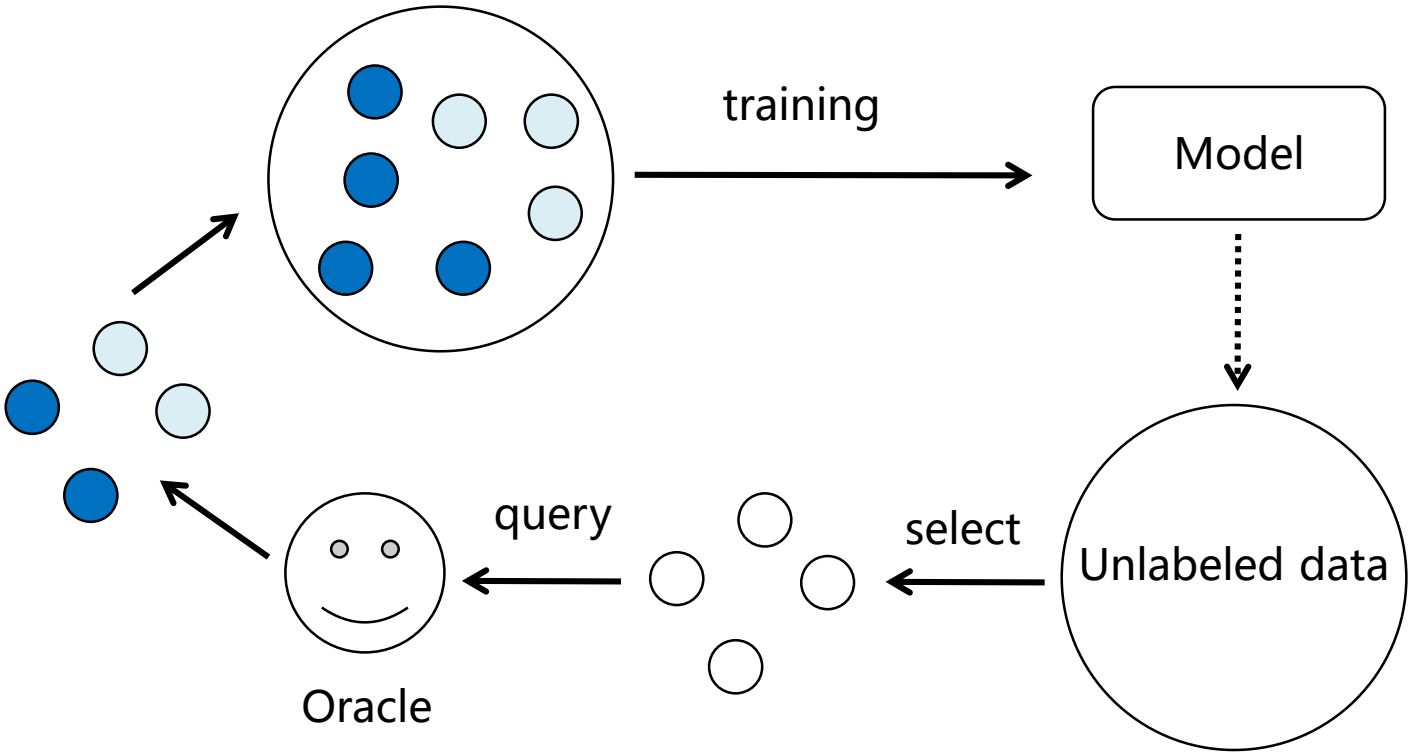
PARNEC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

Towards Understanding the Behaviors of Optimal Deep Active Learning Algorithms

Yilun Zhou* Adithya Renduchintala† Xian Li†
Sida Wang† Yashar Mehdad† Asish Ghoshal†
*MIT CSAIL †Facebook AI

AISTATS 2021

Introduction



Goal: query less for more.

Main Contributions

- **Propose a method to find the optimal AL algorithm**

  Convert the data selection problem into the data permutation form, and search for the optimal order with Simulated Annealing algorithm and a large labeled validation set.

- **Concrete experiments on deep learning tasks**

  Compare the optimal query strategy and heuristic methods with different tasks, networks, stochasticity.

- **Useful insights for both heuristic and optimal AL methods**

  i) AL works well in deep learning. ii) training stochasticity tends to negatively affect the AL performance. iii) the optimal method transfers better than heuristics (across different networks) iv) representativeness is very important for deep learning.

Data Labeling Order

- **Different query strategies lead to different order**

AL \ iter	1	2	3	4	5	6
Uncertainty	36	20	5	98	74	72
MMD	1	19	40	29	22	67
BALD	99	24	65	38	20	93
Coreset	3	41	52	32	83	20

Goal: find a permutation of unlabeled data which leads to the highest average performance.

