

Differentiable Patch Selection for Image Recognition

Jean-Baptiste Cordonnier^{1†*} Aravindh Mahendran^{2†} Alexey Dosovitskiy²
Dirk Weissenborn² Jakob Uszkoreit² Thomas Unterthiner²
¹EPFL, Switzerland ²Google Research, Brain Team

jean-baptiste.cordonnier@epfl.ch

{aravindhm,adosovitskiy,diwe,usz,unterthiner}@google.com

Abstract

Neural Networks require large amounts of memory and compute to process high resolution images, even when only a small part of the image is actually informative for the task at hand. We propose a method based on a differentiable Top-K operator to select the most relevant parts of the input to efficiently process high resolution images. Our method may be interfaced with any downstream neural network, is able to aggregate information from different patches in a flexible way, and allows the whole model to be trained end-to-end using backpropagation. We show results for traffic sign recognition, inter-patch relationship reasoning, and fine-grained recognition without using object/part bounding box annotations during training.

1. Introduction

High-resolution imagery has become ubiquitous nowadays: both consumer devices and specialized sensors routinely capture images and videos with resolution in tens of megapixels. Processing these high-quality images with computer vision models remains challenging: analyzing the images at full resolution can be prohibitively computationally expensive, while simply downsampling them before processing may remove important fine details and substantially hurt performance. It would be desirable to save compute, while retaining the capability to recognize fine details.

Compute can be saved by exploiting the following property of many practical vision tasks: not all parts of the image are equally important for finding the answer. Figure 1 shows examples of tasks where only a small fraction of the full image needs to be processed in detail. Being able to quickly discard uninformative parts of the image would have several benefits. It would reduce the overall computational and memory complexity of the model, and the regions of interest could be processed in more detail and by a more power-



Figure 1: Examples of large images where patch extraction allows (*top-left*) to focus on details for fine-grained recognition, (*bottom-left*) to reason across patches, and (*right*) to efficiently capture very localized information.

ful model than otherwise.

Determining which parts of the image to retain and which to discard is usually nontrivial and highly task dependent. In some applications the solution might be as simple as taking the center crop of the image, but in most cases relevant regions need to be detected first. For instance, in a self-driving car setting, it would be permissible to ignore the sky, but all traffic signs in sight should be correctly identified and must not be ignored. One may formulate this as follows: Given a regular grid of equally sized image patches, decide for each patch whether to process or discard it. This decision is however discrete, which makes it unsuitable for end-to-end learning.

To overcome this limitation, inspired by the work of Katharopoulos & Fleuret [22], we formulate patch selection as a ranking problem, where per-patch relevance scores are predicted by a small ConvNet and the Top K scoring patches are selected for downstream processing. We make this end-to-end trainable with backpropagation using the perturbed maximum method of Berthet *et. al* [5]. We present this as a generic module for patch selection. Our approach is most effective when the majority of patches in the image are irrelevant to the target, but the model a priori

*Work done during internship at Google Research. † Equal contribution.

does not know where in the image the important patches are present. Hence, we do not aim to achieve image coverage such as in semantic segmentation and object detection.

In the remainder of this paper, we will formulate patch selection for image recognition as a Top-K selection problem in section 3, apply the perturbed maximum method to construct an end-to-end model trainable via backpropagation in section 4.2, and demonstrate wide applicability of this method via empirical results in three different domains: (1) street sign recognition, (2) inter-patch relationship reasoning on synthetic data, and (3) fine-grained classification without using object/part bounding box annotations during training and evaluation (Section 5).

2. Related Work

Region proposal methods : Several computer vision methods extract regions of interest from the image. Two stage object detection approaches, for instance, select regions of interest using region proposal networks [35] or hand crafted heuristics [38, 15]. Selected regions are later processed by a separate stage of the model. These methods use the non-differentiable RoI-Pooling [15] or the differentiable RoI-Align [16]. Such architectures require bounding box supervision to train large scale object detection models, whereas our experiments focus on a simpler setting and aim to train with weak supervision using only a single class label per image.

Soft attention : In order to attend to specific parts of an image, an alternative approach is to occlude parts of the input by generating attention masks [45]. While this helps the model focus on relevant features [37, 50, 42, 26], become more interpretable [26], or include external data such as image captions [46, 2], the models will typically still process the whole image on a fixed input resolution. Thus they do not lead to any efficiency gains. Another approach [8] would be to process several image resolutions in parallel and use an attention mechanism to pick features from them. It is also possible to employ adhoc losses to extract meaningful patches [32].

Multiple-Instance Learning : A number of works use attention to solve Multiple-Instance Learning (MIL) problems, which are especially common in medical imaging, where images tend to be very large [20, 27]. Here the goal is to label a set of related input samples, such as slices of organ scans or large images that are decomposed into patches. While, for example, the method of Ilse *et. al* [20] can be used to identify the most relevant patches, this is not leveraged to make computation more efficient, as all image patches are processed in equal detail by this method.

