frontiers | Frontiers in Artificial Intelligence

TYPE Brief Research Report
PUBLISHED 26 October 2022
DOI 10.3389/frai.2022.939967

OPEN ACCESS

EDITED BY
Leslie N. Smith,
Navy Center for Applied Research in AI, United States

REVIEWED BY
Tereza Cristina Cardoso,
Universidade Estadual de São Paulo, Brazil
Luis Rato,
University of Evora, Portugal

*CORRESPONDENCE
Vitor Filipe
vfilipe@utad.pt

SPECIALTY SECTION
This article was submitted to Machine Learning and Artificial Intelligence, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 09 May 2022
ACCEPTED 07 October 2022
PUBLISHED 26 October 2022

CITATION
Moreira da Silva DE, Gonçalves L, Franco-Gonçalo P, Colaço B, Alves-Pimenta S, Ginja M, Ferreira M and Filipe V (2022) Active learning for data efficient semantic segmentation of canine bones in radiographs. *Front. Artif. Intell.* 5:939967. doi: 10.3389/frai.2022.939967

Active learning for data efficient semantic segmentation of canine bones in radiographs

D. E. Moreira da Silva[1], Lio Gonçalves[1,2], Pedro Franco-Gonçalo[3,4], Bruno Colaço[4,5,6], Sofia Alves-Pimenta[4,5,6], Mário Ginja[3,4,6], Manuel Ferreira[7] and Vitor Filipe[1,2]*

[1]School of Science and Technology, University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal, [2]INESC Technology and Science (INESC TEC), Porto, Portugal, [3]Department of Veterinary Science, UTAD, Vila Real, Portugal, [4]Veterinary and Animal Research Centre (CECAV), Vila Real, Portugal, [5]Department of Animal Science, UTAD, Vila Real, Portugal, [6]Centre for the Research and Technology of Agro-Environmental and Biological Sciences (CITAB), Vila Real, Portugal, [7]Neadvance Machine Vision SA, Braga, Portugal

X-ray bone semantic segmentation is one crucial task in medical imaging. Due to deep learning's emergence, it was possible to build high-precision models. However, these models require a large quantity of annotated data. Furthermore, semantic segmentation requires pixel-wise labeling, thus being a highly time-consuming task. In the case of hip joints, there is still a need for increased anatomic knowledge due to the intrinsic nature of the femur and acetabulum. Active learning aims to maximize the model's performance with the least possible amount of data. In this work, we propose and compare the use of different queries, including uncertainty and diversity-based queries. Our results show that the proposed methods permit state-of-the-art performance using only 81.02% of the data, with $\mathcal{O}(1)$ time complexity.

KEYWORDS

X-ray image analysis, deep learning, active learning, cluster based sampling, representative sampling, Monte Carlo Dropout sampling, Shannon's entropy

1. Introduction

Image segmentation is one of the most important yet, challenging tasks in medical image analysis (Shah and Sharma, 2018). In recent years, Deep Learning's (DL) emergence has made it possible to build models that achieve human-like or superior performance in many medical imaging tasks, such as segmentation. However, the DL shortcoming is the need for large quantities of annotated data, often in the order of thousands (Ronneberger et al., 2015). Additionally, gathering such large and high-quality datasets, annotated by medical experts, is often very difficult (Kim et al., 2019; Nguyen et al., 2021) because segmentation requires thorough pixel-wise labeling, hence being a highly time-intensive procedure (Ozdemir et al., 2021). Additionally, collecting medical images might be financially expensive (Kim et al., 2019).

Active Learning's (AL) goal is to identify which unlabeled samples are the most interesting to be labeled by a human expert. In other words, how to maximize the model's performance using the least possible amount of data (Ren et al., 2020). Without AL, this sampling is purely random, which might cause data redundancy. Hence, the use of AL enables the annotation process to be as time and financially efficient as possible.