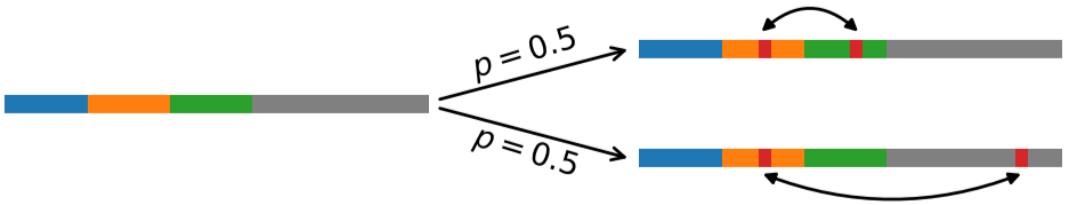
Simulated Annealing

- The space of all labeling orders is prohibitive
- The order space is discrete

✓ **Apply Simulated Annealing Search method**

✓ **Introduce an extra validation set**

$$P = \begin{cases} 1, & E_{t+1} < E_t \\ e^{\frac{-(E_{t+1}-E_t)}{kT}}, & E_{t+1} \geq E_t \end{cases}$$



Either swaps two data points from two different batches

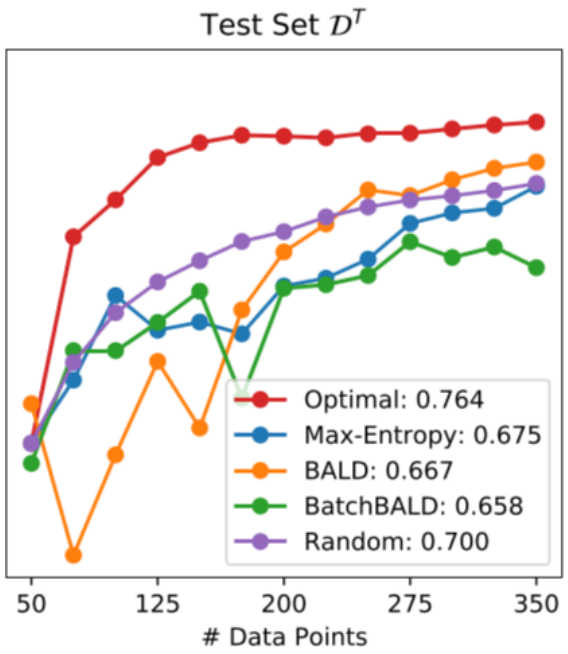
Or replaces a data point in the first K batches with one outside

