Twitter Thread by Davis Blalock

Davis Blalock
@davisblalock

"A Data-Based Perspective on Transfer Learning"

Different classes in a pretraining dataset can have different effects on downstream accuracy. And you can use this to your advantage. [1/9]

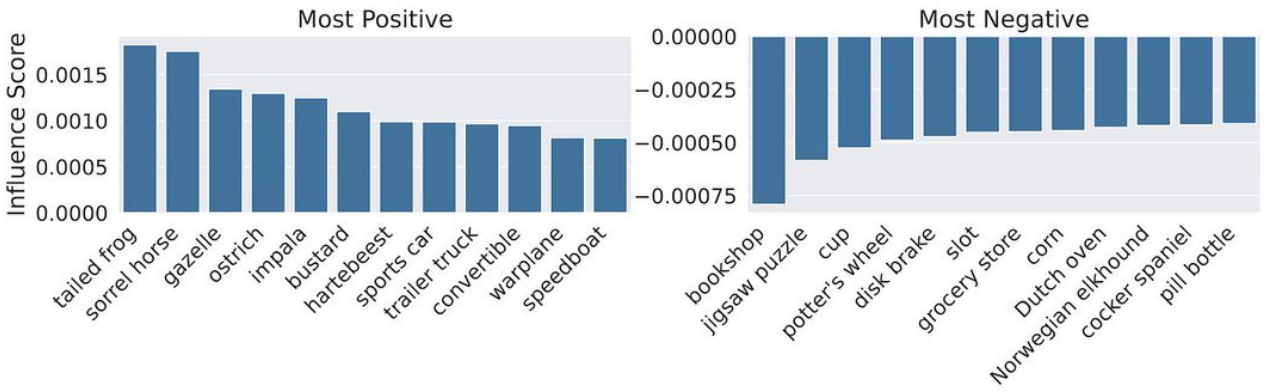


Figure 1: Most positive and negative ImageNet classes ordered based on their overall influence on the CIFAR-10 dataset. The top source classes (e.g., tailed frog and sorrel horse) turn out to be semantically relevant to the target classes (e.g., frog and horse).

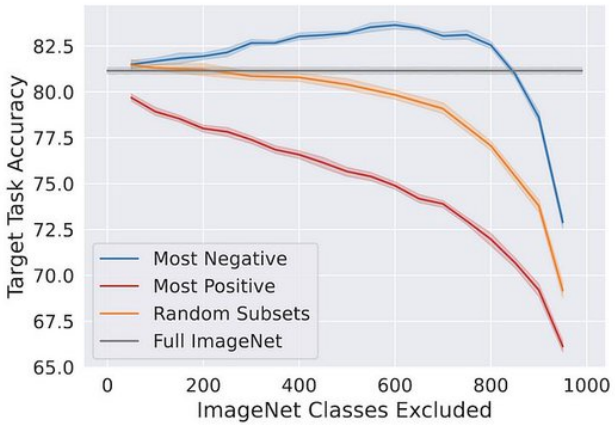
They assess these effects using a simple algorithm that trains different models on different subsets of the data and looks at both the class counts and the predictions for each model on each downstream sample. [2/9]

Algorithm 1 Estimation of source dataset class influences on transfer learning performance.

Require: Source dataset $\mathcal{S} = \cup_{k=1}^{K} \mathcal{C}_k$ (with K classes), a target dataset $\mathcal{T} = (t_1, t_2, \cdots, t_n)$, training algorithm $\mathcal{A}$, subset ratio α, and number of models m

Sample m random subsets $S_1, S_2, \cdots, S_m \subset \mathcal{S}$ of size $\alpha \cdot |\mathcal{S}|$:
for $i \in 1$ to m **do**
    Train model f_i by running algorithm $\mathcal{A}$ on S_i
end for
for $k \in 1$ to K **do**
    for $j \in 1$ to n **do**
        $$\text{Infl}[\mathcal{C}_k \to t_j] = \frac{\sum_{i=1}^{m} f_i(t_j; S_i) \mathbb{1}_{\mathcal{C}_k \subset S_i}}{\sum_{i=1}^{m} \mathbb{1}_{\mathcal{C}_k \subset S_i}} - \frac{\sum_{i=1}^{m} f_i(t_j; S_i) \mathbb{1}_{\mathcal{C}_k \not\subset S_i}}{\sum_{i=1}^{m} \mathbb{1}_{\mathcal{C}_k \not\subset S_i}}$$
    end for
end for
return $\text{Infl}[\mathcal{C}_k \to t_j]$, for all $j \in [n], k \in [K]$