Sequential “glimpses” : There is a long line of work that sequentially processes a sequence of patches (“glimpses”), from a network, until they settle on the most relevant ones [36, 14]. These methods often rely on Reinforcement Learning to train non-differentiable attention mechanisms [33, 4, 28, 12], which typically makes them difficult to train. The Spatial Transformers [21] on the other hand can be deployed as a differentiable attention mechanism, for example for fine-grained recognition of bird species. These have been applied sequentially to extract several regions of interest [13, 24] as part of recurrent neural networks. Training spatial transformers on large images can, however, be difficult because the gradients with respect to the transformation parameters are an accumulation of gradients of sub-pixel bilinear interpolation which can be very local. Angles *et. al* [3] overcome these limitations to train a multiple instance spatial transformers by lifting non-differentiable Top-K by introducing an auxiliary function that creates a heat-map given a set of interest points. Our method, on the other hand, avoids these limitations by computing gradients with respect to all patches in every backward step.

Attention Sampling : Most related to our current work are differentiable methods that sample patches, specifically Attention Sampling (ATS) [22]. ATS requires that the output of the network $f(x; \theta)$ be an expected value over embeddings of all possible patches \mathbf{P} . That is, $f(x, \theta) = \mathbb{E}_{p \sim Z(\mathbf{P})} [f(p; \theta)]$. Using a Monte Carlo approximation of the expectation, f is only applied on a small number of patches sampled according to the distribution $Z(\mathbf{P})$. That is, $f(x, \theta) = \frac{1}{K} \sum_{i=1}^K f(p_i; \theta)$. Thus ATS is restricted to a simplistic average pooling scheme for aggregating information from the extracted patch embeddings. Our method, on the other hand, solves the patch sampling problem using a differentiable Top-K operation, which allows us to combine the per-patch information in a flexible manner.

Differentiable Top-K : The subset sampling operation can be implemented by expanding the Gumbel-Softmax trick [43] or based on optimal-transport formulations for ranking and sorting [44, 10], the latter of which was recently made significantly faster by Blondel *et. al* [6]. Our work uses perturbed optimizers [5] to make a Top-K differentiable, which we found performed better than the Sinkhorn operator [44].

3. Model overview

Our model processes high resolution images by scoring, selecting then processing, some regions of interest. As illustrated in Figure 2, the model consists of a scorer network s_θ , a patch selection module p , a feature network f_ϕ and an aggregation network a_ψ , where θ , ϕ , and ψ are learnable

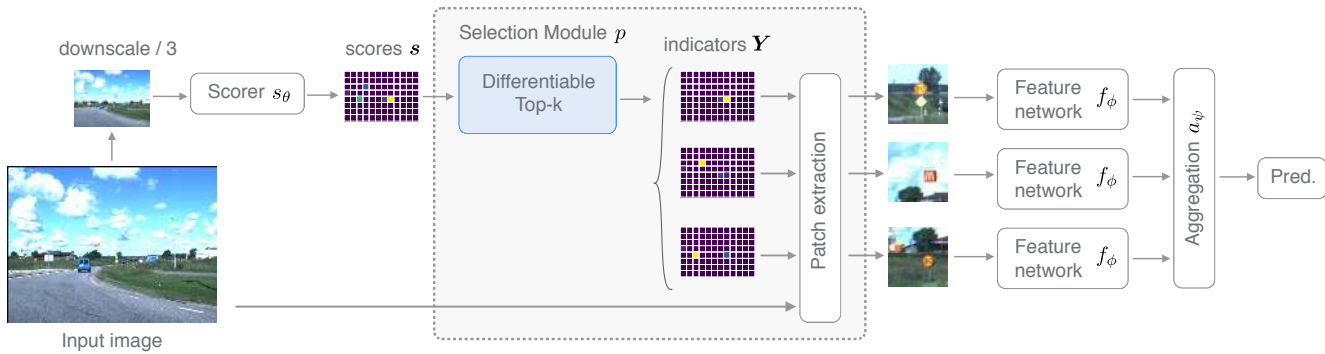


Figure 2: The differentiable Top-K layer uses the scores output by s_θ to extract patches that can be processed by an *arbitrary* downstream network. The whole model is trained end-to-end without any adhoc loss to train the selection module.

parameters. Attention sampling (ATS) [22] is a special case of this where aggregation is an average pool and patch selection is implemented using discrete sampling. To describe the model in one sentence: The scorer scores patches, the patch selection module selects a subset of those, the feature network computes embeddings for each patch, and the aggregation network combines these embeddings to make a final prediction. These modules are described in more detail below.