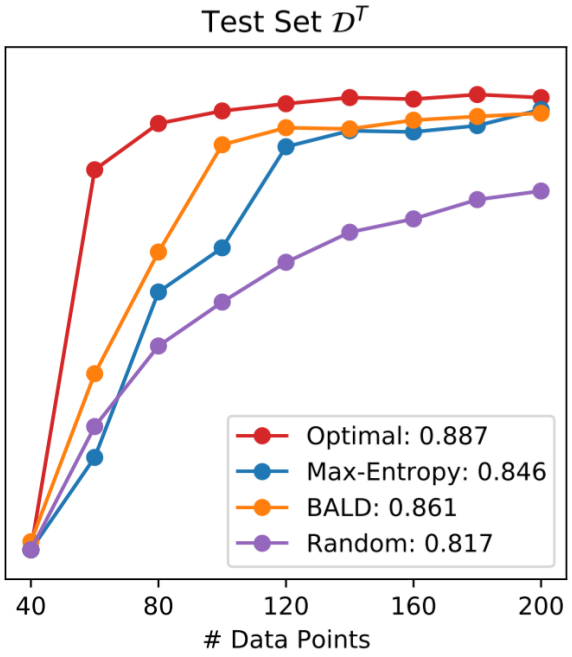
Experiments

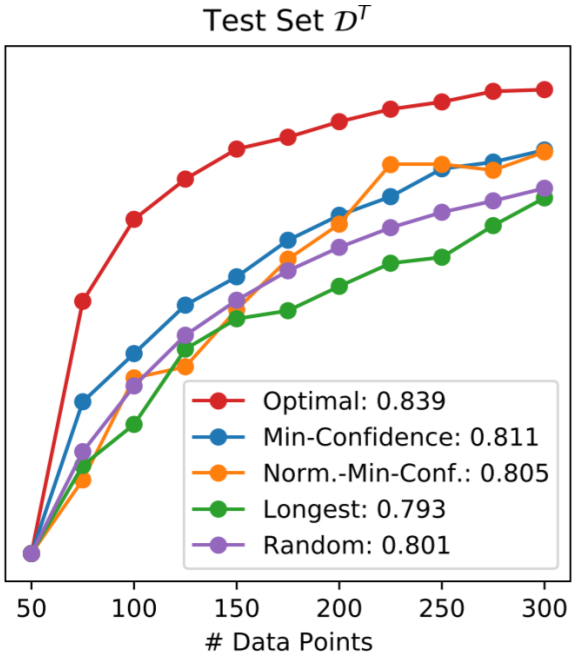
Optimal Deep Active Learning Behaviors

Task	Type	Dataset	$\lvert\mathcal{D}^U\rvert, \lvert\mathcal{D}_0^L\rvert, \lvert\mathcal{D}^M\rvert, \lvert\mathcal{D}^V\rvert, \lvert\mathcal{D}^T\rvert$	$B,\ K$	Metric	Architecture	Heuristics
OC	class.	Fashion-MNIST	2000, 50, 150, 4000, 4000	25, 12	Acc	CNN	Max-Ent., (Batch)BALD
IC	class.	TOPv2 (alarm)	800, 40, 100, 4000, 4000	20, 8	F1	LSTM, CNN, AOE, RoBERTa	Max-Ent., BALD
NER	tagging	MIT Restaurant	1000, 50, 200, 3000, 3000	25, 10	F1	LSTM	(Norm.-)Min-Conf., Longest

Table 1: Summary of experiment settings. Architecture details are in App. C.







Experiments

Effect of Training Stochasticity

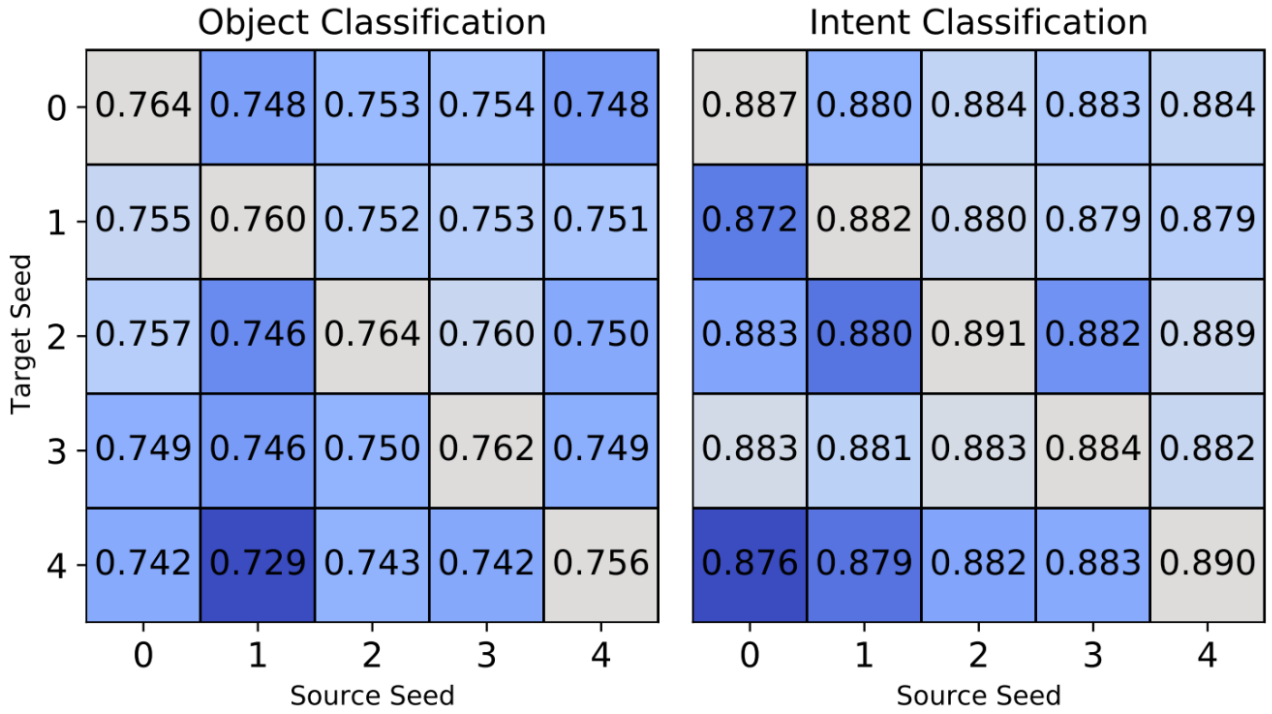


Figure 4: Training stochasticity negatively affects the optimal quality. The number in each cell represents $q_\xi(\widehat{\sigma}_{\xi'})$, with ξ on the row and ξ' on the column. The color represents $q_\xi(\widehat{\sigma}_{\xi'}) - q_\xi(\widehat{\sigma}_\xi)$, with darker colors indicating larger gaps.

