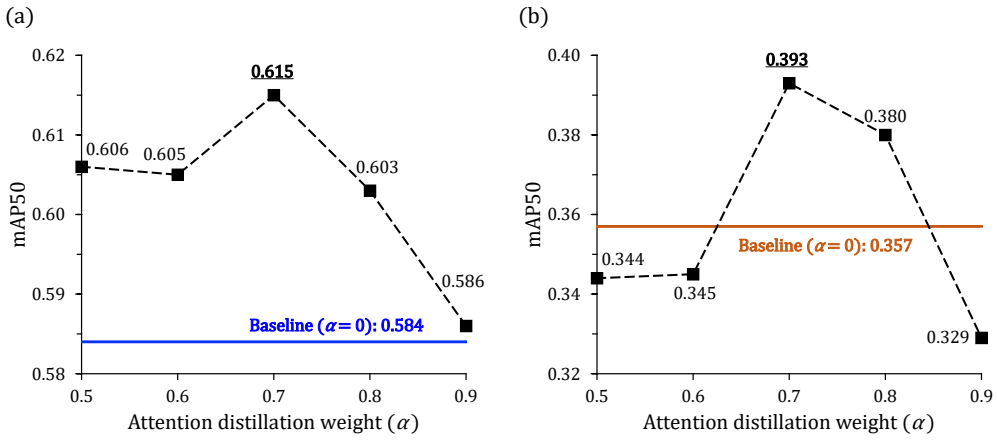

Figure 3: Results for a sweep over α values from 0.5 – 0.9 using a DETR-R50-1/1 student network with attention distillation applied at the final encoder layer. Baseline at $\alpha = 0$ (no distillation) is also shown. (a) mAP results on test dataset. (b) mAP_{short} on a challenging dataset containing only short insertions.

Full video sequences, each typically several hundred frames in length (up to 60 sec in time) were divided into short 30-frame clips spanning about 1 sec in time. Ground-truth labels were generated in the form of bounding box localizations for the last frame of each clip (example labeled frames are shown in Figure 1(b)). 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. In both sets, approximately 60% of frames were labeled with the needle present, and the remaining 40% were negative, representing a clinically realistic prevalence. All 2D models took as input the last frame of each video clip, which contained an associated label. 3D models took all or part of the entire video clip, depending on the model's input size along the third dimension. All networks were trained using PyTorch for 50 total epochs on a P100 GPU.

5.2. 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) consisting of a single encoder and decoder. We perform a sweep over the α hyper-parameter given in Eq. (3) to control how much of the attention distillation loss to include during training. We selected α to be between the values 0.5 – 0.9, as well as $\alpha = 0$, i.e. no attention distillation. To evaluate the exact contribution of attention distillation to the performance of the network, we fix the parameters of the R50 backbone and only train the encoder-decoder transformer sub-network. Furthermore, we only rely on the attention distillation component of the loss function and omit the optional knowledge distillation from class label predictions. The size of the transformer's hidden dimension was 256 and a total of 8 attention heads were used.

Figure 3 shows mAP^{50} (left) and mAP^{50}_{short} (right) results for a series of attention distilled student models where α varies between values 0.5 – 0.9 (x-axis). The mAP^{50}_{short} results refer to the particularly challenging use-case of **short** needle insertions, where the needle tip is just barely entering the ultrasound field of view. In these experiments, we chose to apply attention distillation on the *final layer of the encoder stack*, $A \in \mathbb{R}^{HW \times HW}$. The results of encoder-based attention distillation are compared to the baseline model with α=0 (flat lines).

Figure 3 shows that the use of attention distillation leads to improved mAP^{50} and mAP^{50}_{short} scores compared to a baseline model trained on localization labels alone. In each case α=0.7 gives the best performance (*0.615* vs. *0.584* and *0.393* vs. *0.357* for mAP^{50} and mAP^{50}_{short}, respectively).

Model	Parameters	GMac	FPS	$\mathbf{mAP}^{50}$	$\mathbf{mAP}^{50}_{short}$
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

Table 1: Model sizes, inference speeds, and average precision results of attention distilled student DETR models compared to a large 2D teacher model. 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 $\alpha = 0.7$. For comparison, a baseline model trained without attention distillation ($\alpha = 0$) is also shown, as well as comparison to a Faster R-CNN model. The reported inference speeds (FPS) were based on model evaluation on a P100 GPU.