Using their scoring function, you can intelligently remove subsets of classes from the pretraining dataset in order to significantly raise downstream accuracy. [3/9]



(a) CIFAR-10 results

Target Dataset	*Source Dataset* Full ImageNet	Removing Bottom Infl.	Hand-picked
AIRCRAFT	36.08 ± 1.07	$\mathbf{36.88 \pm 0.74}$	N/A
BIRDSNAP	38.42 ± 0.40	$\mathbf{39.19 \pm 0.38}$	26.74 ± 0.31
CALTECH101	86.69 ± 0.79	$\mathbf{87.03 \pm 0.30}$	82.28 ± 0.40
CALTECH256	74.97 ± 0.27	$\mathbf{75.24 \pm 0.21}$	67.42 ± 0.39
CARS	39.55 ± 0.32	$\mathbf{40.59 \pm 0.57}$	21.71 ± 0.40
CIFAR10	81.16 ± 0.30	$\mathbf{83.64 \pm 0.40}$	75.53 ± 0.42
CIFAR100	59.37 ± 0.58	$\mathbf{61.46 \pm 0.59}$	55.21 ± 0.52
FLOWERS	$\mathbf{82.92 \pm 0.52}$	82.89 ± 0.48	N/A
FOOD	56.19 ± 0.14	$\mathbf{56.85 \pm 0.27}$	39.36 ± 0.39
PETS	83.41 ± 0.55	$\mathbf{87.59 \pm 0.24}$	87.16 ± 0.24
SUN397	50.15 ± 0.23	$\mathbf{51.34 \pm 0.29}$	N/A

(b) Summary of 11 target tasks

Figure 2: Target task accuracies after removing the K most positively or negatively influential ImageNet classes from the source dataset. Mean/std are reported over 10 runs. **(a)** Results with CIFAR-10 as the target task after removing different numbers of classes from the source dataset. We also include baselines of using the full ImageNet dataset and removing random classes. One can note that, by removing negatively influential source classes, we can obtain a test accuracy that is 2.5% larger than what using the entire ImageNet dataset would yield. Results for other target tasks can be found in Appendix C. **(b)** Peak performances when removing the most negatively influential source classes across a range of other target tasks. We compare against using the full ImageNet dataset or a relevant subset of classes (hand-picked, see Appendix A for details).

Another use of their method is identifying more granular subpopulations than what a downstream task has annotated. E.g., you can find which CIFAR-10 images look most like ostriches even though CIFAR-10 only has the label “bird”. [4/9]

Figure 4: **Projecting source labels onto the target dataset.** The CIFAR-10 images that were most positively influenced by the ImageNet classes “ostrich” and “fire engine.” We find that these images look similar to the corresponding images in the source dataset.

You can also use a similar idea to understand model failure modes or identify data leakage. [5/9]

