Appendix

Broader Impact

Our work attacks both the labeling and computational costs of machine learning and will hopefully make machine learning much more affordable. Instead of being limited to a small number of large teams and organizations with the budget to label data and the computational resources to train on it, SEALS dramatically reduces the barrier to machine learning, enabling small teams or individuals to build accurate classifiers. SEALS does, however, introduce another system component, a similarity search index, which adds some additional engineering complexity to build, tune, and maintain. Fortunately, several highly optimized implementations like Annoy [2] and Faiss [3] work reasonably well out of the box. There is a risk that poor embeddings will lead to disjointed components for a given concept. This failure mode may prevent SEALS from reaching all fragments of a concept or take a longer time to do so, as mentioned in Section 6. However, active learning and search methods often involve humans in the loop, which could detect biases and correct them by adding more examples.

Proof of SEALS under idealized conditions

To begin, we introduce the mathematical setting. Assume the input space is a convex set $\mathcal{X} \subset \mathbb{R}^d$ and that the optimal linear classifier $w_* \in \mathbb{R}^d$ satisfies $\|w_*\|_2 = 1$. We assume the *homogenous setting* where $\mathcal{H}_* = \{x \in \mathbb{R}^d : w_*^\top x = 0\}$ is the hyperplane defining the optimal classification. For each $x \in \mathcal{X}$, let y_x denote its associated label and assume that $y_x = 1$ if $w_*^\top x \geq 0$ and $y_x = -1$ if $w_*^\top x < 0$. Define $\mathcal{X}^+ = \{x \in \mathcal{X} : w_*^\top x \geq 0\}$ and $\mathcal{X}^- = \{x \in \mathcal{X} : w_*^\top x < 0\}$. Let $\delta > 0$ and let $\mathcal{G} = (\mathcal{X}, E)$ be a nearest-neighbor graph where we assume that for each $x, x' \in \mathcal{X}$, if $\|x - x'\|_2 \leq \delta$, then $(x, x') \in E$.

Our analysis makes two key assumptions. First, we assume that the classes are linearly separable. Since the SEALS algorithm uses feature embedding, often extracted from a deep neural network, the classes are likely to be almost linearly separable in many applications. Second, we assume a *membership query model* where the algorithm can query any point belonging to the input space $\mathcal{X}$. Since SEALS is typically applied to datasets with billions of examples, this query model is a reasonable approximation of practice. It should be possible to extend our analysis to the pool-based active learning setting. Suppose there are enough points so that for every possible direction there is at least one point in every δ-nearest neighborhood with a component of size $c\delta$, for some $0 < c < 1$, in that direction. Then we believe that a bound like that in Theorem 1 should hold by replacing δ with $c\delta$. Relaxing these two assumptions is left to future theory work.

The main goal of our theory is to quantify the effect of the nearest neighbor restriction. Our analysis considers the modified SEALS procedure described in Algorithm 3. It differs from original SEALS in some minor ways that make it more amenable to analysis, but crucially it too is based on nearest neighbor graph search. First, we introduce some notation. We let $\mathcal{S}_r \subset \mathcal{X} \times \{-1, 1\}$ consist of the examples and their labels queried until round r. We let $\mathcal{S}_r^1 = \{x \in \mathcal{S}_1 : (x, 1) \in \mathcal{S}_r\}$ denote the positive examples queried until round r and $\mathcal{S}_r^{-1} = \{x \in \mathcal{S}_1 : (x, -1) \in \mathcal{S}_r\}$ the negative examples queried until round r. Let $A, B \subset \mathbb{R}^d$. The subroutine MaxMarginSeparator(A, B) finds a maximum margin separator of A and B and returns the hyperplane H and margin γ: $(H, \gamma) \longleftarrow$ MaxMarginSeparator(A, B). This is a support vector machine.

Now, we present Algorithm 3. We suppose that the algorithm has an initial set of seed points $\{x_{1,0}, \ldots, x_{d-1,0}\}$, with which it initiates $d-1$ nearest neighbor searches. We note that the algorithm could perform n nearest neighbor searches and the analysis would still go through provided that $d - 1 \leq n = O(d)$, and that the initial set of labeled points $\mathcal{S}_r$ may be larger than $d - 1$. At each round r, Algorithm 3 queries one unlabeled neighbor from each set $C_{i,r}$, $i = 1, \ldots, d-1$ (at $r = 0$ these sets are the seed points themselves). The decision rule is as follows: first, for each $x \in C_{i,r}$, the algorithm computes a max-margin separating hyperplane H_x separating x from the examples with opposing labels. Second, the algorithm selects the example with the smallest margin $\bar{x}_{i,r}$ and queries a neighbor of $\bar{x}_{i,r}$ that is closest to $H_{\bar{x}_{i,r}}$. This is similar to using *MaxEnt* uncertainty sampling

[2] `https://github.com/spotify/annoy`

[3] `https://github.com/facebookresearch/faiss`

in SEALS. In Algorithm 3, $\arg\min_{x':(\bar{x}_{i,r},x')\in E} \operatorname{dist}(x', H_{\bar{x}_{i,r},r})$ may contain several examples if $\operatorname{dist}(\bar{x}_{i,r}, H_{\bar{x}_{i,r},r}) \leq \delta$. In this case, we tiebreak by letting $\tilde{x}_{i,r}$ be the projection of $\bar{x}_{i,r}$ onto $H_{\bar{x}_{i,r},r}$.

Algorithm 3 Modified SEALS

Input: seed labeled examples $\mathcal{S}_1 \subset \mathcal{X} \times \{-1, 1\}$
$r = 1$
Initialize the clusters $C_{i,r} = \{x_{i,0}\}$ for $i = 1, \ldots, d-1$
for $r = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, d-1$ **do**
    $(H_{x,r}, \gamma_{x,r}) = \text{MaxMarginSeparator}(\mathcal{S}_r^{-y_x}, x)$ for all $x \in C_{i,r}$
    Let $\bar{x}_{i,r} \in \arg\min_{x \in C_{i,r}} \gamma_{x,r}$ and $\widetilde{x}_{i,r} \in \arg\min_{x':(\bar{x}_{i,r},x')\in E} \operatorname{dist}(x', H_{\bar{x}_{i,r},r})$
    Query $\widetilde{x}_{i,r}$
    Update $\mathcal{S}_{r+1} = \mathcal{S}_r \cup \{(\widetilde{x}_{i,r}, y_{\bar{x}_{i,r}})\}$ and $C_{i,r+1} = C_{i,r} \cup \{\widetilde{x}_{i,r}\}$
  end for
  Fit a homogenous max-margin separator with normal vector $\widehat{w}_{r+1}$ to $\mathcal{S}_{r+1}$
end for

We actually restate the Theorem 1 a bit more formally in Theorem 2, below.

Theorem 2. *Let $\epsilon > 0$. Let $x_{1,0}, \ldots, x_{d-1,0}$ denote the seed points. Define $\gamma_i = \operatorname{dist}(x_{i,0}, \texttt{conv}(\mathcal{S}_1^{-y_{x_{i,0}}}))$ for $i \in [d-1]$, where $\texttt{conv}(\mathcal{S}_1^{-y_{x_{i,0}}})$ is the convex hull of the points $\mathcal{S}_1^{-y_{x_{i,0}}}$. Then, after Algorithm 3 makes $\max_{i\in[d-1]} d(\frac{\gamma_i}{\delta} + 2\log(\frac{2d\delta}{\epsilon \min(\sigma,1)}))$ queries, $\|\widehat{w}_r - w_*\| \leq \epsilon$.*

The constant σ is a measure of the diversity of the initial seed examples, which we now define. For $i \in [d-1]$, define the set

$$\mathcal{Z}_i = \{z \in \mathbb{R}^d : \|z - x_{i,0}\| \leq \gamma_i + 2\delta + \epsilon \text{ and } \operatorname{dist}(z, \{x \in \mathbb{R}^d : x^\top w_* = 0\}) \leq \epsilon\}.$$

Define

$$\sigma := \min_{z_i \in \mathcal{Z}_i : \forall i \in [d-1]} \sigma_{d-1}([z_1 \ldots z_{d-1}]).$$

where σ_{d-1} denotes the $(d-1)$th singular value of the matrix $[z_1 \ldots z_{d-1}]$.

Here we give a simple example where $\sigma = \Omega(1)$ so as to provide intuition, although there are a wide variety of such cases.

Example 1. *Let $\epsilon \in (0, 1)$. Let $\mathcal{X} = \mathbb{R}^d$, $w^* = e_1$, $\delta = 1/2$. Let $M \geq 6\sqrt{d-1}$. Suppose the seed examples are $x_{i,0} = e_1 + Me_{i+1}$ for $i \in [d-1]$ and suppose the algorithm is given additional examples $v_i = -e_1 + Me_{i+1} \in \mathcal{S}_0^{-1}$ for $i \in [d-1]$. Then, $\sigma \geq 1$.*

Proof of Example 1. Note that $\gamma_i = 2$ for all $i \in [d-1]$. Define the matrix

$$Z = \begin{pmatrix} z_1^\top \\ \vdots \\ z_{d-1}^\top \end{pmatrix}$$

such that $\|z_i - x_{i,0}\| \leq \gamma_i + 2\delta + \epsilon, \forall i \in [d-1]$. We may write $z_i = x_{i,0} + v_i$ where $\|v_i\| \leq \gamma_i + 2\delta + \epsilon$. Courant-Fisher's min-max theorem implies that $s_{d-1}(Z) = \max_{\dim E = d-1} \min_{u \in \text{span}(E):\|u\|=1} \|Zu\|$ where $E \subset \mathbb{R}^d$. Therefore, taking $E = \{e_2, \ldots, e_d\}$, it suffices to lower bound $\|Zu\|$ for any $u \in \text{span}(e_2, \ldots, e_d)$ with $\|u\| = 1$. Since $\|u\| = 1$, there exists $j \in \{2, \ldots, d\}$ such that $|u_j| \geq \frac{1}{\sqrt{d-1}}$. Suppose wlog that $u_j \geq \frac{1}{\sqrt{d-1}}$ (the other case is

similar). Then, by Cauchy-Schwarz,

$$\begin{aligned}\|Zu\| &\geq \max_{i\in[d-1]} |z_i^\top u| \\ &\geq (e_1 + Me_j + v_{j-1})^\top u \\ &\geq M\frac{1}{\sqrt{d}} - 1 - \gamma_i - 2\delta - \epsilon \\ &\geq \frac{M}{\sqrt{d}} - 5 \\ &\geq 1.\end{aligned}$$

□

Now, we turn to the proof of Theorem 2. In the interest of using more compact notation, we define for all $i \in [d-1]$ and $r \in \mathbb{N}$

$$\rho_{i,r} := \operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-y_{\bar{x}_{i,r}}})).$$

In words, $\rho_{i,r}$ is the distance of the example queried in nearest neighbor search i and at round r, $\bar{x}_{i,r}$, to the convex hull of $\mathcal{S}_r^{-y_{\bar{x}_{i,r}}}$, the examples with opposite labels from $\bar{x}_{i,r}$.

Proof of Theorem 2. **Step 1: Bounding the number of queries to find points near the decision boundary.** Define $\bar{\epsilon} = \frac{\min(\sigma,1)\epsilon^2}{2\sqrt{d}}$, where σ is defined as in the Theorem statement. We assume $\sigma > 0$ for the remainder of the proof. Let $C_{i,r} = \{x_{i,0}, x_{i,1}, \ldots, x_{i,r}\}$ where $x_{i,l}$ is the queried example in the lth round. We show that for all $r \geq \max_{i\in[d-1]} \frac{\gamma_i}{\delta} + \log(\frac{2\delta}{\bar{\epsilon}})$, for all $i \in [d-1]$, $\rho_{i,r} \leq \bar{\epsilon}$.