Model	Parameters	GMac	FPS	$\mathbf{mAP}^{50}$	$\mathbf{mAP}^{50}_{short}$
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

Table 2: Model sizes, inference speeds, and average precision results for attention distilled student DETR models compared to a large 3D teacher model that incorporates a history of temporal information. All student models were trained with $\alpha = 0.7$. The reported inference speeds (FPS) were based on model evaluation on a P100 GPU.

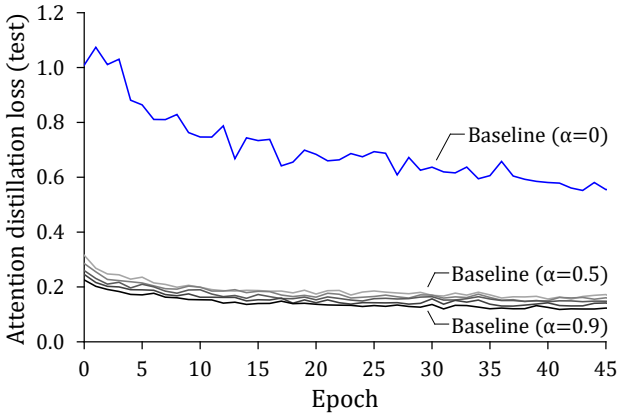

Figure 4: Attention distillation loss, i.e. the $\mathcal{KL}$ divergence between teacher and student self-attention maps, on test data. For α=0 (no distillation), there remains a large gap compared to student networks trained with attention distillation $\alpha \in \{0.5 \dots 0.9\}$.

Table 1 summarizes model sizes, inference speeds, and average precision results for a series of student networks, and compares these to the DETR-R50-6/6 teacher model, a state-of-the-art Faster R-CNN model, and a baseline model trained without attention distillation. Here, we trained student models starting with single (1/1) encoder-decoder pairs up to 3/3. All student models were trained with the attention distillation hyper-parameter fixed to α=0.7. We can see that the student models approach the mAP^{50} and mAP^{50}_{short} performance of the full teacher model while improving upon the processing frame rate. In particular, the DETR-R50-2/2 student model achieves a mAP^{50} of *0.643* vs. *0.655* for the full teacher model, while increasing the frame rate from 26 to 43 FPS. While the Faster R-CNN model achieves strong detection results (*0.773* mAP^{50} and *0.681* and mAP^{50}_{short}, its heavy computational cost and slow processing speeds (134.1 GMac, 19 FPS) make it infeasible for use on resource-limited ultrasound hardware. Furthermore, all the parameters of the Faster R-CNN model are trainable, whereas

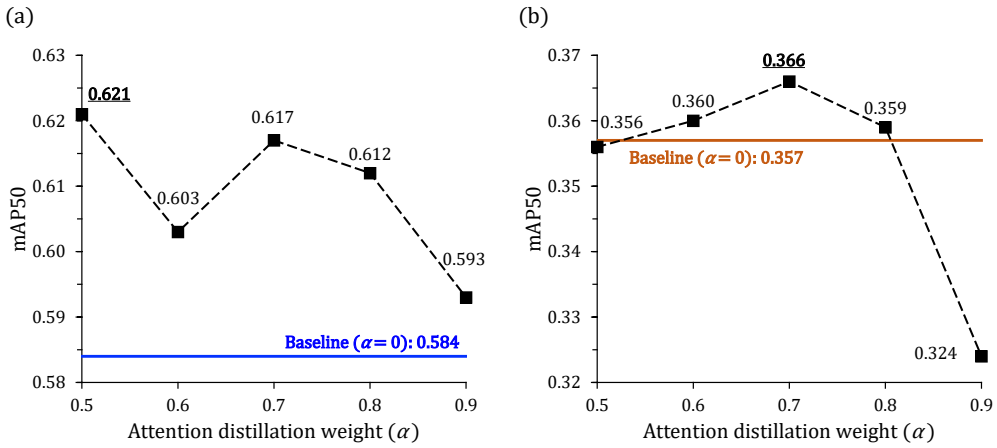

Figure 5: 3D attention distillation results for a sweep over α values from 0.5 – 0.9, again with attention distillation applied at the final encoder layer. Baseline at $\alpha = 0$ (no distillation) is also shown. (a) mAP results on test dataset. (b) mAP_{short} on challenge dataset containing only short insertions.

we fix the weights of the R50 backbone for the DETR models and trained only the transformer sub-networks in order to study the attention distillation mechanism in isolation).