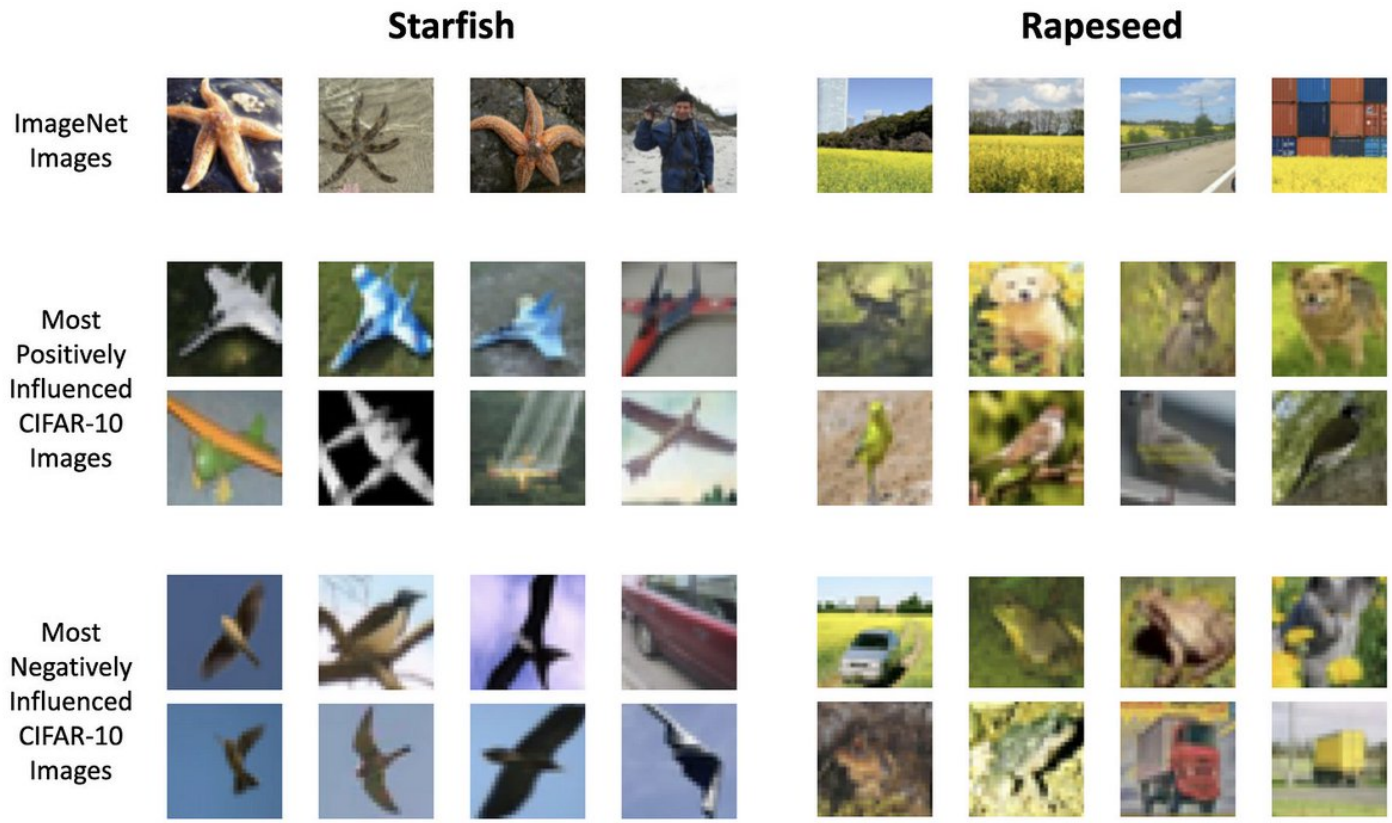


Figure 5: The CIFAR-10 images that were most positively (or negatively) influenced by the ImageNet classes "starfish" and "rapeseed." CIFAR-10 images that are highly influenced by the "starfish" class have similar shapes, while those influenced by "rapeseed" class have yellow-green colors.

And last but not least, you can use it to understand helpful/harmful samples in your pretraining dataset. [6/9]