Fix $i \in [d-1]$. Define

$$E_r = \{\text{at round } r, \rho_{i,r} \leq \bar{\epsilon}\}.$$

We have that

$$\begin{aligned}\sum_{r\in\mathbb{N}} \mathbf{1}\{E_r^c\} = \sum_{r\in\mathbb{N}} &\mathbf{1}\{E_r^c \cap \{\rho_{i,r} \geq 2\delta\}\} \\ &+ \mathbf{1}\{E_r^c \cap \{\rho_{i,r} < 2\delta\}\}\end{aligned}$$

If $\rho_{i,r} \geq 2\delta$, we have by Lemma 1 that

$$\rho_{i,r+1} \leq \rho_{i,r} - \delta.$$

This implies that

$$\sum_{r\in\mathbb{N}} \mathbf{1}\{E_r^c \cap \{\rho_{i,r} \geq 2\delta\}\} \leq \frac{\gamma_i}{\delta}.$$

Now, suppose that $\rho_{i,r} < 2\delta$. Then, by Lemma 1, we have that

$$\rho_{i,r+1} \leq \frac{\rho_{i,r}}{2}.$$

Unrolling this recurrence implies that

$$\sum_{r\in\mathbb{N}} \mathbf{1}\{E_r^c \cap \{\rho_{i,r} < 2\delta\}\} \leq \log(\frac{2\delta}{\bar{\epsilon}}).$$

Putting it altogether, we have that

$$\sum_{r\in\mathbb{N}} \mathbf{1}\{E_r^c\} \leq \frac{\gamma_i}{\delta} + \log(\frac{2\delta}{\bar{\epsilon}}).$$

This implies that after $\max_{i\in[d-1]} \frac{\gamma_i}{\delta} + \log(\frac{2\delta}{\bar{\epsilon}})$ queries, we have that for all $i \in [d-1]$, $\rho_{i,r} := \operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-y_{x_{i,r}}})) \leq \bar{\epsilon}$.

Step 2: Showing $\|x_{i,0} - x_{i,r}\| \leq \gamma_i + 2\delta$ for all $i \in [d-1]$. Fix $i \in [d-1]$. Let $C_{i,r} = \{x_{i,0}, x_{i,1}, \ldots, x_{i,r}\}$ where $x_{i,l}$ is the queried example in the lth round. Note there exists a path of length at most r in the nearest neighbor graph on the nodes in $C_{i,r}$ between $x_{i,0}$ and $x_{i,r}$. In the worst case, this path consists of $x_{i,0}, x_{i,2}, \ldots, x_{i,r}$ with $x_{i,s}$ being the child of $x_{i,s-1}$ and thus we suppose that this is the case wlog.

By the argument in step 1 of the proof, we have that after at most $\bar{k} = \frac{\gamma_i}{\delta}$ rounds,

$$\operatorname{dist}(x_{i,\bar{k}}, \texttt{conv}(\mathcal{S}_r^{-y_{x_{i,\bar{k}}}})) \leq 2\delta.$$

For $s \leq \bar{k}$, we have that $\|x_{i,s} - x_{i,s-1}\| \leq \delta$ by definition of the nearest neighbor graph. Now, consider $s \geq \bar{k}$. By Lemma 2, we have that $\|x_{i,s+1} - x_{i,s}\| \leq \frac{\|x_{i,s} - x_{i,s-1}\|}{2}$ and $\left\|x_{i,\bar{k}+1} - x_{i,\bar{k}}\right\| \leq \delta$. Therefore,

$$\begin{aligned}
\|x_{i,r} - x_{i,0}\| &= \left\| \sum_{s=1}^{r-1} x_{i,s+1} - x_{i,s} \right\| \\
&\leq \sum_{s=1}^{r-1} \|x_{i,s+1} - x_{i,s}\| \\
&\leq \bar{k} \max_{s \leq \bar{k}} \|x_{i,s+1} - x_{i,s}\| + \sum_{s=\bar{k}+1}^{r-1} \|x_{i,s+1} - x_{i,s}\| \\
&\leq \frac{\gamma_i}{\delta}\delta + \delta \sum_{s=\bar{k}+1}^{r-1} \frac{1}{2^{-s}} \\
&\leq \gamma_i + 2\delta.
\end{aligned}$$

Step 3: for all $i \in [d-1]$, there exists $z_i \in \mathcal{Z}_i$ such that $\widehat{w}_r^\top z_i = 0$. We have that $\rho_{i,r} \leq \bar{\epsilon}$ for all $i \in [d-1]$ and $\|x_{i,0} - x_{i,r}\| \leq \gamma_i + 2\delta$ for all $i \in [d-1]$. Now, we show that there exists $z_i \in \mathcal{Z}_i$ such that $\widehat{w}_r^\top z_i = 0$. Fix $i \in [d-1]$. Suppose $\widehat{w}_r^\top \bar{x}_{i,r} > 0$ (the other case is similar). Then, $\rho_{i,r} \leq \bar{\epsilon}$ implies that there exists $\bar{x} \in \texttt{conv}(\mathcal{S}_r^{-1})$ such that $\|\bar{x}_{i,r} - \bar{x}\| \leq \bar{\epsilon}$. Then, since $\widehat{w}_r$ separates $\mathcal{S}_r^{-1}$ and $\mathcal{S}_r^1$, $\widehat{w}_r^\top \bar{x} \leq 0$. Now, there exists $z_i \in \texttt{conv}(\bar{x}_{i,r}, \bar{x})$ such that $z_i^\top \widehat{w}_r = 0$. By the triangle inequality, we have that $\|x_{i,0} - z_i\| \leq \gamma_i + 2\delta + \epsilon$, and $\operatorname{dist}(z_i, \{x : w_*^\top x = 0\}) \leq \bar{\epsilon} \leq \epsilon$, thus, $z_i \in \mathcal{Z}_i$ for all $i \in [d-1]$, completing this step.

Step 4: Pinning down w_*. Since $\sigma > 0$, and each $z_i \in \mathcal{Z}_i$, we have that $z_1, \ldots, z_{d-1}$ are linearly independent. Then, we can write $w_* = \sum_{i=1}^{d-1} \beta_i z_i + \beta_d \widehat{w}_r$ for $\beta_1, \ldots, \beta_d \in \mathbb{R}$ where we used that $\widehat{w}_r$ is orthogonal to $z_1, \ldots, z_{d-1}$ by construction of $z_1, \ldots, z_{d-1}$. Note that $\widehat{w}_r^\top w_* = \widehat{w}_r^\top (\sum_{i=1}^{d-1} \beta_i z_i + \beta_d \widehat{w}_r) = \beta_d$. Let Pw_* denote the projection of w_* onto $\operatorname{span}(z_1, \ldots, z_{d-1})$. Defining the matrix $Z = [z_1\ z_2\ \ldots\ z_{d-1}]$ and $\beta = (\beta_1\ \ldots, \beta_{d-1})^\top$, we can write $Pw_* = Z\beta$. Let $Z = U\Sigma V^\top$ denote the SVD of Z and $Z^\dagger = V\Sigma^\dagger U^\top$ the pseudoinverse. Note that $Z^\dagger P w_* = \beta$. Then,

$$\|\beta\| = \left\|Z^\dagger P w_*\right\| \leq \max_{i=1,2,\ldots,d-1} \sigma_i(Z^\dagger) = \frac{1}{\sigma_{d-1}(Z)}. \tag{1}$$

We note that $\operatorname{dist}(z_i, \{x : w_*^\top x = 0\}) \leq \bar{\epsilon}$ implies that $|w_*^\top z_i| \leq \bar{\epsilon}$ for $i = 1, 2, \ldots, d-1$. Then, we have that

$$\begin{aligned}
1 &= w_*^\top w_* \\
&= \sum_{i=1}^{d-1} \beta_i w_*^\top x_i + \beta_d w_*^\top \widehat{w}_r \\
&= \sum_{i=1}^{d-1} \beta_i w_*^\top x_i + (w_*^\top \widehat{w}_r)^2 && (2) \\
&\leq \|\beta\|_1 \bar{\epsilon} + (w_*^\top \widehat{w}_r)^2 && (3) \\
&\leq \sqrt{d} \|\beta\|_2 \bar{\epsilon} + (w_*^\top \widehat{w}_r)^2 && (4) \\
&\leq \frac{\sqrt{d}\bar{\epsilon}}{\sigma_{d-1}([z_1 \ldots z_{d-1}])} + (w_*^\top \widehat{w}_r)^2 && (5) \\
&\leq \frac{\sqrt{d}\bar{\epsilon}}{\sigma} + (w_*^\top \widehat{w}_r)^2 && (6)
\end{aligned}$$

where equation 2 follows by plugging in the previously derived $\beta_d = w_*^\top \widehat{w}_r$, equation 3 follows by Holder's inequality, equation 4 follows by $\|\cdot\|_1 \leq \sqrt{d} \|\cdot\|_2$, equation 5 follows by equation 1, and equation 6 follows by the definition of σ. Rearranging, we have that

$$\begin{aligned}
w_*^\top \widehat{w}_r &= |w_*^\top \widehat{w}_r| && (7) \\
&\geq \sqrt{1 - \frac{\sqrt{d}\bar{\epsilon}}{\sigma}} \\
&\geq 1 - \frac{\sqrt{d}\bar{\epsilon}}{\sigma} && (8) \\
&\geq 1 - \frac{\epsilon^2}{2} && (9)
\end{aligned}$$

where equation 7 follows by the definition of $\widehat{w}_r$, in equation 8 we use that fact that $1 - \frac{\sqrt{d}\bar{\epsilon}}{\sigma} \leq 1$ and in equation 9 we use the definition of $\bar{\epsilon}$. Now, we have that

$$\|\widehat{w}_r - w_*\|^2 = 2(1 - \widehat{w}_r^\top w_*) \leq \epsilon^2,$$

proving the result.

□

In Lemma 1, we show that at each round $\bar{x}_{i,r}$ moves closer to the labeled examples of the opposite class. The main idea behind the proof is that the algorithm can always choose a point $\tilde{x}_{i,r}$ in the direction orthogonal to the hyperplane $H_{\bar{x}_{i,r},r}$, thus guaranteeing a reduction in $\rho_{i,r}$. Early in the execution of Algorithm 3, the nearest neighbor graph constrains which points are chosen, leading to a reduction in $\rho_{i,r}$ of δ. However, once $\bar{x}_{i,r}$ is close enough to the labeled examples of the opposite class, precisely once $\rho_{i,r} < 2\delta$, the algorithm begins selecting points that lie on $H_{\bar{x}_{i,r},r}$ at each round, halving $\rho_{i,r}$ at each round.

Lemma 1. *Fix $i \in [d-1]$. Fix $r \in \mathbb{N}$.*

1. *If $\rho_{i,r} \geq 2\delta$, then*

$$\rho_{i,r+1} \leq \rho_{i,r} - \delta.$$

2. *If $\rho_{i,r} < 2\delta$, then*

$$\rho_{i,r+1} \leq \frac{\rho_{i,r}}{2}.$$

Proof. 1. Let $\bar{x}_{i,r} = \min_{x \in C_{i,r}} \gamma_{x,r}$ as defined in the algorithm. Wlog, suppose that $\bar{x}_{i,r} \in \mathcal{X}^+$. By Lemma 2, there exists $z \in \texttt{conv}(\mathcal{S}_r^{-1})$ such that $\tilde{x}_{i,r} = \bar{x}_{i,r} + \alpha \frac{(z - \bar{x}_{i,r})}{\|z - \bar{x}_{i,r}\|}$ for some $\alpha \leq \delta$.