It could be expected that without employing attention distillation, the distribution of a student self-attention map would naturally start to resemble that of a fully trained teacher network over training epochs. Figure 4 plots the $\mathcal{KL}$ divergence between student and teacher self-attention matrices, i.e. the attention loss, for a baseline model with α=0 (top **blue** curve) as well as a number of student models trained with $\alpha \in \{0.5 \ldots 0.9\}$ (lower curves). It can be seen that the $\mathcal{KL}$ divergence decreases similarly for all student models where $\alpha > 0$, whereas the attention loss is significantly higher in the baseline model, indicating dissimilar attention distributions compared to the teacher when α=0.

5.3. 3D-to-2D Attention Distillation for Videos

We further extend the concept of attention distillation to compress a 3D detection transformer, which takes into account a temporal history of ultrasound frames, into a 2D student model that processes single frames independently. Here, we utilized a 3D-DETR-R50-6/6 model as the teacher network. The network applies a series of k=6 spatio-temporal 3D convolutions to a stack of ultrasound video frames. The effect is that the temporal dimension is *convolved out* to arrive at a 2D feature map with a singular temporal dimension, k=1, while image width and height remains unchanged. The learned feature map is then fed to the remainder of the network that has the same architecture as the 2D DETR-R50-6/6 teacher network described previously.

We emphasize that the detection problem is exactly the same as in the 2D-to-2D attention distillation experiments (Section 5.2), as is the architecture and number of parameters of the student networks. The only thing that has changed is that the teacher network now has access to a temporal history of frames, which it encodes via spatio-temporal convolution to inform detection predictions. As before, we fix the weights of the R50 backbone and train only the transformer sub-network.

Figure 5 shows the results of applying attention distillation using a DETR-R50-1/1 student network and a sweep of α between 0.5 to 0.9, with the 3D-DETR-R50-6/6 network as the teacher. Also shown is a comparison to the baseline network where no attention distillation is used, i.e. α=0. As in the 2D-to-2D experiments, we applied attention distillation on the *final encoder layer*. Once again, models trained with attention distillation are observed to outperform the baseline model trained on localization labels alone.

Finally, we explored the effects of increasing the capacity of the 3D-to-2D student networks. Table 2 shows speed and average precision results for stu-

dent networks (α=0.7) that range from 1/1 encoder-decoder layers to 3/3. The full 3D-DETR-R50-6/6 teacher model achieved scores of *0.784* and *0.595* for mAP^{50} and mAP$^{50}_{short}$, respectively. Meanwhile, as we increase the capacity of the 2D student networks, mAP^{50} and mAP$^{50}_{short}$ scores approach that of the 3D teacher network. Interestingly, we also notice that the mAP^{50} score of the 3D distilled DETR-R50-3/3 student network outperforms the full 2D-DETR-R50-6/6 teacher network reported in Table 1 (*0.669* vs. *0.655*). Visual detection results for the baseline (α=0), as well as 2D and 3D student and teacher models are given in Appendix A.

6. Discussion and Future Work

Overall, our experiments show consistent improvements in performance when attention distillation is employed compared to baseline models trained only on localization information (α= 0). As expected, student models improve as extra encoder and decoder layers are added to the transformer sub-network. Interestingly, however, the capacity of the network alone does not account for the observed improvements. In particular, the 3D-to-2D DETR-R50-3/3 student network (trained with a 3D temporally-aware teacher) outperforms its counterpart 2D-to-2D student model (trained with a 2D teacher) (*0.669 vs. 0.633* mAP^{50} and *0.450 vs. 0.445* mAP$^{50}_{short}$) despite having the same number of parameters. Furthermore, the student model trained with a 3D teacher achieves higher detection performance than even the full 2D teacher (*0.669 vs. 0.655* mAP^{50}).

There are several limitations in this work, and areas for future investigation:

1. Self-attention maps of the teacher and student network are required to have the same dimension, which limits the flexibility of the backbone network. Adaptation layers, such as proposed in Guo et al. (2021), could be one way to support attention maps of different dimensionality.

2. The current work applied attention distillation at the final *encoder* layer in the detection transformer. Interestingly, we did not see consistent improvements when we instead applied attention distillation to the final *decoder* layer, and more investigation is needed to understand why. Future work will also evaluate whether performance can be further improved for student networks that consist of more than one encoder/decoder layer by applying attention distillation within *each* layer (rather than only the final layer).

3. Presently, dissimilarity between teacher and student self-attention maps was incorporated into the loss function via $\mathcal{KL}$ divergence. Other similarity measures, such as based on optimal transport Peyré et al. (2019), could be attempted.