In this work, we explore the use of AL methodologies in the context of the segmentation of dogs' femur and acetabulum bones in X-ray images. Dogs' coxofemoral joint radiographic examinations are used worldwide for screening hip dysplasia and to select better animals for breeding. Radiographic evaluation is performed by human observation, being considered a time-consuming, expensive and relatively subjective process due to differences in classification between evaluators. Precise segmentation of these bone structures is noteworthy as it allows further automated diagnosis of canine hip dysplasia (Moreira da Silva et al., 2021, 2022). However, the joint regions present high noise, low contrast, overlapping tissue, and a narrow gap between the femur and acetabulum (Lianghui et al., 2019). As such, the annotation of these regions requires increased attention, a greater level of medical specialization and knowledge of these anatomical structures. Consequently, this set of factors leads to an increased expenditure of veterinary medicine professionals' valuable time in the annotation process. Therefore, we aim to assemble and compare the effects of different AL queries to build a high performant U-Net model with low amounts of annotated data. The developed techniques will be integrated into the Dys4Vet[1] web platform, a dedicated software for the automated canine hip dysplasia diagnosis.

The rest of this paper is organized as follows: Related work (Section 2); Methods (Section 3); Results and discussion (Section 4); Conclusions (Section 5).

2. Related work

There are two types of AL queries (Munro, 2021): uncertainty sampling; diversity sampling. The first aims to fix the model's currently known confusion by sampling data that the model presents low predictive confidence. Diversity sampling intends to provide the model with samples of unknown areas of the feature space, thus narrowing the model's knowledge gap. Mahapatra et al. (2018) used AL for X-Ray lung segmentation. The authors suggest generative adversarial networks (GANs) to generate diverse images. Then, using a Bayesian Neural Network, each generated sample's informativeness is calculated. The informativeness is calculated through the combination of epistemic and aleatoric uncertainties (Kendall and Gal, 2017). The most informative samples are added, at each iteration, to the labeled training data. With this method, the authors reach state-of-the-art performance by using only 35% of the full annotated dataset. Ozdemir et al. (2018) used Monte Carlo Dropout (Gal and Ghahramani, 2016) (MCD) to measure sample uncertainty based on inference variance. Then, content distance (Gatys et al., 2016) and layer entropy maximization are used to measure representativeness. The novelty of this work is that instead of applying uncertainty and then sampling for diversity, the authors propose a Borda count approach: samples are ranked for each metric, and sampling is carried out based on combined rank. On a similar note, in later works, Ozdemir et al. (2021) proposed a modification of MCD (Gal and Ghahramani, 2016), where instead of randomly dropping neuron connections, entire convolutional kernels are dropped. The uncertainty of each sample is calculated by averaging each pixel's variance over the multiple inferences. Additionally, a variational autoencoder is used to project gaussian distributions of labeled and unlabeled pools. With both distributions, the authors calculate underrepresented samples in the labeled dataset. By combining uncertainty and representativeness, the authors stay within 2% of the state-of-the-art performance using only 25% of the data. Zhao et al. (2020) modified a U-Net to extract and then upscale segmentation maps from deep and intermediate layers. Then, the authors calculate the dice coefficient between the model's final segmentation map and each upscaled map. The samples with the highest dice scores' average are selected to be labeled. This technique achieves comparable state-of-the-art performance with 43.16% of the data. However, the results do not differ much from random sampling, with a performance difference of $<1\%$. Zhao et al. (2021) introduced DSAL through the reuse of the previously described technique. While high uncertainty samples are annotated by a human expert, in this work the samples with low uncertainty are also provided to weak labelers (i.e., dense conditional random fields) to generate pseudo labels. The authors state the incorporation of pseudo labels further boosts the results. Jin et al. (2022) proposed a one-shot active learning framework based on contrastive learning and diversity sampling. First, contrastive learning is used for feature extraction. Then, this new feature space is clustered using K-Means++ (Arthur and Vassilvitskii, 2007), and sampling is performed using farthest point sampling (FPS). While clustering guarantees inter-cluster diversity, FPS provides intra-cluster diversity. This method was validated in three different datasets, and it delivered dice score gains when compared to others.

3. Methods

This section describes the employed methodologies for the present study. Initially, the used dataset is presented in Section 3.1. Then, Section 3.2 describes the DL segmentation model used for the experiments. Lastly, Section 3.3 details the AL procedure and Section 3.4 details the proposed AL queries.