Experiments

Model Transfer Quality （on IC task）

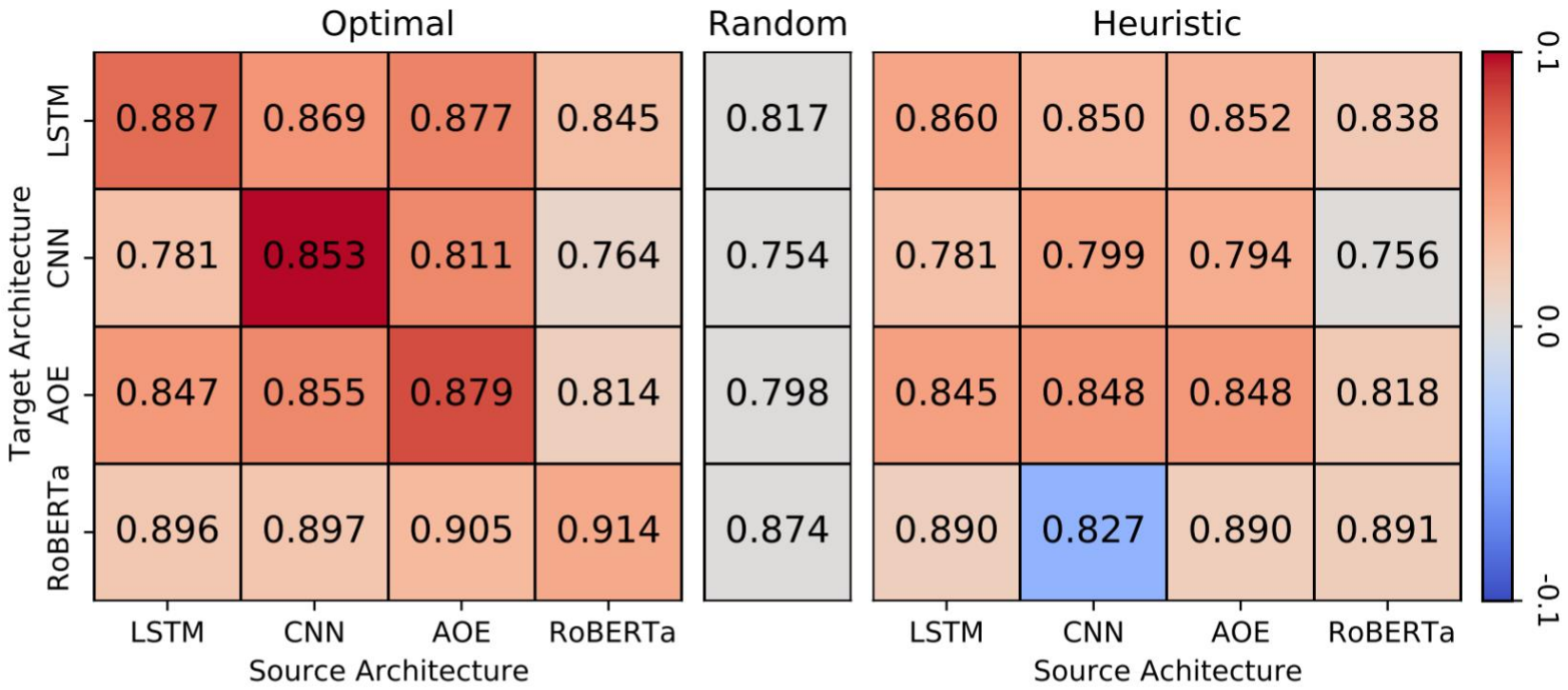


Figure 5: Quality of optimal and max-entropy heuristic order across different architectures. Cell color represents difference to the random baseline.

Experiments

Model Transfer Quality (number of same data the first 160 data)