4. Our models were only required to detect a single needle or no needle within each frame, i.e. $N = 1$. Detection performance for multiple and varied objects should be evaluated.

5. To isolate the contribution of the attention distillation mechanism, we applied a single consistent backbone network (ResNet50) across all experiments, and we fixed the parameters of the backbone such that only the transformer sub-network was influenced by distillation during training. It may be possible to improve performance further by optimizing the full network end-to-end. Generalization to other state-of-art backbone architectures, e.g. Ren et al. (2015); Chollet (2017); Gholami et al. (2018); Howard et al. (2019); Wang et al. (2021), should also be investigated.

6. Further progress toward clinical translation can be achieved by integrating the student models on real ultrasound systems and evaluating inference times during live clinical procedures.

7. Finally, validation of the approach was demonstrated on an ultrasound dataset selected for its clinical relevance and challenging nature. Further studies are needed to assess attention distillation across a wide range of ultrasound datasets as well as data from other imaging modalities.

7. Conclusions

We have presented an approach for distilling transformer-based object detectors using a novel attention distillation learning mechanism. Empirical results showed notable gains in accuracy and speed over baseline models trained on localization labels only. We demonstrated the utility of attention distillation for the challenging and clinically-relevant problem of medical instrument detection in ultrasound video. However, we posit that the approach may have broader applicability to other biomedical image and video detection tasks where speed is critical and computational resources are constrained.

References

Parmida Beigi, Robert Rohling, Tim Salcudean, Victoria Lessoway, and Gary C Ng. Detection of an invisible needle in ultrasound using a probabilistic svm and time-domain features. *Ultrasonics*, 78: 18–22, 2017.

Parmida Beigi, Septimiu Salcudean, Gary C Ng, and Robert Rohling. Enhancement of needle visualization and localization in ultrasound. *IJCARS*, 16: 169–178, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *31st Conference on Neural Information Processing Systems (NIPS 2017)*, pages 742–751, 2017.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *arXiv preprint arXiv:2106.01345*, 2021.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 606–613, 2020.

François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. MARS: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7882–7891, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

KJ Draper, CC Blake, L Gowman, DB Downey, and A Fenster. An algorithm for automatic needle localization in ultrasound-guided breast biopsies. *Medical Physics*, 8:1971–1979, 2000.

Amir Gholami, Kiseok Kwon, Bichen Wu, Zizheng Tai, Xiangyu Yue, Peter Jin, Sicheng Zhao, and Kurt Keutzer. Squeezenext: Hardware-aware neural network design. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1638–1647, 2018.

Jianyuan Guo, Kai Han, Yunhe Wang, Han Wu, Xinghao Chen, Chunjing Xu, and Chang Xu. Distilling object detectors via decoupled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2154–2164, 2021.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016.

Ilker Hacihaliloglu, Parmida Beigi, Gary Ng, Robert Rohling, Salcudean Septimiu, and Purang Abolmaesumi. Projection-based phase features for localization of a needle tip in 2dcurvilinear ultrasound. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9349, 2015.

Charles Hatt, Gary C Ng, and Vijay Buyyounouski, Parthasarathy. Enhanced needle localization in ultrasound using beam steering and learning-based segmentation. *Computerized Medical Imaging and Graphics*, 41:46–54, 2015.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL http://arxiv.org/abs/1503.02531.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019.

Jia Yi Lee, Mobarako Islam, Jing Ru Woh, Mohamed Washeem, Lee Ying Clara Ngoh, Weng Kin Wong, and Hongliang Ren. Ultrasound needle segmentation and trajectory prediction using excitation network. *International Journal of Computer Assisted Radiology and Surgery*, 15:437–443, 2020.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Miao Liu, Xin Chen, Yun Zhang, Yin Li, and James M Rehg. Attention distillation for learning video representations. In *BMVC*, 2020.

Zelun Luo, Jun-Ting Hsieh, Lu Jiang, Juan Carlos Niebles, and Li Fei-Fei. Graph distillation for action detection with privileged modalities. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 166–183, 2018.

Cosmas Mwikirize, John L Nosher, and Ilker Hacihaliloglu. Convolution neural networks for real-time needle detection and localization in 2d ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, 13:647–657, 2018.

Cosmas Mwikirize, John L Nosher, and Ilker Hacihaliloglu. Learning needle tip localization from digital subtraction in 2d ultrasound. *International Journal of Computer Assisted Radiology and Surgery*, 14:1017–1026, 2019.

Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*, 2014.