1 https://www.citab.utad.pt/projects/780/show

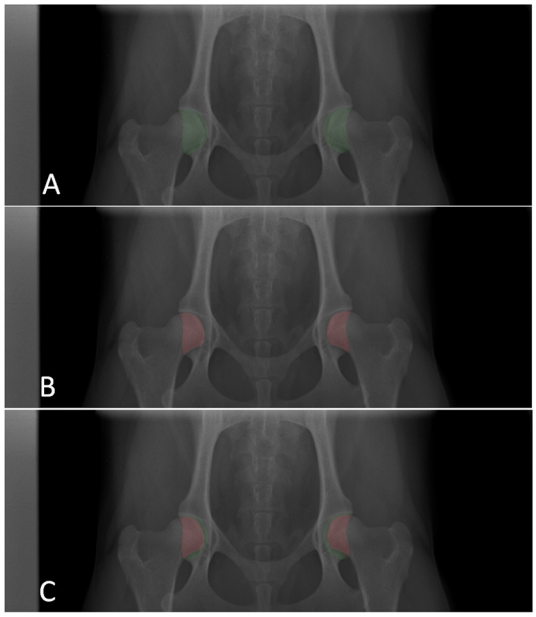

FIGURE 1
Example of annotations made for each image. **(A)** Acetabulum annotation. **(B)** Femoral head (that overlaps the acetabulum) annotation. **(C)** Combined femoral head and acetabulum annotations. The X-ray was cut on the y axis for better visualization. For model training, full images are used.

3.1. Dataset

For this work, DICOM images were collected from Veterinary Teaching Hospital of the University of Trás-os-Montes and Alto Douro and from the Danish Kennel Club, totaling 202 images. Please note that each image corresponds to a unique patient, avoiding data correlation in subsequent splitting processes. Then, manual annotation was carried out for every DICOM. In detail, the acetabulum and femoral head acetabulum intersection were annotated (Figure 1). With these annotations, three-channel masks were generated, where each channel is a binary mask for each class: background, femur, and acetabulum. Then the images were resized to 544 × 448. The masks underwent the same resizing through nearest neighbor interpolation. Finally, a test (15%) and a validation set (15%) are created, which remain constant throughout the AL cycles. Additionally, 3% is used as initial training data $\mathcal{L}_0$ and the remaining as the initial unlabeled pool $\mathcal{U}_0$.

3.2. Segmentation model

The DL segmentation model we use is the same we propose in previous works (Moreira da Silva et al., 2022), a U-Net with EfficientNet (Tan and Le, 2019) modules as the feature-extractor backbone.

For quantitative results we evaluated the dice score (Equation 1), as it is the common metric in medical image segmentation (Siddique et al., 2021).

$$Dice(P, G) = \frac{2 \times |G \cap P|}{|G| + |P|} \tag{1}$$

where

P: Predicted Segmentation
G : Ground Truth

We also use the same loss function of the previous work (Moreira da Silva et al., 2022), a combination of dice and focal loss (Equation 2). For this work, we set $\alpha = 0.25$ and $\gamma = 2$.

$$L(P, G) = (1 - Dice(P, G)) - \alpha(1 - P)^{\gamma} log(P) \tag{2}$$

3.3. Active learning procedure

In this section, we present and explain our AL procedures. We define the unlabeled pool as $\mathcal{U} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ is the available feature space to sample from, and $\mathcal{Y}$ the corresponding labels. Please note that we have $\mathcal{Y}$ because we generate the unlabeled pool artificially, as detailed in Section 3.1. In real-world AL, $\mathcal{Y}$ would not be present, so the labels would need to be provided by an oracle (i.e., human expert) in real-time.

Our AL cycle is formally defined as follows: Given the initial training data $\mathcal{L}_0$, and the initial unlabeled pool $\mathcal{U}_0$, at each AL iteration t, a given acquisition query Q will sample n images from $\mathcal{U}_t$ and then this new subset $\mathcal{U}_t^n \subseteq \mathcal{U}_t$ is added to the training dataset $\mathcal{L}_t = \mathcal{L}_{t-1} \cup \mathcal{U}_t^n$, removed from the next iteration's unlabeled pool $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \mathcal{U}_t^n$, and the model is re-trained. After training, the model is evaluated on the test dataset, and the resulting metrics are saved. The iteration is incremented $t = t + 1$, and this process repeats until the unlabeled pool is empty $\mathcal{U} = \emptyset$. For each query $\mathcal{Q}$ under study, this entire cycle is repeated 10 times. Later, we will provide each metric's average at each AL iteration t, for each query $\mathcal{Q}$. This way our results are more statistically significant.