Let $\beta = \frac{\alpha}{\|z-\bar{x}_{i,r}\|}$. Then,

$$\begin{aligned}\|\tilde{x}_{i,r} - z\| &= \|\bar{x}_{i,r} + \beta(z - \bar{x}_{i,r}) - z\| \\ &= \|(1-\beta)\bar{x}_{i,r} - z\| \\ &= (1-\beta)\,\|\bar{x}_{i,r} - z\| \\ &= \|\bar{x}_{i,r} - z\| - \alpha.\end{aligned}$$

If $\operatorname{dist}(\bar{x}_{i,r}, \mathcal{S}_r^{-1}) \geq 2\delta$, Lemma 2 implies that $\alpha = \delta$ and we have that

$$\|\tilde{x}_{i,r} - z\| = \|\bar{x}_{i,r} - z\| - \delta = \operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-1})) - \delta$$

Thus, if $\tilde{x}_{i,r} \in \mathcal{X}^+$, using the definition of $\bar{x}_{i,r+1}$, we have that

$$\begin{aligned}\rho_{i,r+1} &= \operatorname{dist}(\bar{x}_{i,r+1}, \texttt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r+1}}})) \\ &\leq \operatorname{dist}(\tilde{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-1})) \\ &\leq \operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-1})) - \delta \\ &= \rho_{i,r} - \delta.\end{aligned}$$

On the other hand, if $\tilde{x}_{i,r} \in \mathcal{X}^-$, we have that

$$\begin{aligned}\rho_{i,r+1} &= \operatorname{dist}(\bar{x}_{i,r+1}, \texttt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r+1}}})) \\ &\leq \operatorname{dist}(\tilde{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{1})) \\ &\leq \delta \\ &\leq \operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-1})) - \delta\end{aligned}$$

which also shows the claim.

2. Now, suppose that $\operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_{r+1}^{-1})) < 2\delta$. Then, by Lemma 2, we have that $\alpha = \frac{\|z-\bar{x}_{i,r+1}\|}{2} < \delta$, implying that

$$\begin{aligned}\rho_{i,r+1} &= \operatorname{dist}(\bar{x}_{i,r+1}, \texttt{conv}(\mathcal{S}_r^{-y_{\bar{x}_{i,r+1}}}) \\ &\leq \|\tilde{x}_{i,r} - z\| \\ &= \|\tilde{x}_{i,r} - \bar{x}_{i,r}\| \\ &= \|\bar{x}_{i,r} - z\|\,/2 \\ &= \frac{\operatorname{dist}(\bar{x}_{i,r}, \texttt{conv}(\mathcal{S}_r^{-1}))}{2} \\ &= \frac{\rho_{i,r}}{2}\end{aligned}$$

and therefore the result.

□

Lemma 2 characterizes the example, $\tilde{x}_{i,r}$, queried by Algorithm 3. It shows that $\tilde{x}_{i,r}$ always belongs to a line segment connecting $\bar{x}_{i,r}$ to some point, z, in the convex hull of labeled points of the opposite class, $\texttt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r}}})$. If $\rho_{i,r} \geq 2\delta$, then $\tilde{x}_{i,r}$ moves δ along this line segment towards z and if $\rho_{i,r} < 2\delta$, $\tilde{x}_{i,r}$ is the midpoint of this line segment.

Lemma 2. *Fix round $r \in \mathbb{N}$. There exists $z \in \texttt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r}}})$ such that the following holds. If $\rho_{i,r} \geq 2\delta$, then $\tilde{x}_{i,r} = \bar{x}_{i,r} + \delta\frac{(z-\bar{x}_{i,r})}{\|z-\bar{x}_{i,r}\|}$. If $\rho_{i,r} < 2\delta$, then $\tilde{x}_{i,r} = \frac{\bar{x}_{i,r}+z}{2}$.*

Proof. **Step 1: A formula for the max-margin separator.**

Without loss of generality, suppose that $\bar{x}_{i,r} \in \mathcal{X}^+$ (the other case is similar). Let $\bar{w} \in \mathbb{R}^d$, $\bar{b} \in \mathbb{R}$, and $\bar{t} \in \mathbb{R}$ denote the optimal solutions of the optimization problem

$$\begin{aligned} \max_{w \in \mathbb{R}^d, b \in \mathbb{R}, t \in \mathbb{R}} \quad & t \\ \text{s.t. } & \bar{x}_{i,r}^\top w - b \leq -t \\ & x^\top w - b \geq t \, \forall x \in \mathcal{S}_{r+1}^{-1} \\ & \|w\|_2 \leq 1. \end{aligned} \tag{10}$$

Then, $\bar{w} \in \mathbb{R}^d$ and $\bar{b} \in \mathbb{R}$ define the max-margin separator separating $\bar{x}_{i,r}$ from $\mathcal{S}_{r+1}^{-1}$. We note that by Lemma 3, $\bar{w}$ and $\bar{b}$ are the unique solutions up to scaling. By Section 8.6.1. in [8], equation 10 has the same value as

$$\begin{aligned} \min_{\alpha_j} \frac{1}{2} & \left\| \sum_{j : x_j \in \mathcal{S}_{r+1}^{-1}} \alpha_j x_j - \bar{x}_{i,r} \right\| \\ \text{s.t. } & \alpha_j \geq 0 \, \forall j \\ & \sum_j \alpha_j = 1 \end{aligned} \tag{11}$$

Let $\{\tilde{\alpha}_j\}$ attain the maximum in the above optimization problem, which exists since the domain is compact and the objective function is continuous. Define $\tilde{x} = \sum_{j : x_j \in \mathcal{S}_{r+1}^{-1}} \tilde{\alpha}_j x_j$ and

$$\begin{aligned} \tilde{w} &= \frac{\tilde{x} - \bar{x}_{i,r}}{\|\tilde{x} - \bar{x}_{i,r}\|} \\ \tilde{b} &= \frac{\|\tilde{x}\|^2 - \|\bar{x}_{i,r}\|^2}{2 \|\tilde{x} - \bar{x}_{i,r}\|} \end{aligned}$$

We claim that $\tilde{w} = \bar{w}$ and $\tilde{b} = \bar{b}$. First, we show that that there exists $\tilde{t} \in \mathbb{R}$ such that $(\tilde{w}, \tilde{b}, \tilde{t})$ satisfy the constraints in equation 10. Arithmetic shows that $\tilde{w}^\top \tilde{x} - \tilde{b} > 0$ and $\tilde{w}^\top \bar{x}_{i,r} - \tilde{b} < 0$. Since $\tilde{x}$ is the projection of $\bar{x}_{i,r}$ onto $\texttt{conv}(\mathcal{S}_r^+)$, by the Projection Lemma, we have that

$$(\tilde{x} - \bar{x}_{i,r})^\top x \geq (\tilde{x} - \bar{x}_{i,r})^\top \tilde{x}$$

for all $x \in \texttt{conv}(\mathcal{S}_r^+)$. Thus, for all $x \in \texttt{conv}(\mathcal{S}_r^+)$, we have that

$$\tilde{w}^\top x - \tilde{b} \geq \tilde{w}^\top \tilde{x} - \tilde{b} > 0.$$

We conclude that there exists $\tilde{t} \in \mathbb{R}$ such that $(\tilde{w}, \tilde{b}, \tilde{t})$ satisfy the constraints in equation 10.

We have that

$$\tilde{w}^\top \bar{x}_{i,r} - \tilde{b} = \frac{1}{\|\tilde{x} - \bar{x}_{i,r}\|} (\bar{x}_{i,r}^\top (\tilde{x} - \bar{x}_{i,r}) - \frac{\|\tilde{x}\|^2 - \left\|\bar{x}_{i,r}^2\right\|}{2}) = -\frac{1}{2} \|\bar{x}_{i,r} - \tilde{x}\| .$$

A similar calculation shows that $\tilde{w}^\top x - \tilde{b} \geq \frac{1}{2} \|\bar{x}_{i,r} - \tilde{x}\|$ for all $x \in \texttt{conv}(\mathcal{S}_r^+)$. Thus, putting $\tilde{t} = \frac{1}{2} \|\tilde{x} - \bar{x}_{i,r}\|$, $(\tilde{w}, \tilde{b}, \tilde{t})$ is feasible to equation 10. By the equivalence in value of equation 10 and equation 11 and the definition of $\tilde{x}$, we have that $\frac{1}{2} \|\tilde{x} - \bar{x}_{i,r}\| = \bar{t}$. Thus, by uniqueness of $\bar{w}$ and $\bar{b}$, we have that $\tilde{w} = \bar{w}$ and $\tilde{b} = \bar{b}$.

Step 2: Putting it together. We have shown that $\tilde{w}$ and $\tilde{b}$ define the max-margin separator. Let $P\bar{x}_{i,r}$ denote the projection of $\bar{x}_{i,r}$ onto $H := \{z \in \mathbb{R}^d : \tilde{w}^\top z = \tilde{b}\}$. We have that

$$\begin{aligned} P\bar{x}_{i,r} &= \bar{x}_{i,r} + (\frac{\|\tilde{x}\|^2 - \|\bar{x}_{i,r}\|^2}{2 \|\tilde{x} - \bar{x}_{i,r}\|} - (\frac{\tilde{x} - \bar{x}_{i,r}}{\|\tilde{x} - \bar{x}_{i,r}\|})^\top \bar{x}_{i,r} \frac{\tilde{x} - \bar{x}_{i,r}}{\|\tilde{x} - \bar{x}_{i,r}\|} \\ &= \frac{\tilde{x} + \bar{x}_{i,r}}{2}. \end{aligned}$$

Suppose that $\mathrm{dist}(\bar{x}_{i,r}, \mathtt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r}}})) \geq 2\delta$. Then, we have that $\mathrm{dist}(\bar{x}_{i,r}, H) > 2\delta$. It can be easily seen that $\tilde{x}_{i,r} = \arg\min_{x:\|x-\bar{x}_{i,r}\|\leq\delta} \mathrm{dist}(x, H) = \bar{x}_{i,r} + \delta\frac{\tilde{x}-\bar{x}_{i,r}}{\|\tilde{x}-\bar{x}_{i,r}\|}$. Note that $\bar{x}_{i,r} + \delta\frac{\tilde{x}-\bar{x}_{i,r}}{\|\tilde{x}-\bar{x}_{i,r}\|} \in \mathcal{X}$, since $\bar{x}_{i,r}, \tilde{x} \in \mathcal{X}$ and $\mathcal{X}$ is convex.

Similarly, if $\mathrm{dist}(\bar{x}_{i,r}, \mathtt{conv}(\mathcal{S}_{r+1}^{-y_{\bar{x}_{i,r}}})) < 2\delta$, Then, we have that $\mathrm{dist}(\bar{x}_{i,r}, H) < \delta$. Then, by definition of $\tilde{x}_{i,r}$ we have that $\tilde{x}_{i,r} = \frac{\tilde{x}+\bar{x}_{i,r}}{2} \in \mathcal{X}$, where we have that $\frac{\tilde{x}+\bar{x}_{i,r}}{2} \in \mathcal{X}$ by convexity of $\mathcal{X}$.

□

The following result shows that the max-margin separator is unique, and is a standard result on SVMs.