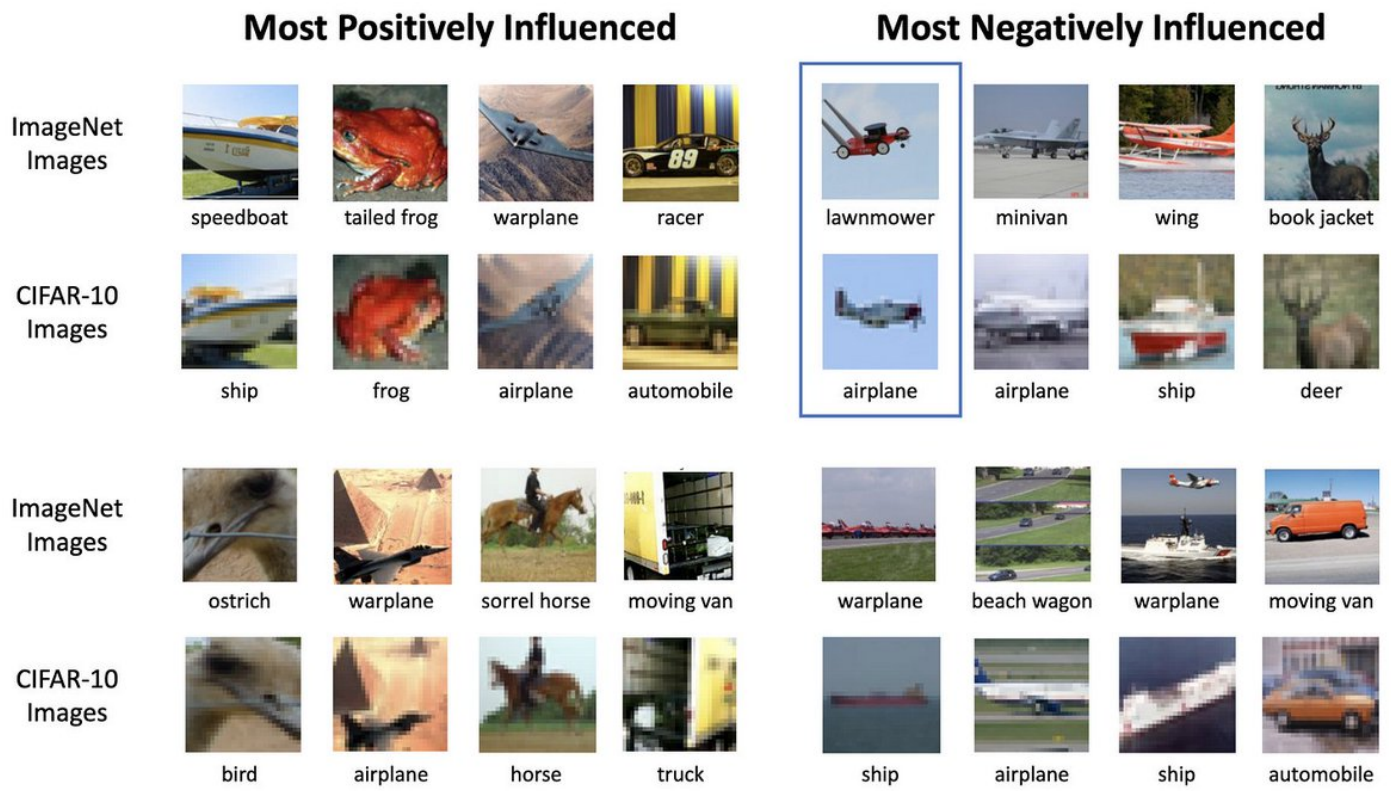


Figure 7: ImageNet training images with highest positive (**left**) or negative (**right**) example-wise (average) influences on CIFAR-10 test images. We find that ImageNet images that are highly positively influential often correspond to data leakage, while ImageNet images that are highly negatively influential are often either mislabeled, ambiguous, or otherwise misleading. For example, the presence of a flying lawnmower in the ImageNet dataset hurts the downstream performance on a similarly shaped airplane (boxed).

Overall their algorithm seems like a great tool to have in the toolbox. [7/9]

Paper: https://t.co/CKg0nxmSxE

If you like this paper, consider RTing this (or another!) thread to publicize the authors' work, or following the authors: @saachi_jain_ @hadisalmanX @Alaa_Khaddaj… [8/9]

@saachi_jain_ @hadisalmanX @Alaa_Khaddaj …@RICEric22 @ssung_mminn @aleks_madry

For more paper summaries, you might like following @mosaicml, me, or my newsletter: https://t.co/5BMBC84xY8

As always, comments and corrections welcome! [9/9] https://t.co/8VRLAGmrfQ