The **scorer network** s_θ predicts a relevance score for each region of the image with respect to the task at hand. This network is meant to be shallow and typically operates on downsampled images for applications where the original image size is too big. It outputs relevance scores, $\mathbf{S} = s_\theta(\mathbf{X}) \in \mathbb{R}^{h \times w}$, which constitute a $h \times w$ grid. This can be easily generalized to a scorer network that predicts scores for patches at different scales and aspect ratios, much like the region proposal network in Faster-RCNN [35]. Crucially though, our model selects only a small number of regions of interest (e.g. 10), when compared to the 2000 used in Faster-RCNN as necessitated by the high recall requirement for object detection. A **selection module** $p(\mathbf{X}, \mathbf{S})$ uses this information to extract the K most relevant patches from the image. The output of the selection module are K patches of dimension $P_h \times P_w$ denoted by $\tilde{\mathbf{X}} \in \mathbb{R}^{K \times P_h \times P_w \times C}$ ¹. To simplify the notation, we use a single patch size but in practice (see Section 5.3) one can also consider patches at different resolutions and aspect ratios. For each individual patch, a **feature network** f_ϕ calculates a D_h -dimensional representation, which for all patches together is a matrix $\mathbf{H} \in \mathbb{R}^{K \times D_h}$. f_ϕ is typically a computationally large and expensive network. One may use different feature networks for different patches but we did not explore this direction. Finally, an **aggregation network** a_ψ pools this information into the model output

¹Patch sizes need not coincide with the scorer network’s receptive field size. For example, the scorer network might score 32×32 pixel patches, while the selection module may extract patches of size 50×50 pixels.

$\mathbf{y} = a_\psi(\mathbf{H}) \in \mathbb{R}^{D_o}$. This network can be a simple mean pooling operation as in ATS, or an equally simple max pooling operation, or a sophisticated transformer network or anything in between.

Several prior works can be cast in this manner: *Spatial transformers* [21] have localisation networks that are similar to our scorer network, albeit instead of scoring patches they predict localisation parameters such as an affine transformation. Thus scoring and patch selection are effectively done together. Their grid sampler can be interpreted as a patch extraction module. *Attention sampling* [22] sample patches (with or without replacement) according to normalized scores output by a scorer network and use a ResNet for feature extraction f_ϕ . The aggregation network a_ψ is restricted to taking the average of the patch embeddings to be able to back propagate through the discrete sampling operation. *Vision transformer* [11, 9]: the patch selection module is exhaustive and extracts all $P \times P$ patches in the image with a stride P . The feature network f_ϕ and the aggregation network a_ψ are merged into a large transformer that process all the flattened patches with self-attention and output the representation of the CLS token or an average of the tokens’ representations.

4. Patch selection as differentiable Top-K

We aim to train our model end-to-end without introducing any auxiliary losses for individual components. The core technical novelty of our method is the patch selection module. Roughly speaking, we formalize patch selection as the Top-K problem: $\text{Top-K}(x \in \mathbb{R}^N) = y \in \mathbb{N}^K$ where y contains the indices of the K largest entries in x . We then apply the perturbed maximum method [5] to construct our differentiable patch selection module.

4.1. Patch selection as Top-K

Given scores from the scorer network, we select the K highest scores and extract the corresponding patches.

Specifically, given scores $\mathbf{S} \in \mathbb{R}^{h \times w}$ for $N = h \times w$ patches, Top-K returns the indices of the K most salient patches in the image. For reasons that will be clear in the next subsection, we define Top-K such that the indices are sorted. That is, $y_1 < y_2 < \dots < y_K$. Without this constraint Top-K's output could be permuted arbitrarily breaking the perturbed optimizer method. Furthermore, we represent y_i 's as one-hot N dimensional indicator vectors $\{I_{y_1}, I_{y_2}, \dots, I_{y_K}\}$. The intuition here is that if we had a large tensor of all patches in the image, we could "extract" all chosen patches using a single matrix tensor multiplication. Specifically, given all patches, $\mathbf{P} \in \mathbb{R}^{N \times P_h \times P_w \times C}$, and $\mathbf{Y} = [I_{y_1}, I_{y_2}, \dots, I_{y_K}] \in \{0, 1\}^{N \times K}$, patches may be extracted as $\mathbf{X} = \mathbf{Y}^\top \mathbf{P}$ ². This operation is non-differentiable because both Top-K and one-hot operations are non-differentiable.

4.2. Differentiable Top-K

To learn the parameters of the scorer network using back-propagation, we need to differentiate through the patch selection method. We employ the perturbed maximum method [5] for this purpose. Given a non-differentiable module, whose forward pass can be represented as a linear program of the form

$$\arg \max_{\mathbf{Y} \in \mathcal{C}} \langle \mathbf{Y}, \eta \rangle. \quad (1)$$

with inputs η , optimization variable \mathbf{Y} , and convex polytope constraint set \mathcal{C} , the perturbed maximum method defines a differentiable module with forward and backward operations as below.

Forward: Sample uniform Gaussian noise Z and perturb the input η , to generate several perturbed inputs. Solve the linear program for each of these, and then average their results.

$$\mathbf{Y}_\sigma = \mathbb{E}_Z \left[\arg \max_{\mathbf{Y} \in \mathcal{C}} \langle \mathbf{Y}, \eta + \sigma \mathbf{Z} \rangle \right] \quad (2)$$

where σ is a hyper-parameter. In practice one computes the expectation using an empirical mean with n independent samples for Z . We fix $n = 500$ in all our experiments and tune $\sigma \in \{0.01, 0.05, 0.5\}$. This does not require one to solve 500 linear programs in every forward pass, instead as the linear program is chosen to be equivalent to Top-K, we run the Top-K algorithm 500 times, one for each perturbed input, which is very fast in practice.