Lemma 3. *Let $A, B \subset \mathbb{R}^d$ be disjoint, closed, and convex. The max-margin separator separating A and B is unique.*

Proof. There exists a separating hyperplane between A and B by the separating hyperplane Theorem. By a standard argument, the optimization problem for the max-margin separator can be stated as

$$\begin{aligned} \min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \|w\|^2 \\ \text{s.t. } w^\top x + b \geq 1 \,\forall x \in A \\ w^\top x + b \leq -1 \,\forall x \in B. \end{aligned}$$

This is a convex optimization problem with a strongly convex objective and, therefore, has a unique solution. □

SEALS with Querying Anywhere Capability

Algorithm 4 Modified SEALS: Project onto Hyperplane

Input: seed labeled points $\mathcal{S}_1 \subset \mathcal{X} \times \{-1, 1\}$ $r = 1$
Initialize the clusters $C_{i,r} = \{x_{i,0}\}$ for $i = 1, \ldots, d-1$
for $r = 1, 2, \ldots$ **do**
  for $i = 1, \ldots, d-1$ **do**
    $(H_{x,r}, \gamma_{x,r}) = \text{MaxMarginSeparator}(\mathcal{S}_r^{-y_x}, x)$ for all $x \in C_{i,r}$
    Let $\bar{x}_{i,r} \in \arg\min_{x\in C_{i,r}} \gamma_{x,r}$ and $\widetilde{x}_{i,r} \in \arg\min_{x'} \mathrm{dist}(x', H_{\bar{x}_{i,r},r})$
    Query $\widetilde{x}_{i,r}$
    Update $\mathcal{S}_{r+1} = \mathcal{S}_r \cup \{(\widetilde{x}_{i,r}, y_{\bar{x}_{i,r}})\}$ and $C_{i,r+1} = C_{i,r} \cup \{\widetilde{x}_{i,r}\}$
  end for
  Fit a homogenous max-margin separator with normal vector $\widehat{w}_{r+1}$ to $\mathcal{S}_{r+1}$
end for

The data structure used in SEALS enables queries of the k-nearest neighbors of any point in the input space. Therefore, a natural question is whether we can leverage this querying capability to improve the sample complexity of Algorithm 3, *at the cost of being less generic than SEALS*. To address this question, we consider the Algorithm 4, which queries the nearest neighbor of the projection onto the max-margin separator. The key takeaway is that this modification of the algorithm removes the slow phase in the sample complexity of SEALS. The analysis requires a slightly different definition of $\mathcal{Z}_i$:

$$\mathcal{Z}_i = \{z \in \mathbb{R}^d : \|z - x_{i,0}\| \leq 2\gamma_i + \epsilon \text{ and } \mathrm{dist}(z, \{x \in \mathbb{R}^d : x^\top w_* = 0\}) \leq \epsilon\}.$$

Theorem 3. *Let $\epsilon > 0$. Let $x_{1,0}, \ldots, x_{d-1,0}$ denote the seed points. Define $\gamma_i = \mathrm{dist}(x_{i,0}, \mathtt{conv}(\mathcal{S}_1^{-y_{x_{i,0}}}))$ for $i \in [d-1]$, where $\mathtt{conv}(\mathcal{S}_1^{-y_{x_{i,0}}})$ is the convex hull of the points $\mathcal{S}_1^{-y_{x_{i,0}}}$. Then, after Algorithm 4 makes $\max_{i\in[d-1]} d(2\log(\frac{2d\gamma_i}{\epsilon\min(\sigma,1)}))$ queries, we have that $\|\widehat{w}_r - w_*\| \leq \epsilon$.*

Proof Sketch. The proof is very similar to the of Theorem 2. Step 1 of Theorem 2 is essentially the same, but it does not matter whether $\rho_{i,r} \geq 2\delta$ since the algorithm is not constrained by the nearest

neighbor graph. A similar argument shows that after $\max_{i\in[d-1]} \log(\frac{2\gamma_i}{\bar{\epsilon}})$ queries, we have that for all $i \in [d-1]$, $\rho_{i,r} := \mathrm{dist}(\bar{x}_{i,r}, \mathtt{conv}(\mathcal{S}_r^{-y_{x_{i,r}}})) \leq \bar{\epsilon}$.

Step 2 is also quite similar, except that we now have using a similar argument about the geometric series that

$$\|x_{i,r} - x_{i,0}\| \leq 2\gamma_i.$$

Step 3 is exactly the same.

□

Impact of Embedding model (G_z) on SEALS

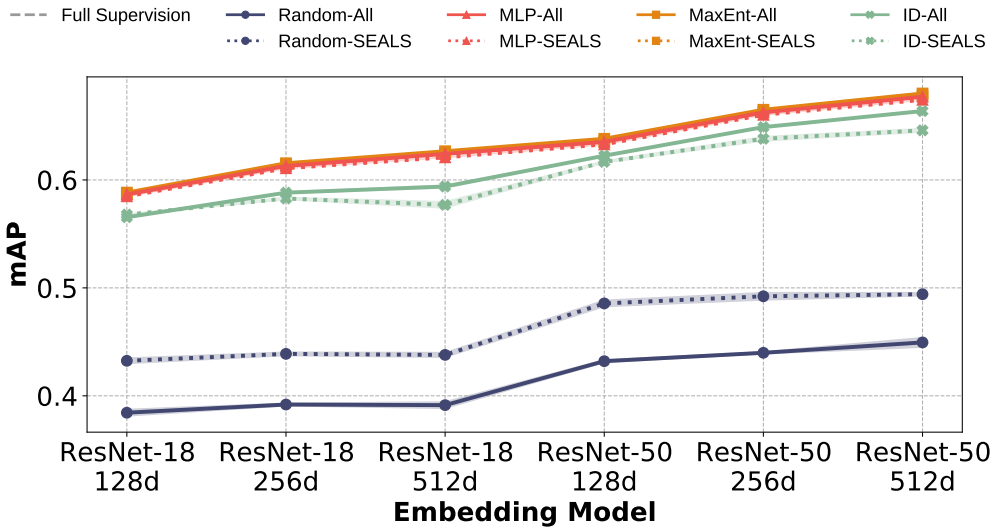


Figure 4: Active learning on ImageNet with varying embedding models (ResNet-18 or ResNet-50) and dimensions (128, 256, or 512). Performance increases with larger models and higher dimensional embeddings. However, SEALS achieves similar performance to the baseline approach regardless of the choice of model and dimension, empirically demonstrating SEALS' robustness to the embedding function.

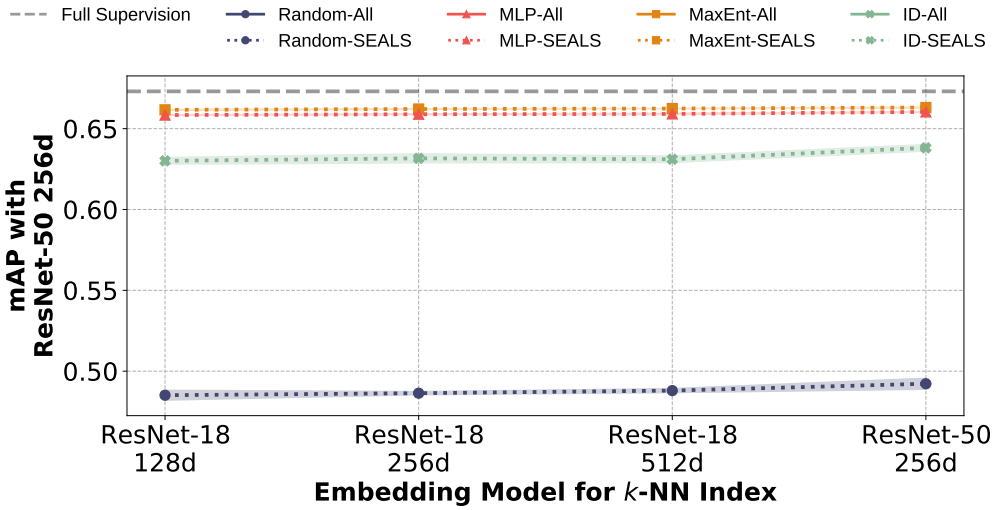


Figure 5: Active learning on ImageNet with 256-dimensional ResNet-50 embeddings and varying k-NN indices. Different embeddings might be used for learning rare concepts than the embeddings used for similarity search in practice. Fortunately, SEALS performs similarly for varying k-NN indices, as shown above. This can also be exploited to reduce further the cost of constructing the index by using a smaller, cheaper model to generate the embedding for similarity search and only applying the larger model to examples added to the candidate pool.

Impact of k on SEALS

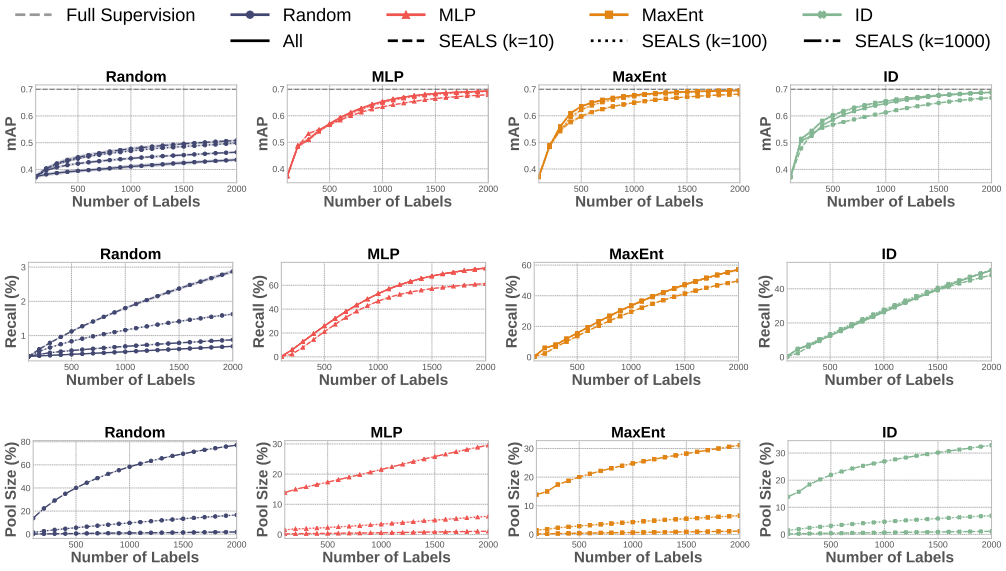


Figure 6: **Impact of increasing k on ImageNet ($|U|$=639,906).** Larger values of k help to close the gap between SEALS and the baseline approach for active learning (top) and active search (middle). However, increasing k also increases the candidate pool size (bottom), presenting a trade-off between labeling efficiency and computational efficiency.