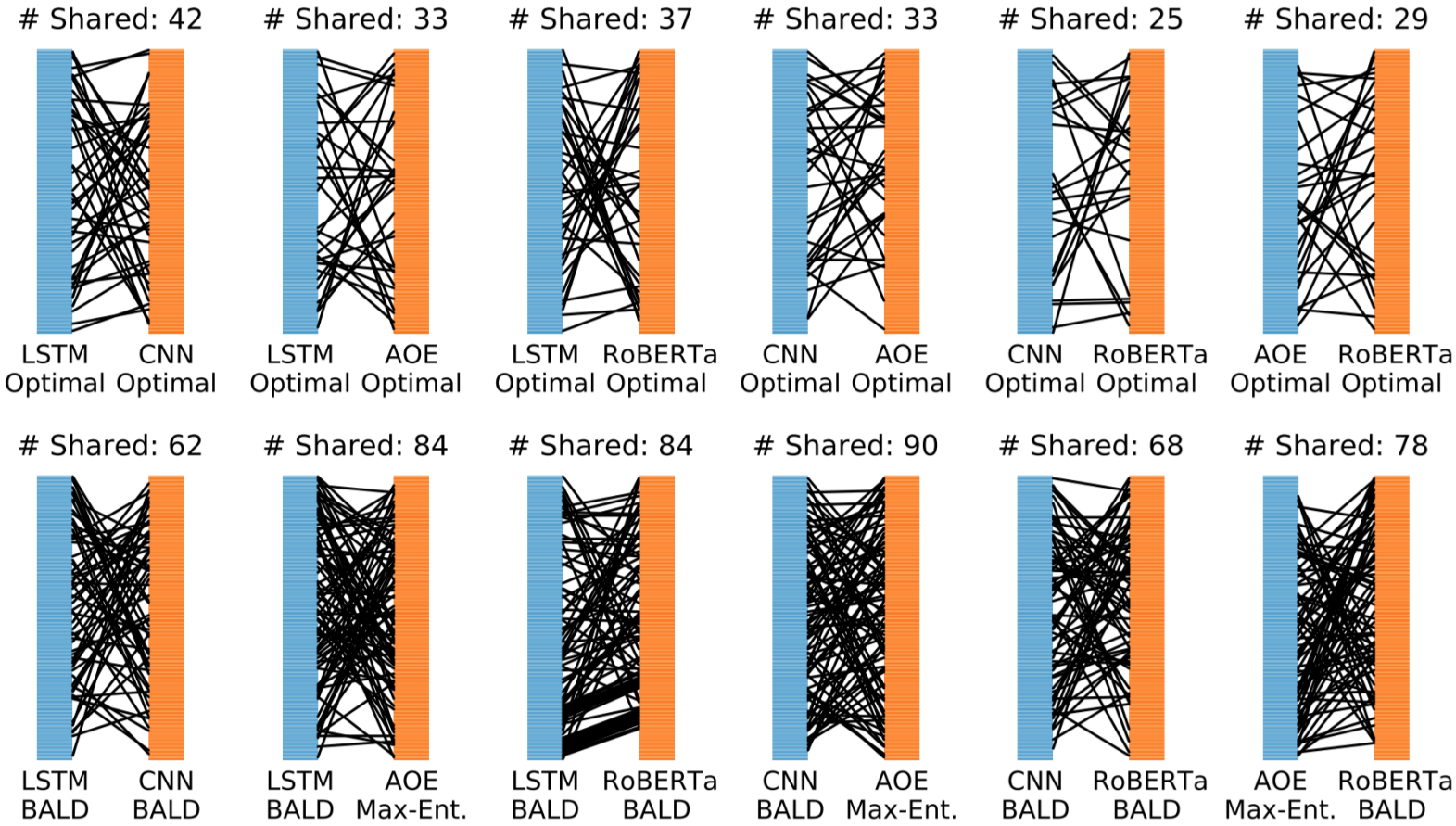
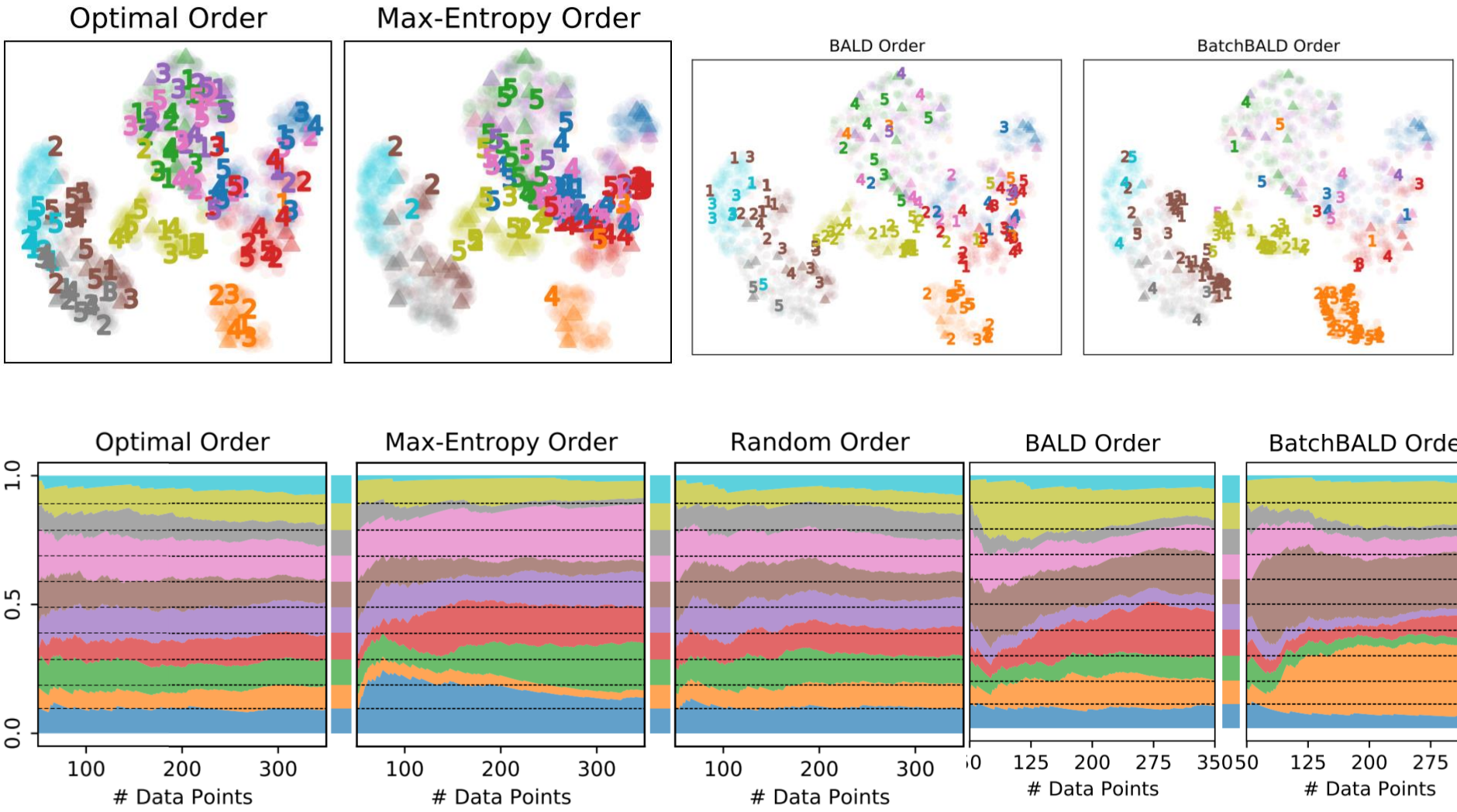