Jonathan Rubin, Alvin Chen, Anumod Odungattu Thodiyil, Raghavendra Srinivasa Naidu, Ramon Erkamp, Jon Fincke, and Balasundar Raju. Efficient video-based deep learning for ultrasound guided needle insertion. *Medical Imaging with Deep Learning (MIDL)*, 2021.

Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. D3D: Distilled 3d networks for video action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 625–634, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021.

Appendix A. Detection Results

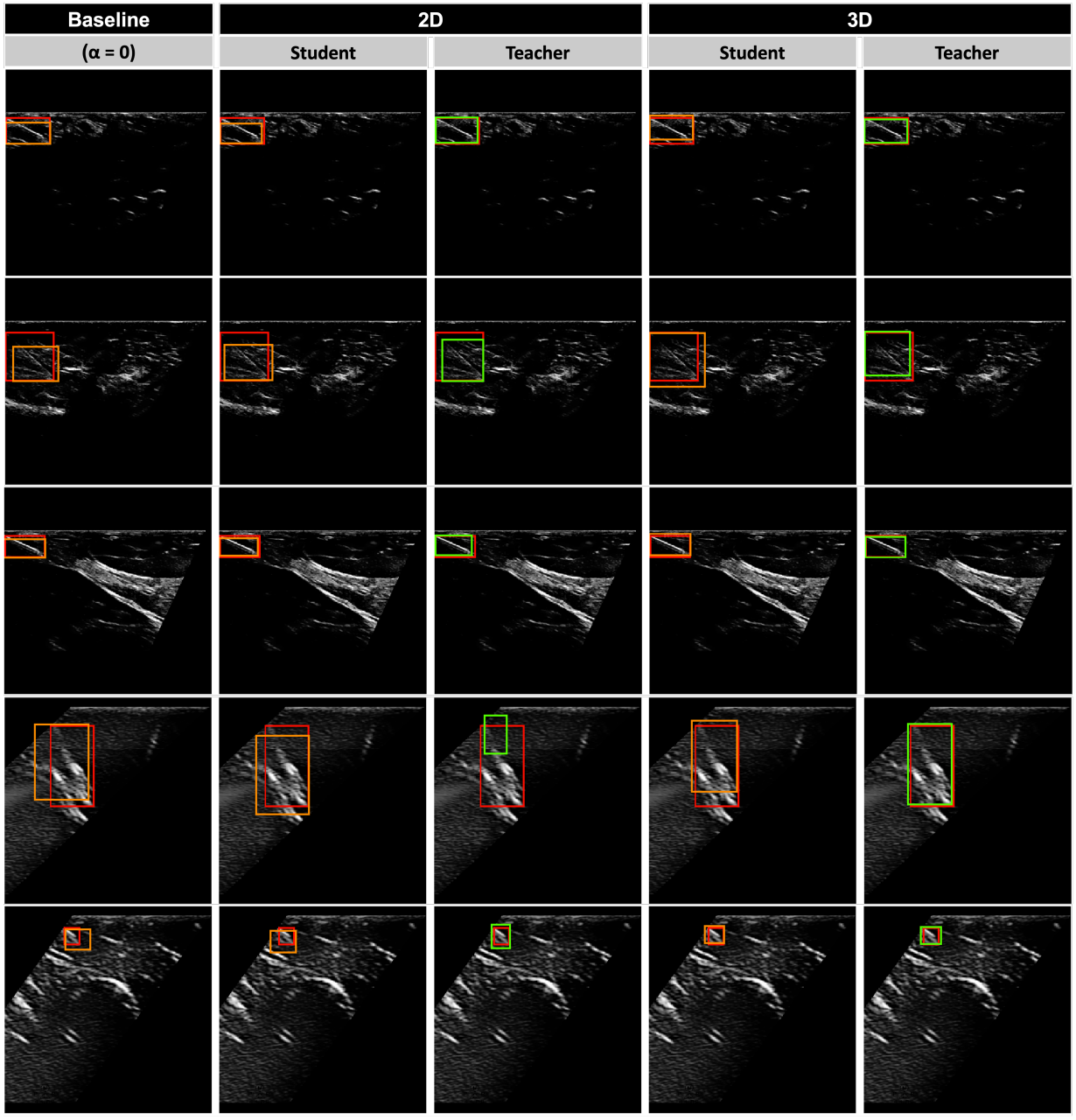

Figure A.1: Qualitative detection results for baseline (α=0), 2D student and teacher models and 3D student and teacher models, respectively. Ground truth bounding boxes are shown in red, student predictions in orange and teacher predictions in green.