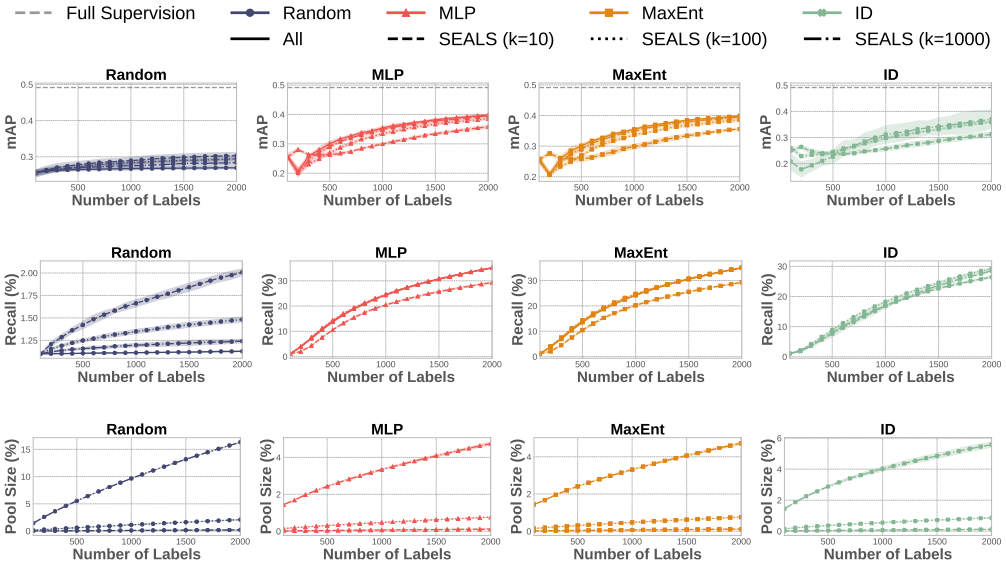
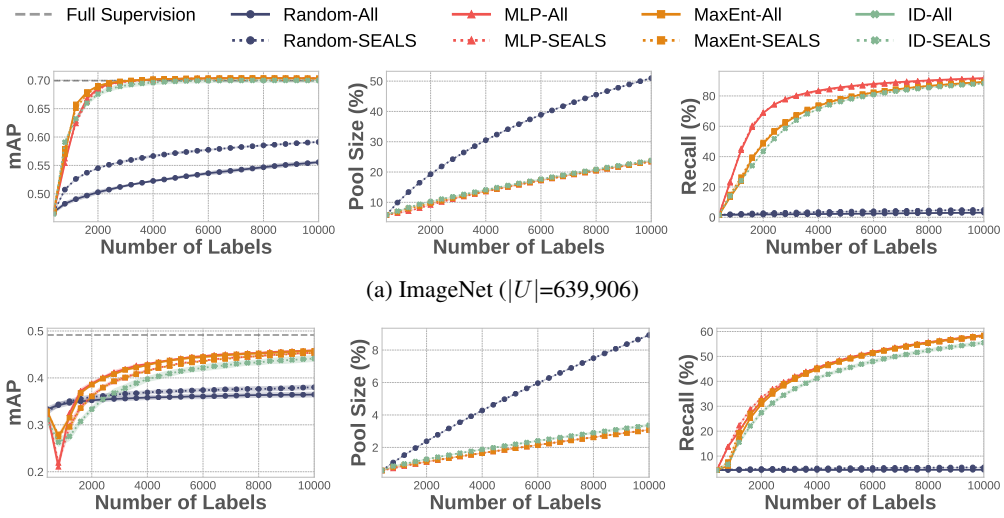


Figure 7: **Impact of increasing k on OpenImages ($|U|$=6,816,296).** Larger values of k help to close the gap between SEALS and the baseline approach for active learning (top) and active search (middle). However, increasing k also increases the candidate pool size (bottom), presenting a trade-off between labeling efficiency and computational efficiency.

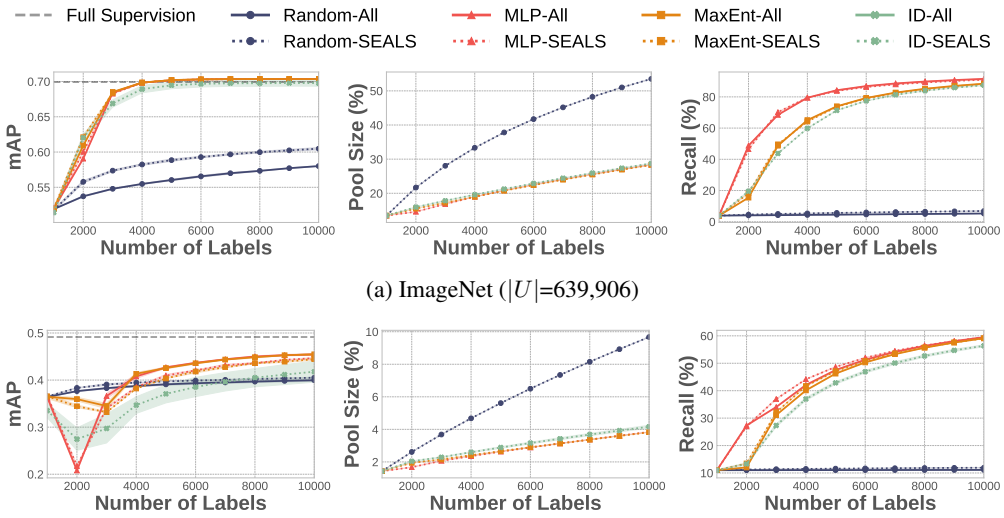
Impact of the number of initial positives on SEALS



(a) ImageNet ($|U|$=639,906)

(b) OpenImages ($|U|$=6,816,296)

Figure 8: **Active learning and search with 20 positive seed examples** and a labeling budget of 10,000 examples on ImageNet (top) and OpenImages (bottom). Across datasets and strategies, SEALS with $k = 100$ performs similarly to the baseline approach in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right), while only considering a fraction of the unlabeled data U (middle).

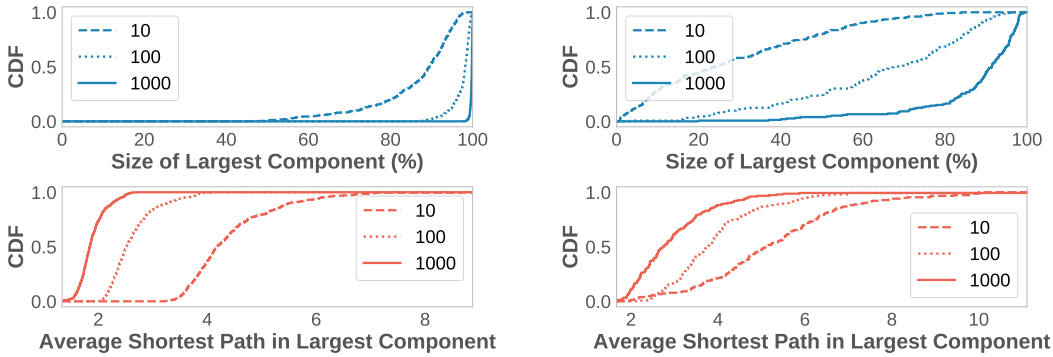


(a) ImageNet ($|U|$=639,906)

(b) OpenImages ($|U|$=6,816,296)

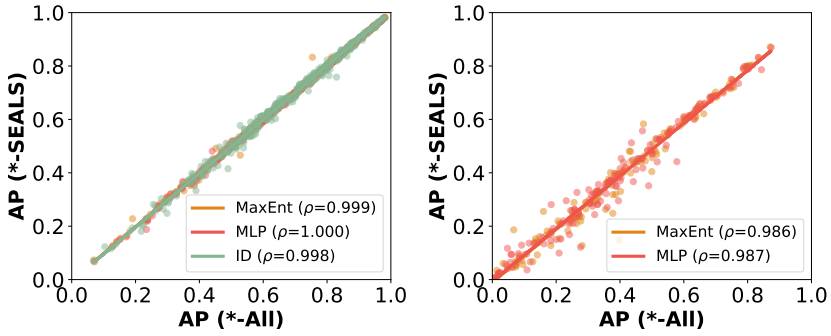
Figure 9: **Active learning and search with 50 positive seed examples** and a labeling budget of 10,000 examples on ImageNet (top) and OpenImages (bottom). Across datasets and strategies, SEALS with $k = 100$ performs similarly to the baseline approach in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right), while only considering a fraction of the unlabeled data U (middle).

Latent structure



(a) ImageNet (b) OpenImages

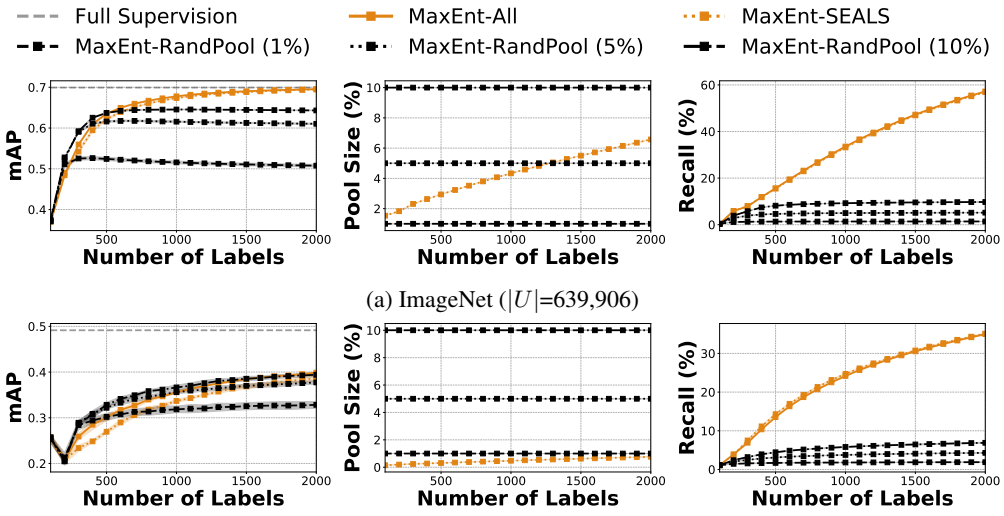
Figure 10: Measurements of the latent structure of unseen concepts in ImageNet and OpenImages. The 10B images dataset was excluded because only a few thousand examples were labeled. The largest connected component gives a sense of how much of the concept SEALS can reach, while the average shortest path serves as a proxy for how long it will take to explore.



(a) ImageNet (b) OpenImages

Figure 11: The per-class APs of SEALS ($k = 100$) were highly correlated to the baseline approaches (*-All) for active learning on ImageNet (right) and OpenImages (left). On OpenImages with a budget of 2,000 labels, the Pearson's correlation (ρ) between the baseline and SEALS for the average precision of individual classes was 0.986 for MaxEnt and 0.987 for MLP. The least-squares fit had a slope of 0.99 and y-intercept of -0.01. On ImageNet, the correlations were even higher.

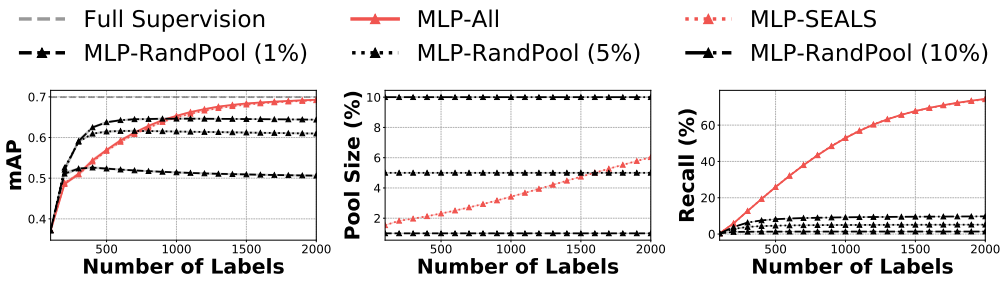
Comparison to pool of randomly selected examples



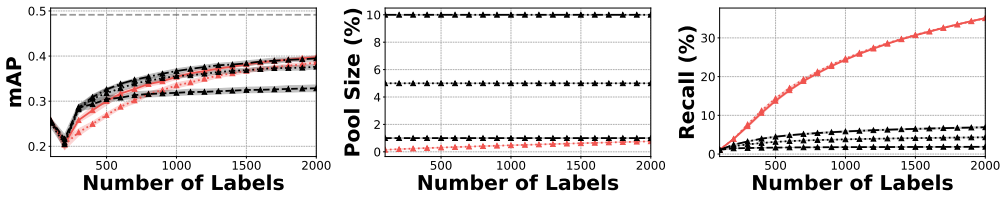
(a) ImageNet ($|U|$=639,906)

(b) OpenImages ($|U|$=6,816,296)

Figure 12: MaxEnt-SEALS ($k = 100$) versus MaxEnt applied to a candidate pool of randomly selected examples (RandPool). Because the concepts we considered were so rare, as is often the case in practice, randomly chosen examples are unlikely to be close to the decision boundary, and a much larger pool is required to match SEALS. On ImageNet (top), MaxEnt-SEALS outperformed MaxEnt-RandPool in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right) even with a pool containing 10% of the data (middle). On Openimages (bottom), MaxEnt-RandPool needed at least $5\times$ as much data to match MaxEnt-SEALS for active learning and failed to achieve similar recall even with $10\times$ the data.