Figure 6: A visual comparison of optimal (top) and heuristic (bottom) orders, for every pair of architectures, showing the number of shared data points acquired under both architectures and their relative ranking.

Experiments

Distributional Characteristics （OC task）



Experiments

Distributional Characteristics （IC task）

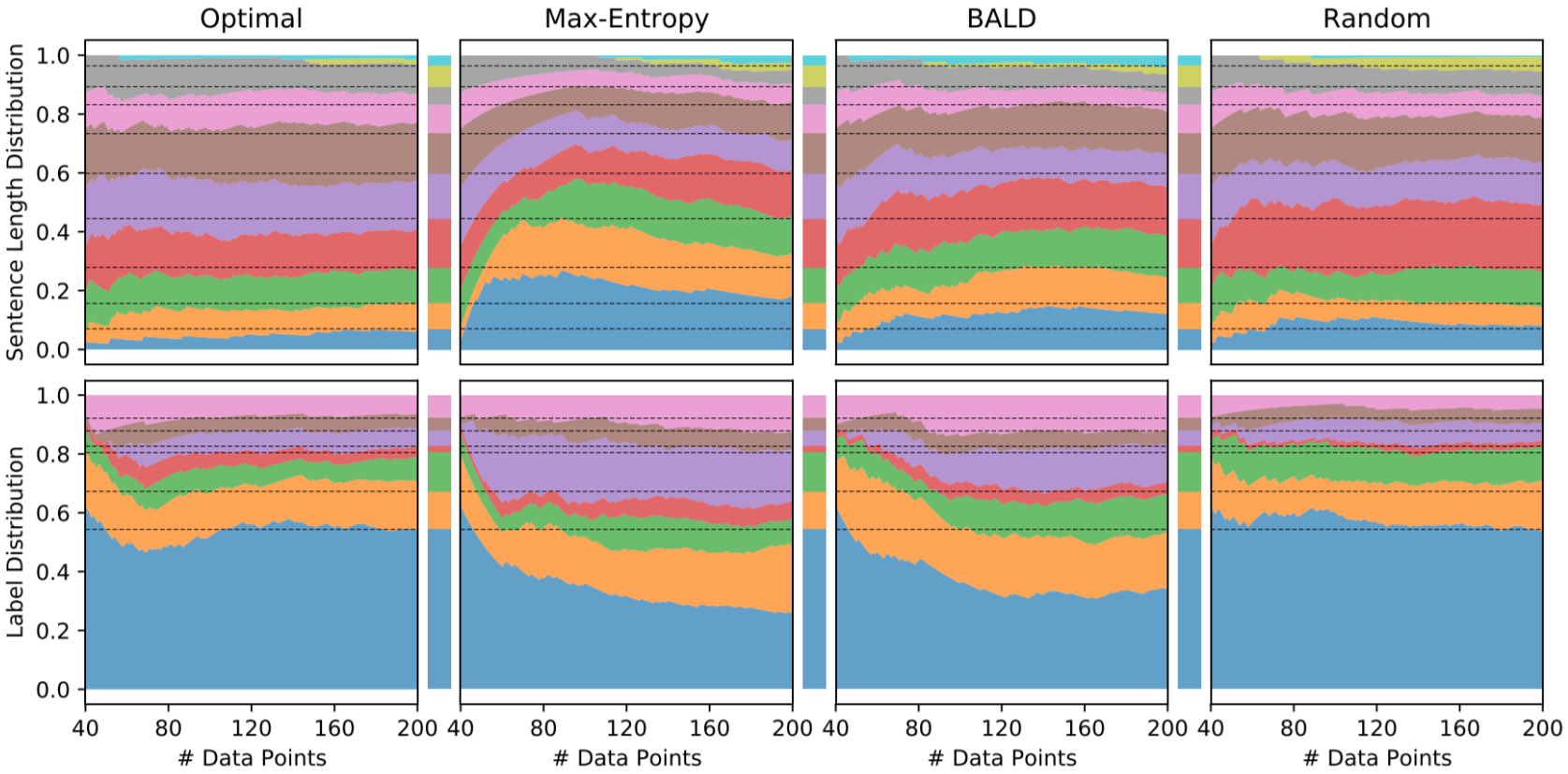
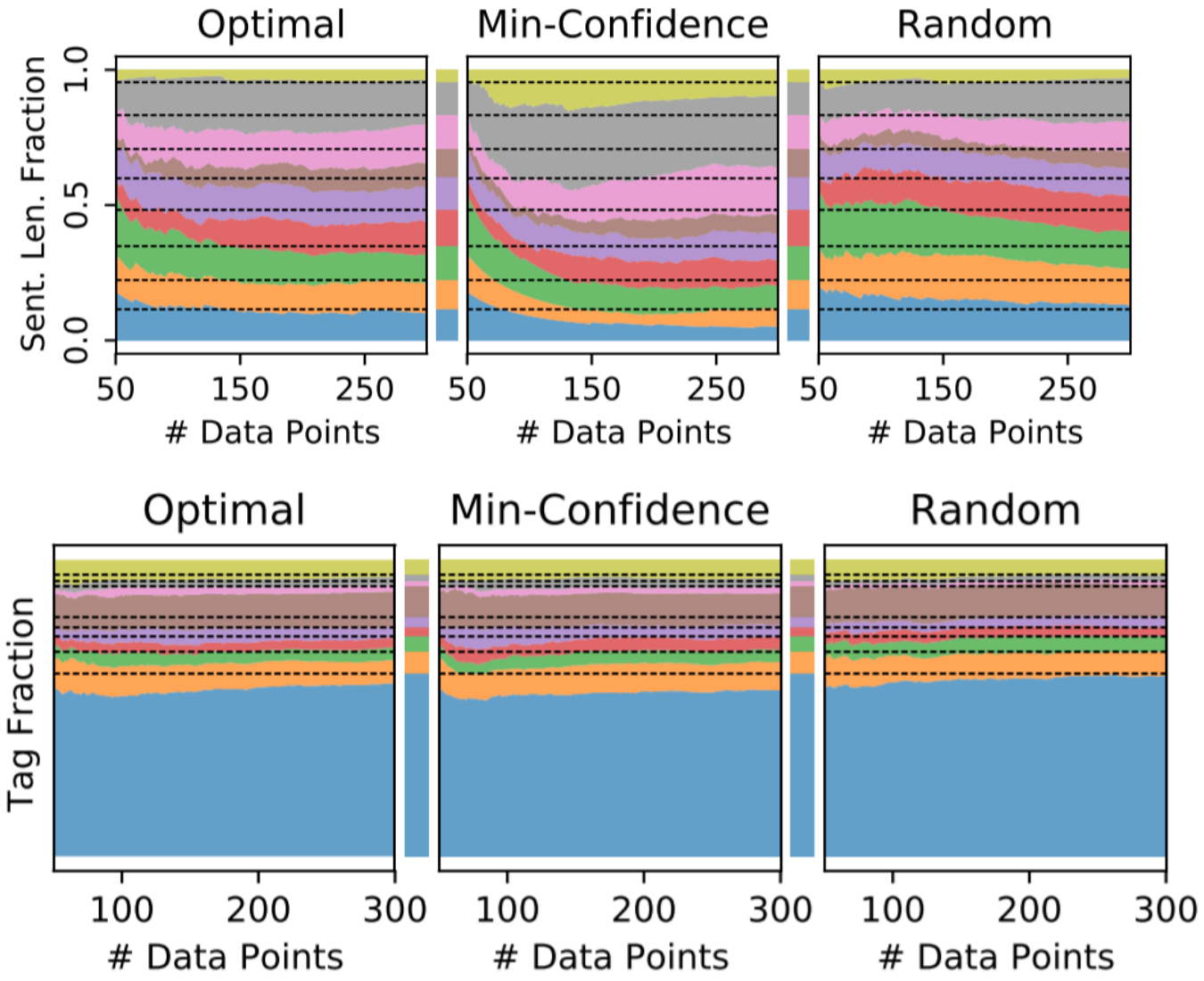


Figure 8: Sentence length and label distribution w.r.t. labeled set size for intent classification. For sentence lengths, the colors represent, from bottom to top: 1–3, 4, 5, 6, 7, 8, 9, 10, 11–12, and 13+. For labels, the colors represent, from bottom to top: `create-alarm`, `get-alarm`, `delete-alarm`, `other`, `silence-alarm`, `snooze-alarm`, and `update-alarm`.