Backward: Following [1], the Jacobian associated with the above forward pass is

$$J_s \mathbf{Y} = \mathbb{E}_Z \left[\arg \max_{\mathbf{Y} \in \mathcal{C}} \langle \mathbf{Y}, \eta + \sigma \mathbf{Z} \rangle \mathbf{Z}^\top / \sigma \right]. \quad (3)$$

²In practice, we use `jax.lax.scan` as this is memory efficient.

The equations above have been simplified for the special case of normal distributed Z ³.

Patch selection as Top-K with sorted indices is equivalent to the following linear program.

$$\max_{\mathbf{Y} \in \mathcal{C}} \langle \mathbf{Y}, \mathbf{s} \mathbf{1}^\top \rangle. \quad (4)$$

where \mathbf{s} are scores output by the scorer network. $\mathbf{s} \mathbf{1}^\top \in \mathbb{R}^{N \times K}$ are scores replicated K times. $\langle \rangle$ flattens the matrices before taking a dot product. The constraint set is defined as

$$\mathcal{C} = \{ \mathbf{Y} \in \mathbb{R}^{N \times K} : \mathbf{Y}_{n,k} \geq 0, \mathbf{1}^\top \mathbf{Y} = \mathbf{1}, \mathbf{Y} \mathbf{1} \leq \mathbf{1}, \quad (5) \\ \sum_{i \in [N]} i \mathbf{Y}_{i,k} < \sum_{j \in [N]} j \mathbf{Y}_{j,k'} \quad \forall k < k' \}.$$

Note how \mathbf{Y} has the same shape as the concatenation of indicator vectors \mathbf{Y} in the previous subsection. They serve the same purpose. The first conditions encourages that \mathbf{Y} is an assignment where each of the columns has a total weight of one. The last condition results in sorting the indices. This linear program has infinitely many optimal solutions, one of which is the required integer solution corresponding to the index-sorted "Top-K" operator.

Note that in theory, noise should be applied to $\mathbf{s} \mathbf{1}^\top$. The equivalence between this linear program and Top-K crucially relies on all columns of $\mathbf{s} \mathbf{1}^\top + \sigma Z$ being identical. We therefore apply noise directly to \mathbf{s} . Our experiments show that this departure from theory works in practice. We normalize the scorer output \mathbf{s} to lie in $[0, 1]$ with a small $\epsilon = 10^{-5}$ to avoid any division by zero.

While Top-K for each perturbed input will result in one-hot indicators \mathbf{Y} , their perturbed average may be far from one-hot. Thus early in training, when the scores are still non-decisive, extracted patches resemble a weighted average over all image patches (Figure 3). Mixing images like this might contribute to model generalization [48]. Another side benefit is that backpropagated gradients take into account all image patches and can update their weights from the very first step. A discrete sampler on the other hand is solving a combinatorial search over patches and gradients may be non-informative until the right patches have been sampled consistently for a few iterations of training. The average can, however, become meaningless if the Top-K solver returned a different permutation of the K most salient patches for each perturbed input. Sorting of indices overcomes this issue and is an integral part of our "Top-K" operator.

For efficiency reasons we use hard top-K during inference. Hard top-K processes 3-10% more images per second because only a single Top-K operation has to be performed (instead of n perturbed repetitions) and patch extraction by

³Other distributions could be used. Please see [5] for details

slicing the tensor is more efficient than weighted combination with the indicator vectors. Furthermore, hard Top-K is deterministic which might be a desirable property at inference. However, using hard top-K at inference results in a train-test gap. To bridge this we linearly decay σ to zero during training. At $\sigma = 0$, no noise is added and the differentiable top-k operation is numerically identical to hard top-K. The gradients flowing into the scorer network vanish at $\sigma = 0$.

We also experimented with another differentiable Top-K formulation based on Sinkhorn operators [44]. We found the above perturbed-optimizer formulation to give superior results. Sinkhorn based experiments are therefore deferred to the appendix (Section D).

5. Experiments and results

Differentiable patch selection is a generic tool that can benefit a variety of computer vision problems. In this work, we focus on selecting a small number of patches from high resolution images for image classification.

5.1. Traffic Signs Recognition

Our first task is to recognize speed limits signs in large images. This is a key task to enable autonomous driving and is a natural fit for our method since the relevant pixels in the image are very localized and the model must rely on the high resolution images to read speed indications at significant distances. We use the Swedish traffic signs dataset [25], replicating the setup of [22] for their Attention Sampling (ATS) method. ATS uses a subset of the dataset consisting of 747 training images and 684 test images of dimension 960×1280 pixels, and the goal is to classify whether each image contains a limit sign of 50, 70 or 80 kilometers per hour or no speed limit. We apply the same data augmentation as [22], specifically a random translation and a random affine color scaling per image. For a fair comparison, we use ATS’s scorer: the scorer is a 4-layers CNN followed by a stride 8 max-pooling layer. We apply the scorer on a $3 \times$ downscaled version of the image. We obtain scores for $39 \times 52 = 2028$ candidate patches of dimensions 100×100



Figure 3: Scores computed at initialization (*top-right*) and extracted patches (*bottom-right*) displaying interpolation.