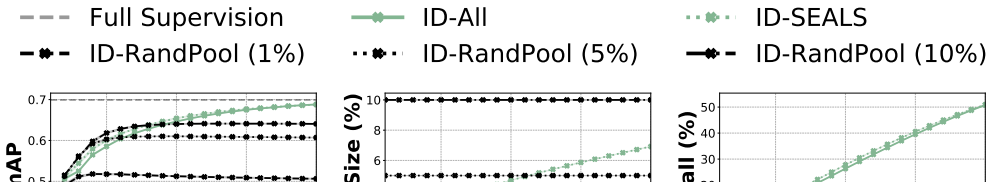


(a) ImageNet ($|U|$=639,906)

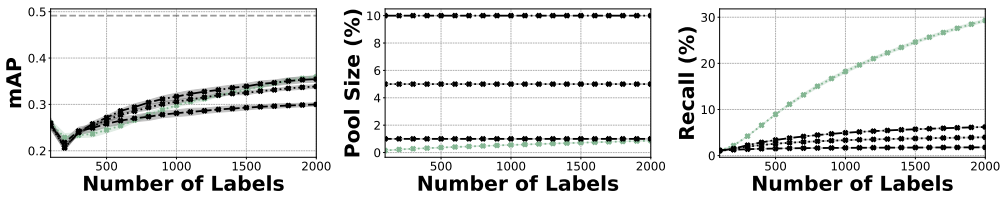


(b) OpenImages ($|U|$=6,816,296)

Figure 13: MLP-SEALS ($k = 100$) versus MLP applied to a candidate pool of randomly selected examples (RandPool). Because the concepts we considered were so rare, as is often the case in practice, randomly chosen examples are unlikely to be close to the decision boundary, and a much larger pool is required to match SEALS. On ImageNet (top), MLP-SEALS outperformed MLP-RandPool in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right) even with a pool containing 10% of the data (middle). On Openimages (bottom), MLP-RandPool needed at least $5\times$ as much data to match MLP-SEALS for active learning and failed to achieve similar recall even with $10\times$ the data.



(a) ImageNet ($|U|$=639,906)



(b) OpenImages ($|U|$=6,816,296)

Figure 14: ID-SEALS ($k = 100$) versus ID applied to a candidate pool of randomly selected examples (RandPool). Because the concepts we considered were so rare, as is often the case in practice, randomly chosen examples are unlikely to be close to the decision boundary, and a much larger pool is required to match SEALS. On ImageNet (top), ID-SEALS outperformed ID-RandPool in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right) even with a pool containing 10% of the data (middle). On Openimages (bottom), ID-RandPool needed at least $5\times$ as much data to match ID-SEALS for active learning and failed to achieve similar recall even with $10\times$ the data.

Active learning on each selected class from OpenImages

Table 3: **Top $\frac{1}{3}$ of classes from Openimages for active learning.** (1 of 3) Average precision and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MaxEnt-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MaxEnt (SEALS)	MaxEnt (All)	Full Supervision
Citrus	796	65	3.34	0.34	0.87	0.87	0.87
Cargo ship	219	84	2.85	0.70	0.83	0.83	0.86
Blackberry	245	87	2.64	0.67	0.80	0.80	0.79
Galliformes	674	82	3.98	0.72	0.80	0.82	0.92
Rope	618	59	3.48	0.29	0.80	0.81	0.74
Hurdling	269	92	2.48	0.26	0.80	0.79	0.80
Roman temple	345	89	2.72	0.63	0.79	0.79	0.82
Monster truck	286	84	2.84	0.41	0.79	0.80	0.81
Pasta	954	91	3.21	0.42	0.75	0.75	0.79
Chess	740	83	3.39	0.53	0.73	0.74	0.86
Bowed string instrument	728	78	3.05	0.72	0.72	0.74	0.79
Parrot	1546	89	2.85	0.59	0.72	0.76	0.92
Calabaza	870	82	3.15	0.50	0.71	0.75	0.81
Superhero	968	58	5.28	0.17	0.70	0.70	0.67
Drums	741	69	3.30	0.52	0.70	0.72	0.83
Shooting range	189	57	3.06	0.38	0.69	0.69	0.68
Ancient roman architecture	589	76	3.34	0.61	0.68	0.70	0.77
Cupboard	898	88	3.41	0.53	0.68	0.69	0.75
Ibis	259	93	2.53	0.29	0.68	0.69	0.66
Cattle	5995	93	3.22	0.37	0.67	0.68	0.74
Galleon	182	74	2.54	0.45	0.66	0.66	0.61
Kitchen knife	360	63	3.52	0.32	0.66	0.65	0.66
Grapefruit	506	83	3.06	0.50	0.65	0.65	0.69
Deacon	341	80	2.80	0.48	0.64	0.64	0.67
Rye	128	75	2.63	0.51	0.64	0.64	0.65
Chartreux	147	91	2.59	0.50	0.63	0.63	0.69
San Pedro cactus	318	76	3.32	0.17	0.62	0.63	0.71
Skateboarding Equipment	862	57	5.92	0.20	0.62	0.66	0.66
Electric piano	345	56	4.15	0.24	0.61	0.60	0.48
Straw	547	65	2.85	0.33	0.61	0.62	0.61
Berry	874	82	3.78	0.30	0.61	0.61	0.69
East-european shepherd	206	86	2.16	0.61	0.61	0.62	0.65
Ring	676	75	3.87	0.15	0.61	0.64	0.64
Rat	1151	94	2.50	0.32	0.60	0.60	0.61
Coral reef fish	434	90	3.07	0.51	0.60	0.64	0.79
Concert dance	357	61	3.91	0.37	0.60	0.60	0.70
Whole food	708	73	3.66	0.18	0.58	0.60	0.57
Modern pentathlon	772	43	2.59	0.13	0.58	0.47	0.51
Gymnast	235	77	2.39	0.39	0.57	0.59	0.65
California roll	368	84	3.49	0.05	0.56	0.56	0.58
Shrimp	907	85	3.82	0.07	0.56	0.56	0.58
Log cabin	448	70	3.62	0.44	0.55	0.55	0.62
Formula racing	351	88	3.38	0.33	0.55	0.54	0.60
Herd	648	75	3.88	0.42	0.54	0.55	0.67
Embroidery	356	81	3.41	0.32	0.53	0.53	0.60
Shelving	810	66	3.41	0.27	0.53	0.53	0.51
Downhill	194	84	2.64	0.42	0.53	0.51	0.59
Daylily	391	87	3.25	0.20	0.51	0.50	0.49
Automotive exterior	1060	23	2.74	0.65	0.49	0.54	0.69
Ciconiiformes	426	88	3.47	0.33	0.49	0.51	0.48
Monoplane	756	81	4.70	0.13	0.48	0.43	0.48

Table 4: **Middle $\frac{1}{3}$ of classes from Openimages for active learning.** (2 of 3) Average precision and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MaxEnt-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MaxEnt (SEALS)	MaxEnt (All)	Full Supervision
Seafood boil	322	85	2.73	0.31	0.48	0.49	0.51
Landscaping	789	32	4.71	0.26	0.48	0.51	0.63
Skating	561	77	4.04	0.17	0.48	0.43	0.40
Floodplain	567	50	4.81	0.61	0.47	0.52	0.66
Knitting	409	71	3.10	0.61	0.46	0.50	0.73
Elk	353	84	2.40	0.15	0.46	0.48	0.45
Bilberry	228	75	3.77	0.10	0.45	0.45	0.32
Goat	1190	88	3.72	0.17	0.44	0.45	0.61
Fortification	287	66	3.96	0.43	0.44	0.46	0.52
Annual plant	677	38	6.07	0.39	0.44	0.43	0.58
Mcdonnell douglas f/a-18 hornet	160	88	3.51	0.11	0.44	0.47	0.37
Tooth	976	49	4.77	0.16	0.44	0.48	0.56
Briefs	539	78	3.68	0.15	0.43	0.44	0.46
Sirloin steak	297	60	4.97	0.14	0.42	0.42	0.46
Smoothie	330	78	3.22	0.15	0.41	0.41	0.38
Glider	393	82	3.94	0.08	0.40	0.40	0.48
Bathroom cabinet	368	95	2.39	0.29	0.40	0.39	0.37
White-tailed deer	238	87	3.24	0.34	0.40	0.43	0.43
Bird of prey	712	78	3.81	0.76	0.40	0.50	0.91
Egg (Food)	1193	85	4.31	0.14	0.40	0.37	0.63
Soldier	1032	74	3.80	0.62	0.40	0.41	0.72
Cranberry	450	63	4.10	0.13	0.39	0.39	0.37
Estate	667	51	4.03	0.47	0.39	0.40	0.54
Chocolate truffle	288	58	5.47	0.10	0.39	0.40	0.42
Town square	617	58	3.69	0.31	0.38	0.36	0.47
Bakmi	191	76	3.34	0.27	0.37	0.37	0.36
Trail riding	679	90	3.15	0.21	0.37	0.37	0.38
Aerial photography	931	63	3.99	0.39	0.37	0.37	0.66
Lugger	103	62	3.14	0.35	0.37	0.37	0.42
Paddy field	468	70	4.02	0.17	0.36	0.36	0.43
Pavlova	195	86	2.60	0.19	0.36	0.36	0.34
Steamed rice	580	75	4.54	0.10	0.35	0.37	0.48
Pancit	385	86	3.16	0.21	0.33	0.33	0.31
Factory	333	61	5.59	0.17	0.33	0.34	0.35
Fur	834	42	4.31	0.08	0.33	0.33	0.31
Stallion	598	70	3.58	0.32	0.33	0.40	0.64
Optical instrument	649	79	3.91	0.15	0.33	0.33	0.28
Thumb	895	26	4.18	0.07	0.32	0.39	0.41
Meal	1250	60	5.68	0.52	0.32	0.38	0.59
American shorthair	2084	94	3.32	0.12	0.32	0.32	0.24
Bracelet	770	46	4.13	0.09	0.31	0.33	0.24
Vehicle registration plate	5697	76	5.89	0.28	0.31	0.33	0.53
Ice	682	50	4.87	0.23	0.30	0.32	0.55
Lamian	257	80	3.57	0.23	0.29	0.32	0.28
Multimedia	741	46	4.12	0.45	0.29	0.31	0.53
Belt	467	41	3.26	0.06	0.29	0.31	0.31
Prairie	792	44	3.92	0.37	0.29	0.26	0.57
Boardsport	673	62	4.08	0.26	0.29	0.29	0.53
Asphalt	1026	40	4.53	0.23	0.29	0.29	0.45
Costume design	818	52	3.44	0.07	0.26	0.26	0.28
Cottage	670	51	4.13	0.36	0.26	0.36	0.61