Experiments

Distributional Characteristics （NER task）



Experiments

Distributional matching regularization

Algorithm 2: Input Dist.-Matching Reg. (IDMR)

Input: $\mathcal{A}\left(m_\theta, \mathcal{D}^L, \mathcal{D}^U\right)$ that returns the next data point in $\mathcal{D}^U$ to label

$d_{\text{ref}} = \texttt{bin-distribution}\left(\mathcal{D}^L_{0,X} \cup \mathcal{D}^U_X \cup \mathcal{D}^M\right);$

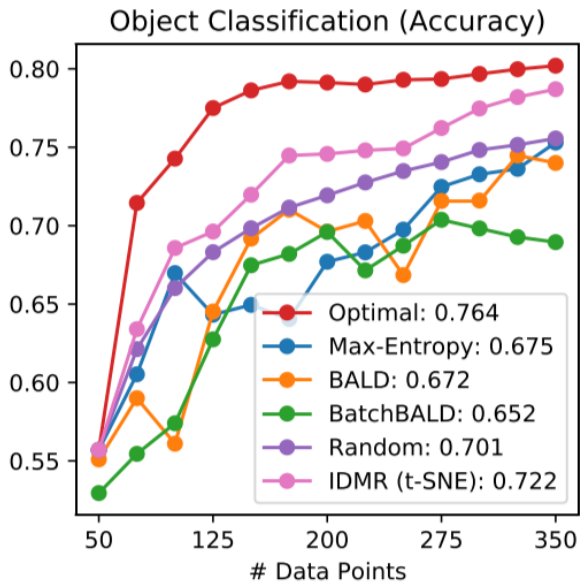
$d_{\text{cur}} = \texttt{bin-distribution}\left(\mathcal{D}^L_X\right);$

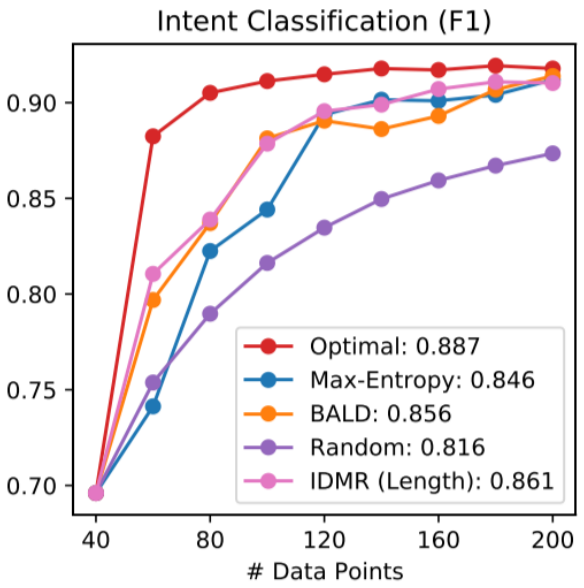
$b^* = \arg\min_b \left(d_{\text{cur}} - d_{\text{ref}}\right)_b;$

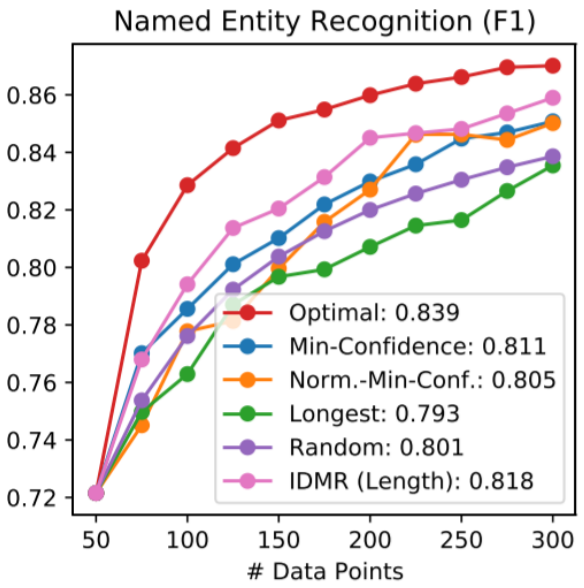
$\mathcal{D}^U_{b^*} = \{x \in \mathcal{D}^U_X : \texttt{bin}(x) = b^*\};$

return $\mathcal{A}\left(m_\theta, \mathcal{D}^L, \mathcal{D}^U_{b^*}\right)$

For OC, a K-means algorithm identified five clusters in the t-SNE space, which are used as the five bins. For IC and NER, sentences are binned by length.







Conclusion

- This paper proposes a Simulated Annealing Search method to obtain the optimal query queue.
- Optimal strategy is model-dependent.
- Optimal strategy is more transferable than heuristics.
- Optimal strategy suggests to matches the data distribution
- Existing heuristics can be further improved by a distributional matching regularization

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

THANKS