	K	test acc. [%]
Top-K (Ours)	5	91.7 ± 2.2
Top-K (Ours)	10	89.3 ± 1.4
ATS [22]	5	91.1 ± 0.2
ATS [22]	10	90.5 ± 0.8
ATS [†]	5	88.6 ± 1.1
ATS [*]	5	87.6 ± 1.4
CNN	-	63.0 ± 2.6

Table 1: Performance on the traffic signs dataset. ATS results are reported from [22]. We report the mean and the standard deviation of 9 runs except for CNN which uses 5 runs. ATS^{*}: Our reproduction of ATS following the hyper-parameters in [22], ATS[†]: the same but with learning rate drop at 12000 epochs.

and select $K \in \{5, 10\}$ of them. The feature network f_ϕ is the modified thin ResNet used by ATS. We use mean-pooling to aggregate per-patch representations.

Results in Table 1 show that we match the mean performance reported by ATS in their paper [22], albeit with a larger variance. We tried reproducing ATS results using their publicly released code, but were unsuccessful. The rows marked as ATS[†] and ATS^{*} show the best stable results we could obtain. These closely follow the hyper-parameters reported in their paper using an entropy regularization of 0.05 and training for $300k$ iterations. We compare the performance of our method and these reproduced ATS results in the box and whiskers plot of fig. 4. ATS can achieve higher accuracy with shorter training schedules by using a lower entropy regularization coefficient of 0.01. However, in this setting, approximately 20 – 25% of runs fail on the test set with accuracy less than approximately 60%.

We ablate against not having a patch selection procedure and directly apply the feature network CNN to the full image. Beside wasting computation over constant parts of images (e.g. the blue sky), this small ResNet completely overfits the training set: reaching 100% training accuracy while not achieving better accuracy than just predicting the majority class on test. It seems that the inductive bias introduced by patch extraction is crucial in this very low data regime.

We manually investigated the mistakes made by our model. Anecdotally the scoring network was able to extract the most relevant patches in most cases. Its only failure mode was a tendency to extract false-positive patches that exhibit the same colors as the traffic signs in question. Most miss-classifications are likely due to either the feature- or the aggregation-network. One could potentially further improve our results by pre-training the feature network.

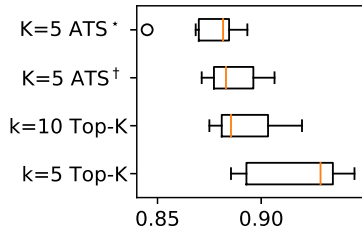


Figure 4: Performance on the traffic signs dataset across 9 repeats.



Figure 5: *Left*: Example of a 960×1280 image from the traffic signs dataset. *Top-right*: scores predicted for patch extraction. *Right*: Patches extracted by perturbed Top-K at inference.

5.2. Inter-patch reasoning

The work most closely related to ours, ATS, requires averaging representations obtained from sampled patches. We hypothesize that such mean aggregation limits expressivity required for modelling relationships between extracted patches. We investigate this using a synthetic dataset inspired by MegaMNIST [22], but which goes beyond scattered MNIST numbers on a Megapixel image to instead consist of billiard balls as presented in Figure 6. Each image contains four to eight randomly colored balls randomly placed on the table. Ball numbers are sampled uniformly from $\{1, \dots, 9\}$ and face the camera. We ensure that balls do not completely obstruction each other. We define the following classification task: report the higher of two numbers extracted from the leftmost and rightmost balls. All the other balls can be ignored. This task exhibits three interesting properties: (i) the information on the image is very localized (around the balls), (ii) downsampling the image severely degrades the readability of ball numbers, (iii) the task may be difficult to solve using a simple mean-pooling approach.

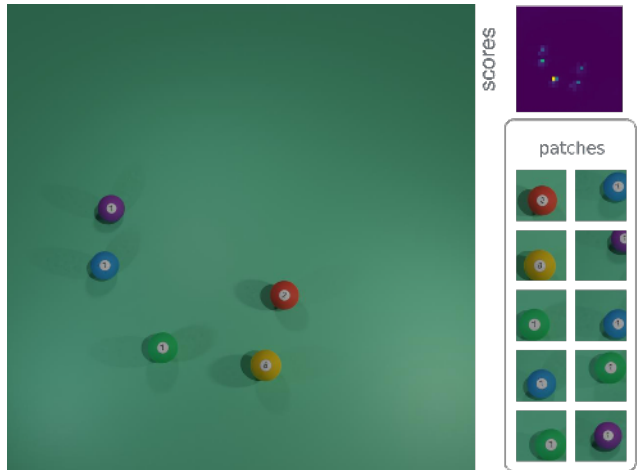


Figure 6: *Left*: example image from the billiard balls dataset. *Top-right*: scorer network output for patch selection. *Right*: Patches extracted at inference.