Table 5: **Bottom $\frac{1}{3}$ of classes from Openimages for active learning.** (3 of 3) Average precision and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MaxEnt-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MaxEnt (SEALS)	MaxEnt (All)	Full Supervision
Stele	450	70	3.74	0.12	0.26	0.25	0.35
Mode of transport	1387	24	4.50	0.15	0.26	0.16	0.54
Temperate coniferous forest	328	59	4.23	0.30	0.26	0.29	0.40
Bumper	985	37	6.65	0.49	0.25	0.38	0.64
Interaction	924	15	6.05	0.04	0.24	0.25	0.37
Plumbing fixture	2124	89	3.19	0.31	0.24	0.27	0.38
Shorebird	234	80	2.76	0.32	0.23	0.26	0.37
Icing	1118	74	4.20	0.13	0.23	0.25	0.46
Wilderness	1225	30	4.12	0.29	0.23	0.24	0.39
Construction	515	63	4.99	0.13	0.23	0.26	0.34
Carpet	644	50	6.98	0.05	0.23	0.28	0.43
Maple	2301	90	4.19	0.06	0.22	0.21	0.36
Rural area	921	41	4.63	0.33	0.22	0.28	0.50
Singer	604	56	4.06	0.12	0.21	0.21	0.40
Delicatessen	196	52	2.80	0.14	0.21	0.22	0.27
Canal	726	62	4.78	0.22	0.21	0.26	0.46
Organ (Biology)	1156	25	3.80	0.23	0.19	0.07	0.44
Laugh	750	19	6.22	0.06	0.18	0.17	0.26
Plateau	452	37	3.88	0.41	0.18	0.24	0.46
Algae	426	57	4.52	0.15	0.18	0.19	0.26
Cactus	377	51	4.11	0.05	0.17	0.18	0.22
Engine	656	82	3.43	0.16	0.17	0.17	0.26
Marine mammal	2954	91	3.58	0.19	0.16	0.15	0.21
Frost	483	60	4.73	0.20	0.15	0.21	0.47
Paper	969	23	3.18	0.16	0.15	0.14	0.41
Cirque	347	29	5.77	0.43	0.15	0.40	0.55
Pork	464	64	4.44	0.06	0.14	0.14	0.15
Antenna	545	73	3.66	0.10	0.14	0.13	0.29
Portrait	2510	67	6.38	0.23	0.13	0.18	0.43
Flooring	814	38	3.87	0.10	0.13	0.14	0.20
Cycling	794	63	5.00	0.53	0.13	0.28	0.66
Chevrolet silverado	115	62	4.82	0.05	0.09	0.08	0.12
Tool	1549	64	4.51	0.08	0.09	0.10	0.13
Liqueur	539	51	5.98	0.26	0.09	0.14	0.38
Pleurotus eryngii	140	84	3.10	0.11	0.08	0.08	0.14
Organism	1148	21	3.49	0.05	0.07	0.13	0.26
Pelecaniformes	457	85	3.96	0.30	0.07	0.09	0.32
Icon	186	15	3.26	0.05	0.07	0.07	0.16
Stadium	1654	77	5.77	0.35	0.06	0.10	0.48
Space	1006	23	4.63	0.03	0.06	0.03	0.14
Performing arts	1030	29	6.97	0.12	0.05	0.06	0.53
Mural	649	41	5.24	0.13	0.05	0.07	0.34
Brown	1427	16	3.49	0.02	0.05	0.07	0.20
Wall	1218	27	3.13	0.11	0.05	0.05	0.27
Tournament	841	47	9.90	0.15	0.05	0.07	0.16
White	1494	3	2.79	0.02	0.03	0.01	0.10
Mitsubishi	511	37	5.14	0.01	0.02	0.02	0.04
Exhibition	513	40	3.87	0.03	0.02	0.02	0.14
Scale model	667	45	5.64	0.05	0.02	0.02	0.13
Teal	975	16	4.08	0.01	0.01	0.01	0.04
Electric blue	1180	19	3.70	0.01	0.00	0.01	0.06

Active search on each selected class from OpenImages

Table 6: **Top $\frac{1}{3}$ of classes from Openimages for active search.** (1 of 3) Recall (%) of positives and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MLP-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MLP (SEALS)	MLP (All)
Chartreux	147	91	2.59	3.5	83.9	84.6
Ibis	259	93	2.53	2.0	83.9	83.9
Hurdling	269	92	2.48	1.9	83.5	86.2
East-european shepherd	206	86	2.16	2.4	78.2	78.3
Blackberry	245	87	2.64	2.0	77.5	78.5
Bathroom cabinet	368	95	2.39	1.4	76.8	77.1
Rat	1151	94	2.50	0.5	75.1	75.2
Rye	128	75	2.63	3.9	74.7	74.5
Elk	353	84	2.40	1.5	73.4	74.3
Pavlova	195	86	2.60	2.6	70.8	71.3
Seafood boil	322	85	2.73	1.6	70.4	70.6
Roman temple	345	89	2.72	1.5	69.2	68.3
Monster truck	286	84	2.84	1.7	68.1	67.8
Downhill	194	84	2.64	2.6	67.2	69.0
Shorebird	234	80	2.76	2.1	66.8	66.4
Mcdonnell douglas f/a-18 hornet	160	88	3.51	3.2	66.0	67.9
San Pedro cactus	318	76	3.32	1.6	65.8	64.9
Pleurotus eryngii	140	84	3.10	3.6	65.7	66.1
California roll	368	84	3.49	1.4	65.3	68.0
Gymnast	235	77	2.39	2.2	64.0	64.0
Galleon	182	74	2.54	2.7	62.4	61.5
Cargo ship	219	84	2.85	2.3	61.1	61.7
Trail riding	679	90	3.15	0.8	59.7	60.7
Daylily	391	87	3.25	1.3	59.4	59.5
Grapefruit	506	83	3.06	1.0	59.4	60.4
Bilberry	228	75	3.77	2.2	58.9	55.2
Smoothie	330	78	3.22	1.5	58.0	59.8
Embroidery	356	81	3.41	1.5	57.6	57.2
Deacon	341	80	2.80	1.5	57.1	57.9
Shooting range	189	57	3.06	2.6	56.3	55.6
Glider	393	82	3.94	1.3	55.8	57.6
White-tailed deer	238	87	3.24	2.2	55.8	55.9
Coral reef fish	434	90	3.07	1.3	54.8	54.9
Chevrolet silverado	115	62	4.82	4.3	54.1	54.6
Lugger	103	62	3.14	4.9	53.8	53.8
Pancit	385	86	3.16	1.3	52.8	53.1
Chess	740	83	3.39	0.7	51.9	50.9
Bakmi	191	76	3.34	2.6	51.8	51.2
Kitchen knife	360	63	3.52	1.5	50.9	53.9
Straw	547	65	2.85	1.0	50.3	51.0
Ancient roman architecture	589	76	3.34	0.8	48.5	47.1
Lamian	257	80	3.57	1.9	47.8	48.2
Antenna	545	73	3.66	1.0	47.4	48.0
Calabaza	870	82	3.15	0.6	46.0	45.8
Ring	676	75	3.87	0.7	45.2	45.4
Ciconiiformes	426	88	3.47	1.2	45.2	45.2
Log cabin	448	70	3.62	1.1	44.9	45.7
Bowed string instrument	728	78	3.05	0.7	44.4	44.7
Pasta	954	91	3.21	0.5	43.7	43.8
Knitting	409	71	3.10	1.3	43.5	42.8
Rope	618	59	3.48	0.8	43.0	42.8

Table 7: **Middle $\frac{1}{3}$ of classes from Openimages for active search.** (2 of 3) Recall (%) of positives and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MLP-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MLP (SEALS)	MLP (All)
Formula racing	351	88	3.38	1.4	42.6	41.4
Paddy field	468	70	4.02	1.1	42.6	44.2
Engine	656	82	3.43	0.8	41.7	40.6
Electric piano	345	56	4.15	1.5	40.9	42.1
Shrimp	907	85	3.82	0.6	40.4	40.8
Goat	1190	88	3.72	0.4	39.6	39.6
Chocolate truffle	288	58	5.47	1.8	39.6	39.9
Cupboard	898	88	3.41	0.6	39.6	39.6
Citrus	796	65	3.34	0.7	39.3	39.6
Parrot	1546	89	2.85	0.4	39.2	38.8
Delicatessen	196	52	2.80	2.6	38.2	39.0
Berry	874	82	3.78	0.6	37.8	37.6
Briefs	539	78	3.68	1.0	37.1	37.2
Concert dance	357	61	3.91	1.4	36.6	36.1
Modern pentathlon	772	43	2.59	0.6	35.9	32.6
Fortification	287	66	3.96	1.7	35.7	37.6
Stallion	598	70	3.58	0.9	35.7	36.3
Belt	467	41	3.26	1.1	35.2	34.9
Sirloin steak	297	60	4.97	1.8	33.9	32.7
Stele	450	70	3.74	1.1	33.9	32.7
Galliformes	674	82	3.98	0.7	33.9	33.9
Algae	426	57	4.52	1.2	33.8	33.1
Herd	648	75	3.88	0.8	33.5	33.7
Pelecaniformes	457	85	3.96	1.1	33.4	37.5
Cactus	377	51	4.11	1.3	33.4	35.2
Shelving	810	66	3.41	0.7	33.2	33.3
Drums	741	69	3.30	0.7	32.9	32.7
Cranberry	450	63	4.10	1.2	32.9	33.7
Factory	333	61	5.59	1.5	32.0	31.7
Costume design	818	52	3.44	0.6	30.9	30.6
Optical instrument	649	79	3.91	0.8	30.3	32.8
Construction	515	63	4.99	1.0	30.1	31.1
Temperate coniferous forest	328	59	4.23	1.5	30.1	27.6
Skating	561	77	4.04	1.0	28.8	30.4
Egg (Food)	1193	85	4.31	0.4	28.8	28.6
Steamed rice	580	75	4.54	0.9	28.1	30.2
Plumbing fixture	2124	89	3.19	0.3	27.9	27.9
Whole food	708	73	3.66	0.7	27.7	27.5
Boardsport	673	62	4.08	0.8	26.8	26.5
Pork	464	64	4.44	1.1	26.3	26.6
Aerial photography	931	63	3.99	0.6	25.8	26.1
Town square	617	58	3.69	0.8	25.7	26.1
Estate	667	51	4.03	0.9	24.8	25.9
Maple	2301	90	4.19	0.2	24.3	24.4
Cattle	5995	93	3.22	0.1	23.8	23.6
Superhero	968	58	5.28	0.6	23.4	23.3
Bracelet	770	46	4.13	0.6	23.2	24.8
Frost	483	60	4.73	1.0	23.1	22.5
Scale model	667	45	5.64	0.8	22.9	23.7
Plateau	452	37	3.88	1.1	22.7	19.1
Bird of prey	712	78	3.81	0.7	22.4	22.0

Table 8: **Bottom $\frac{1}{3}$ of classes from Openimages for active search.** (3 of 3) Recall (%) of positives and measurements of the largest component (LC) for each selected class (153 total) from OpenImages with a labeling budget of 2,000 examples. Classes are ordered based on MLP-SEALS.

Display Name	Total Positives	Size of the LC (%)	Average Shortest Path in the LC	Random (All)	MLP (SEALS)	MLP (All)
Canal	726	62	4.78	0.7	22.4	20.9
Exhibition	513	40	3.87	1.0	21.9	23.1
Carpet	644	50	6.98	0.8	21.9	22.7
Monoplane	756	81	4.70	0.7	21.8	20.1
Ice	682	50	4.87	0.8	21.6	23.1
Fur	834	42	4.31	0.6	21.2	17.3
Icing	1118	74	4.20	0.4	20.5	20.1
Flooring	814	38	3.87	0.6	20.4	16.9
Icon	186	15	3.26	2.7	19.9	17.2
Prairie	792	44	3.92	0.6	19.0	19.2
Tooth	976	49	4.77	0.5	18.6	18.0
Skateboarding Equipment	862	57	5.92	0.6	18.1	19.3
Automotive exterior	1060	23	2.74	0.5	17.7	11.9
Cottage	670	51	4.13	0.7	17.6	17.3
Soldier	1032	74	3.80	0.5	17.3	16.8
Marine mammal	2954	91	3.58	0.2	17.3	17.2
Tool	1549	64	4.51	0.3	17.0	16.9
Multimedia	741	46	4.12	0.7	16.8	17.1
American shorthair	2084	94	3.32	0.3	16.5	16.7
Asphalt	1026	40	4.53	0.5	15.1	11.5
Singer	604	56	4.06	0.9	14.6	13.6
Floodplain	567	50	4.81	0.9	14.6	14.0
Rural area	921	41	4.63	0.6	14.2	13.2
Mitsubishi	511	37	5.14	1.0	12.6	11.8
Organ (Biology)	1156	25	3.80	0.5	12.1	15.9
Paper	969	23	3.18	0.5	12.0	14.8
Annual plant	677	38	6.07	0.7	11.8	10.7
Electric blue	1180	19	3.70	0.5	11.5	9.4
Stadium	1654	77	5.77	0.3	10.8	9.3
Mural	649	41	5.24	0.8	10.4	10.3
Teal	975	16	4.08	0.5	9.9	10.4
Cirque	347	29	5.77	1.5	9.9	9.8
Wall	1218	27	3.13	0.4	9.3	12.0
Thumb	895	26	4.18	0.6	9.3	13.8
Landscaping	789	32	4.71	0.7	9.2	9.3
Vehicle registration plate	5697	76	5.89	0.1	8.7	8.3
Meal	1250	60	5.68	0.4	8.5	9.1
Wilderness	1225	30	4.12	0.4	8.5	9.8
Liqueur	539	51	5.98	1.0	8.0	12.8
Space	1006	23	4.63	0.5	7.8	6.3
Cycling	794	63	5.00	0.6	7.3	7.8
Brown	1427	16	3.49	0.4	7.2	2.6
Organism	1148	21	3.49	0.4	6.8	2.0
Laugh	750	19	6.22	0.7	6.6	8.6
Bumper	985	37	6.65	0.5	5.9	8.3
Portrait	2510	67	6.38	0.2	5.8	5.3
Mode of transport	1387	24	4.50	0.4	5.1	3.6
Interaction	924	15	6.05	0.6	4.5	4.6
Tournament	841	47	9.90	0.6	4.3	5.1
Performing arts	1030	29	6.97	0.5	2.3	2.5
White	1494	3	2.79	0.3	2.0	0.5

Self-supervised embeddings (SimCLR) on ImageNet

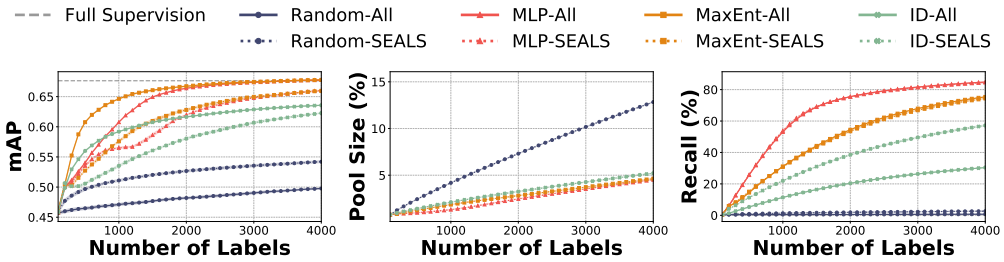


Figure 15: Active learning and search on ImageNet with self-supervised embeddings from SimCLR [11]. Because the self-supervised training for the embeddings did not use the labels, results are average across all 1,000 classes and $|U|$=1,281,167. To compensate for the larger unlabeled pool, we extended the total labeling budget to 4,000 compared to the 2,000 used in Figure 1. Across strategies, SEALS with $k = 100$ substantially outperforms random sampling in terms of both the mAP the model achieves for active learning (left) and the recall of positive examples for active search (right), while only considering a fraction of the data U (middle). For active learning, the gap between the baseline and SEALS approaches is slightly larger than in Figure 1, which is likely due to the larger pool size and increased average shortest paths (see Figure 16).

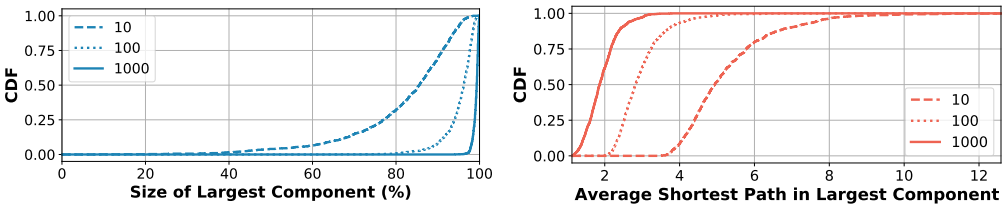


Figure 16: Measurements of the latent structure of unseen concepts in ImageNet with self-supervised embeddings from SimCLR [11]. In comparison to Figure 10a, the k-nearest neighbor graph for unseen concepts was still well connected, forming large connected components (left) for even moderate values of k, but the average shortest path between examples was slightly longer (right). The increased path length is not too surprising considering the fully supervised model still outperformed the linear evaluation of the self-supervised embeddings in Chen et al. [11].

Self-supervised embedding (Sentence-BERT) on Goodreads

We followed the same general procedure described in Section 5.1, aside from the dataset specific details below. Goodreads spoiler detection [46] had 17.67 million sentences with binary spoiler annotations. Spoilers made up 3.224% of the data, making them much more common than the rare concepts we evaluated in the other datasets. Following Wan et al. [46], we used 3.53 million sentences for testing (20%), 10,000 sentences as the validation set, and the remaining 14.13 million sentences as the unlabeled pool. We also switched to the area under the ROC curve (AUC) as our primary evaluation metric for active learning to be consistent with Wan et al. [46]. For G_z, we used a pre-trained Sentence-BERT model (SBERT-NLI-base) [36], applied PCA whitening to reduce the dimension to 256, and performed l^2 normalization.

Active search

SEALS achieved the same recall as the baseline approaches, but only considered less than 1% of the unlabeled data in the candidate pool, as shown in Figure 17. At a labeling budget of 2,000, MLP-ALL and MLP-SEALS recalled $0.15 \pm 0.02\%$ and $0.17 \pm 0.05\%$, respectively, while MaxEnt-All and MaxEnt-SEALS achieved $0.14 \pm 0.04\%$ and $0.11 \pm 0.06\%$ recall respectively. Increasing the labeling budget to 50,000 examples, increased recall to ~3.7% for MaxEnt and MLP but maintained a similar relative improvement over random sampling, as shown in Figure 18. ID-SEALS performed worse than the other strategies. However, all of the active selection strategies outperformed random sampling by up to an order of magnitude.

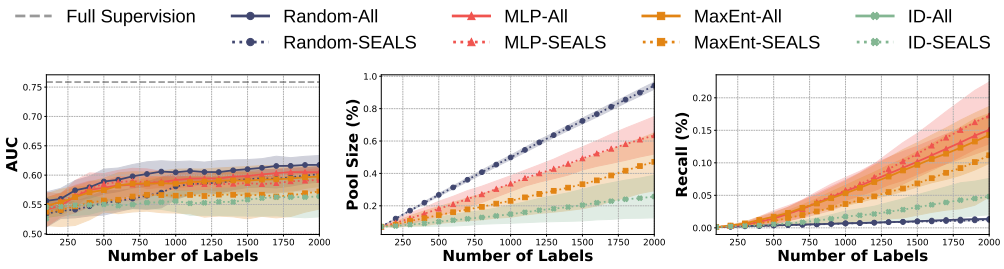


Figure 17: Active learning and search on Goodreads with Sentence-BERT embeddings. Across datasets and strategies, SEALS with $k = 100$ performs similarly to the baseline approach in terms of both the error the model achieves for active learning (left) and the recall of positive examples for active search (right), while only considering a fraction of the data U (middle).

Active learning

At a labeling budget of 2,000 examples, all the selection strategies were indistinguishable from random sampling. Increasing the labeling budget did not help, as shown in Figure 18. Unlike ImageNet and OpenImages, Goodreads had a much higher fraction of positive examples (3.224%), and the examples were not tightly clustered as described in Section 6.

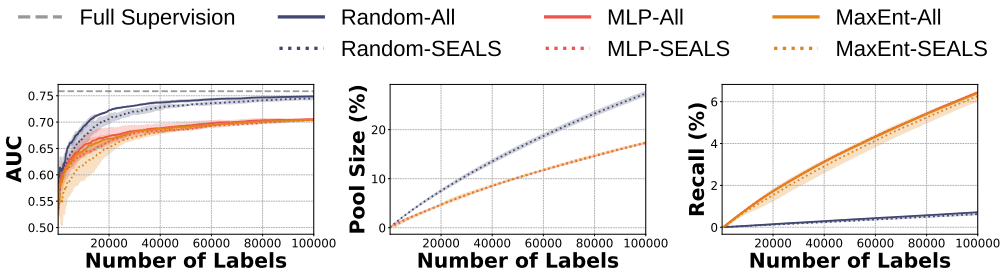


Figure 18: Active learning and search on Goodreads with a labeling budget of 100,000 examples. Across strategies, SEALS with $k = 100$ performed similarly to the baseline approach in terms of both the error the model achieved for active learning (left) and the recall of positive examples for active search (right), while only considering a fraction of the data U (middle). ID was excluded because of the growing pool size and computation. For active search, MaxEnt and MLP continued to improve recall. For active learning, all the selection strategies (both with and without SEALS) performed worse than random sampling despite the larger labeling budget. This gap was likely due to spoilers being book specific and the higher fraction of positive examples in the unlabeled pool, causing relevant examples to be spread almost uniformly across the space (see Section 6).

Latent structure

The large number of positive examples in the Goodreads dataset limited the analysis we could perform. We could only calculate the size of the largest connected component in the nearest neighbor graph (Figure 19). For $k = 10$, only 28.4% of the positive examples could be reached directly, but increasing k to 100 improved that dramatically to 96.7%. For such a large connected component, one might have expected active learning to perform better in Section 6. By analyzing the embeddings, however, we found that examples are spread almost uniformly across the space with an average cosine similarity of 0.004. For comparison, the average cosine similarity for concepts in ImageNet and OpenImages was 0.453 ± 0.077 and 0.361 ± 0.105 respectively. This uniformity was likely due to the higher fraction of positive examples and spoilers being book specific while Sentence-BERT is trained on generic data. As a result, even if spoilers were tightly clustered within each book, the books were spread across a range of topics and consequently across the embedding space, illustrating a limitation and opportunity for future work.

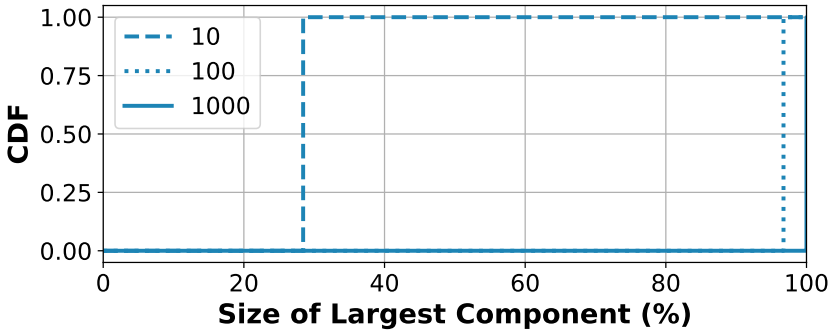


Figure 19: Cumulative distribution function (CDF) for the largest connected component in the Goodreads dataset with varying values of k.