Using the Kubric⁴ software, we generated 20k images of size 1000×1000 . We split them into 8k samples for training, 2k for validation, and 10k for test. The generated dataset is available for download⁵. Code for our experiments is available on GitHub⁶.

We apply our model to this task. Architecture details are as follows. The scorer is a 4-layer CNN and processes the downsampled image of size 250×250 where the balls are visible but the numbers cannot be read. The feature network is ResNet18 [17]. We compare different aggregation schemes: mean-pooling, max-pooling, and a small Transformer [39]. The latter consists of 3 self-attention layers with 8 heads, taking the sequence of patch representations as input augmented with a learned additive positional encoding. We compare against ATS as well as a simple ResNet18 baseline. All the methods are trained using Adam [23] with decoupled weight decay of 10^{-4} [31] and tuned learning rate 10^{-4} . We repeated each experiment several times, but noted that not all runs were successful. For ATS, only 1 out of 4 runs were able to meaningfully solve the problem by reaching an accuracy of 53.25%, while the remaining three attempts performed no better than predicting the majority class. We also tried concatenating $[0, 1]$ normalized fixed positional encodings as two extra channels in the input (‘concat. position’) as the task requires the model to consider ball position. This affords a fairer comparison between our Transformer variant and the ATS and ResNet baselines. Positional encoding considerably im-

⁴<https://github.com/google-research/kubric>

⁵http://storage.googleapis.com/gresearch/ptokp_patch_selection/billiard.tar.xz

⁶https://github.com/google-research/google-research/tree/master/ptokp_patch_selection

	pooling	Test Acc. [%]
Top $K = 10$	transformer	93.8 ± 0.3
	max	90.8 ± 0.1
	mean	66.1 ± 0.7
ATS-10 [22]	mean	20.7 ± 0.00
	+ concat. position	42.44 ± 10.85
ResNet18		81.6 ± 0.35
	+ concat. position	83.8 ± 0.30
	+ downscale by 4	17.1 ± 0.81
Majority class		21.0

Table 2: Predicting the highest number between the rightmost and leftmost ball on the billiard balls dataset. Median and Median Absolute Deviation (MAD) for five runs (except for ATS where we have 4 runs and our method where we have 9 runs) of the Top-1 accuracy is reported.

proved baseline performances but were not as effective as the transformer. With concatenated position encodings, 3 out of 4 ATS runs meaningfully solved the problem while 1 run always predicted the majority class. This failure mode of majority class prediction also happened to 2 out of 9 runs of our model when using max-pooling aggregation. Both mean-pooling and transformer aggregation were relatively stable. We report robust estimates of performance (median and median absolute deviation) in Table 2.

As the results show, a standard CNN is able to solve this task, but the training time increase by 24% compared to differentiable Top-K with a transformer on top. This difference in running time would be even more pronounced in higher resolutions (see supplementary material). One may reduce CNN training time by downsampling the input image. As demonstrated in the results, this does not work, as the numbers become too small to be readable. ATS is not able to solve this task reliably, most likely because of mean pooling per-patch embeddings. Our own method also performs poorly when using mean-pooling, but when using transformers and max-pooling aggregation we are able to solve this task with high accuracy, outperforming all other methods. We compare the three aggregation schemes using a box-whiskers plot in fig. 7.

5.3. Fine-grained bird classification

Another way to extract salient regions of the image is to use a teacher-student approach to learn how to rank patches [47] (NTS-Net). We argue that NTS-Net’s ranking loss is essential only due to the non-differentiability of their patch selection method, and show here how their model can be simplified using our differentiable Top-K module. We explore this using the Caltech-UCSD Birds (CUB-200) dataset [40], a common dataset for fine-grained image clas-

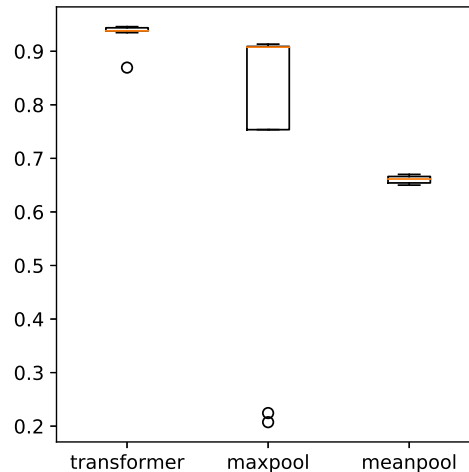


Figure 7: The effect of aggregation method in our model is ablated using 9 repetitions for each setting. Transformer and mean aggregation are relatively stable and transformer performs best.

Method	acc [%]
ResNet50	84.5
MG-CNN [41]	81.7
Bilinear-CNN [29]	84.1
Spatial Transformer [21]	84.1
FCAN [30]	84.5
PDFR [49]	84.5
RA-CNN [14]	85.3
HIHCA [7]	85.3
Boost-CNN [34]	85.6
DT-RAM [28]	86.0
MA-CNN [50]	86.5
NTS-Net $k = 2$ [47]	87.3
NTS-Net $k = 4$ [47]	87.5
NTS-Top-K $k = 2$ (ours) [Concat]	84.7 ± 0.8
NTS-Top-K $k = 4$ (ours) [Concat]	85.9 ± 0.4
NTS-Top-K $k = 2$ (ours) [Mean]	85.8 ± 0.3
NTS-Top-K $k = 4$ (ours) [Mean]	86.7 ± 0.4

Table 3: Results on CUB-200 dataset. Error bars are standard deviation over 5 repeats. ‘Concat’ and ‘mean’ refer to the aggregation method. NTS-Net uses ‘Concat’. Numbers for prior work were taken from Table 2 of [47].

sification. This dataset is a bird classification task with 11,788 images from 200 bird species.

We briefly describe the training setup of NTS-Net: The original image is resized and cropped to (448×448) , from which ResNet activations R_I are computed. A scorer net-

work processes these activations to score patches at multiple scales and aspect ratios. The scorer is trained using a ranking loss which we skip here because it is not required in our formulation. K top scoring (Top-K) patches are selected post greedy non-maxima suppression, resized to (224×224) , and encoded using the same ResNet backbone. Call these activations R_{P_1}, \dots, R_{P_K} . NTS-Net tries to predict the class label from R_I , each of R_{P_j} , and a concatenation of all these embeddings $[R_I, R_{P_1}, \dots, R_{P_K}]$. The model minimize a softmax cross entropy loss on each prediction head.

In this section, the emphasis is not on efficiently processing high resolution input but rather on being able to process salient parts of the object at a level of detail that is computationally infeasible when processing the entire input. We adapt our architecture to closely match NTS-Net, modulo four differences. First, NTS-Net uses greedy non-maximum suppression (NMS) in its patch selection module. Hard greedy NMS cannot be made differentiable using the perturbed maximum method as it does not correspond to the optimum of any linear program. Non-greedy-NMS, on the other hand, can be easily incorporated into the constraint set (eq. (5)). It would be computationally infeasible to solve the resulting linear program $n = 500$ times in each forward pass. One could learn a network to do NMS [18] to overcome this problem. We instead resort to making our scorer network more expressive using Squeeze-Excitation layers [19], so that it may learn to select non-overlapping patches if necessary. This way of doing feature modulation enables global communication between all spatial locations, allowing us to model a crude form of global context. Second, we use entropy regularization with a coefficient of 0.05 as in other experiments above ⁷. Third, we do not need and therefore eschew the ranking loss. Instead gradients can flow from the per-patch ResNet into the region proposal head through our differentiable patch selection module. Lastly, two aggregation methods are explored: (a) ‘concat’ matching that in NTS-Net and (b) ‘mean’ $R_I + \frac{(R_{P_1} + \dots + R_{P_K})}{K}$ where per-patch embeddings are averaged and added to the whole image embedding. ‘Concat’ results in a massive linear classifier over excessively high dimensional embeddings. ‘Mean’ ensures that aggregated embeddings are 2048 dimensional.

We tuned optimization hyper-parameters and regularization strength by splitting the training set into 5000 training images and 994 validation images. This was to avoid meta-overfitting on test data. We took the best hyper-parameter combination from this train-val split and then trained the model on the entire 5994 training images and tested on the official test set. Test results are reported in Table 3.

⁷Regularization is applied on the softmax of concatenated unnormalized scores. The final RPN outputs are normalized to lie between $[0, 1]$ as in the other experiments.

We note that our method performs slightly worse than the NTS-Net baseline. This may be due to the following: (a) Our model may be more prone to overfitting. It can select patches that directly optimize the training loss. We combat this using ‘mean’ aggregation which improves performance as shown in the last two rows in Table 3. (b) We report average performance across 5 seeds. Our best performing ‘mean’-aggregation model achieves 87.3%. (c) Despite using SE modules in the scorer network, our model still selects overlapping patches around the bird head. This likely also makes overfitting worse. Please see supplementary material for qualitative results.

6. Discussion and future work

We introduced a differentiable Top-K module for patch selection in large images. This enabled end-to-end learning of the patch selection module in a task dependent manner. We observed competitive performance on a task where information relevant for classification was very localized within the image. We significantly advanced the state-of-the-art in a problem requiring inter-patch relationship reasoning, as evidence by our use of a Transformer module on patch embeddings in the billiard balls dataset. Our method allows for arbitrary downstream neural processing of patches without using auxiliary losses for training the scorer network.

We are delighted by the above progress but note that there is still significant room for improvement and identify the following areas for future work. The patch selection problem is a chicken-and-egg problem. One can pick optimal patches by first knowing the contents of the image, but to efficiently know the contents of the image one should pick the right patches. This makes it inherently difficult to properly account for global context. Training patch selection models remains difficult. Some failure modes include predicting the majority class, or the scorer becoming very confident of the wrong patches either early in training or diverging to this behavior in the middle of training. More adaptive optimization might help improve learning stability. Patch selection lead to less overfitting on the Traffic street signs dataset. However, the same model overfits on the CUB-200 dataset. The scorer was able to identify patches associated with watermarks and background in order to memorize the training set. Fortunately when combined with the NTS-Net framework, overfitting was considerably reduced. Future work could address this by training on larger dataset or using self-supervised learning.

Acknowledgements

We thank Thomas Kipf, Francesco Locatello, Georg Heigold, Klaus Greff, Sindy Löwe, and Quentin Berthet for helpful discussions.

References

- [1] Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. *Perturbations, Optimization, and Statistics*, page 223, 2016. [4](#)
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. [2](#)
- [3] Baptiste Angles, Yuhe Jin, Simon Kornblith, Andrea Tagliasacchi, and Kwang Moo Yi. MIST: multiple instance spatial transformer network. In *CVPR*, 2021. [2](#)
- [4] Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015. [2](#)
- [5] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *NeurIPS*, 2020. [1, 2, 3, 4](#)
- [6] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. Fast differentiable sorting and ranking. *ICML*, 2020. [2](#)
- [7] S. Cai, W. Zuo, and L. Zhang. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *ICCV*, 2017. [7](#)
- [8] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L. Yuille. Attention to scale: Scale-aware semantic image segmentation. *CVPR*, 2016. [2](#)
- [9] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. On the relationship between self-attention and convolutional layers. In *ICLR*, 2020. [3](#)
- [10] Marco Cuturi, Olivier Teboul, and Jean-Philippe Vert. Differentiable ranking and sorting using optimal transport. In *NeurIPS*, 2019. [2](#)
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2020. [3](#)
- [12] Gamaleldin F. Elsayed, Simon Kornblith, and Quoc V. Le. Saccader: Improving accuracy of hard attention models for vision. *NeurIPS*, 2019. [2](#)
- [13] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. *NIPS*, 2016. [2](#)
- [14] J. Fu, H. Zheng, and T. Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, 2017. [2, 7](#)
- [15] R. Girshick. Fast r-cnn. In *ICCV*, 2015. [2](#)
- [16] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. [2](#)
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. [6](#)
- [18] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *CVPR*, pages 4507–4515, 2017. [8](#)
- [19] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. [8](#)
- [20] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, 2018. [2](#)
- [21] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *NIPS*, 2015. [2, 3, 7](#)
- [22] Angelos Katharopoulos and François Fleuret. Processing megapixel images with deep attention-sampling models. In *ICML*, 2019. [1, 2, 3, 5, 6, 7](#)
- [23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. [6](#)
- [24] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *CVPR*, 2016. [2](#)
- [25] Fredrik Larsson and Michael Felsberg. Using fourier descriptors and spatial models for traffic sign recognition. In *Image Analysis*, pages 238–249, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg. [5](#)
- [26] Kunpeng Li, Ziyang Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *CVPR*, 2018. [2](#)
- [27] Weijian Li, Viet-Duy Nguyen, Haofu Liao, Matt Wilder, Ke Cheng, and Jiebo Luo. Patch transformer for multi-tagging whole slide histopathology images. In *MICCAI*, 2019. [2](#)
- [28] Z. Li, Y. Yang, X. Liu, F. Zhou, S. Wen, and W. Xu. Dynamic computational time for visual attention. In *ICCV Workshops*, 2017. [2, 7](#)
- [29] T. Lin, A. RoyChowdhury, and S. Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015. [7](#)
- [30] Xiao Liu, Tian Xia, Jiang Wang, and Yuanqing Lin. Fully convolutional attention localization networks: Efficient attention localization for fine-grained recognition. *CoRR*, abs/1603.06765, 2016. [7](#)
- [31] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *ICLR*, 2019. [6](#)
- [32] S. Ma, J. Liu, and C. W. Chen. A-lamp: Adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment. In *CVPR*, 2017. [2](#)
- [33] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *NIPS*, 2014. [2](#)
- [34] Mohammad Moghimi, Mohammad Saberian, Jian Yang, Li-Jia Li, Nuno Vasconcelos, and Serge Belongie. Boosted convolutional neural networks. In *BMVC*, 2016. [7](#)
- [35] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149, 2017. [2, 3](#)
- [36] Jürgen Schmidhuber and Rudolf Huber. Learning to generate artificial fovea trajectories for target detection. *International Journal of Neural Systems*, 1991. [2](#)

- [37] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 2
- [38] J. R. Uijlings, K. E. Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *Int. J. Comput. Vision*, 2013. 2
- [39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 6
- [40] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 7
- [41] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *ICCV*, 2015. 7
- [42] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, 2017. 2
- [43] Sang Michael Xie and Stefano Ermon. Reparameterizable subset sampling via continuous relaxations. *IJCAI*, 2019. 2
- [44] Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable top-k operator with optimal transport. *NeurIPS*, 2020. 2, 5
- [45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 2
- [46] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, 2016. 2
- [47] Ze Yang, Tiange Luo, Dong Wang, Zhiqiang Hu, Jun Gao, and Liwei Wang. Learning to navigate for fine-grained classification. *ECCV*, 2018. 7
- [48] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 4
- [49] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian. Picking deep filter responses for fine-grained image recognition. In *CVPR*, 2016. 7
- [50] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *ICCV*, 2017. 2, 7