3.4. Active learning queries

For this study, we employ and compare five different queries:

3.4.1. Random sampling

Randomly sampling n elements from the unlabeled pool $\mathcal{U}^n \in_R \mathcal{U}$. This query serves as a baseline.

3.4.2. Cluster based sampling

First, the unlabeled pool $\mathcal{U}$ is normalized according to (3):

$$\mathcal{U}_N = \left\{ \frac{x - \mathcal{U}_{min}}{\mathcal{U}_{max} - \mathcal{U}_{min}} : x \in \mathcal{U} \right\} \tag{3}$$

Then the dimensionality of $\mathcal{U}_N$ is reduced by applying Principal Component Analysis (PCA) (Pearson, 1901), with 99% explained data variance. We define the reduced pool as $\mathcal{U}_R$. Then, we use K-Means++ (Arthur and Vassilvitskii, 2007) to create n clusters inside $\mathcal{U}_R$. Then, for each centroid of $\mathcal{U}_\mathcal{R}$, the closest element is found. The corresponding n elements in $\mathcal{U}$ are sampled. Thus, this query ensures maximum diversity between the samples.

3.4.3. Representative sampling

This query requires both the unlabeled pool $\mathcal{U}$ and the training data $\mathcal{L}$. We then apply the same two initial steps of the previous query to both $\mathcal{U}$ and $\mathcal{L}$. Then, also using K-Means++, two clusters $\mathcal{C}$ are created: a training cluster $\mathcal{C}_\mathcal{L}$, and an unlabeled cluster $\mathcal{C}_\mathcal{U}$. For each item in $\mathcal{C}_\mathcal{U}$ it's representativeness $\mathcal{R}$ is calculated according to (4):

$$\mathcal{R} = \left\{ d(x, \mu_{\mathcal{C}_\mathcal{L}}) - d(x, \mu_{\mathcal{C}_\mathcal{U}}) : x \in \mathcal{C}_\mathcal{U} \right\} \tag{4}$$

where

μ: Cluster centroid

d: Euclidean distance

$\mathcal{R}$ measures the difference between the training and the unlabeled data. The first n elements of $\mathcal{U}$ that have the highest corresponding $\mathcal{R}$ values are sampled. In short, we select the items that best represent the unlabeled pool population and look the most different from the current training data.

3.4.4. Monte Carlo Dropout sampling

Monte Carlo Dropout (MCD) is a Bayesian ANN approximation of the Gaussian process, introduced by Gal and Ghahramani (2016). It uses the dropout layers (Srivastava et al., 2014) of a DL model to measure its predictive uncertainty. It works by turning on dropout during inference, and by running inference k times, each dropout configuration corresponds to a Monte Carlo sample from the available models' space. Thus, we obtain a predictive distribution enabling the inspection of the predictive uncertainty. For this study we set $k = 30$. Then we obtain each sample's average prediction $\overline{\mathcal{P}}$ according to (5):

$$\overline{\mathcal{P}} = \left\{ \frac{1}{k} \sum_{i=1}^{k} F(x) : x \in \mathcal{U} \right\} \tag{5}$$

where

F: DL model

Afterwards, each sample's uncertainty is calculated using Shannon's entropy (Shannon, 1948), according to Equation (6). Since the used model uses a sigmoid activation function, we calculate the entropy for the femur (output channel 1) and the acetabulum (output channel 2) class separately, averaging them thereafter.

$$\mathcal{E} = \left\{ \frac{1}{2} \sum_{c=1}^{2} \left[\frac{1}{H \times W} \sum_{h=0}^{H} \sum_{w=0}^{W} (-x_{hwc} \, log_2 \, x_{hwc}) \right] : x \in \overline{\mathcal{P}} \right\} \tag{6}$$

The first n elements of $\mathcal{U}$ that have the highest corresponding $\mathcal{E}$ values are sampled. As such, we select the items that the model presents a higher level of uncertainty. We call this method CWE-MCD (Class-wise Entropy Monte Carlo Dropout).

3.4.5. Representative CWE-MCD

All the previously mentioned queries sample from one of the following perspectives: uncertainty or diversity. This method proposes a combination of two techniques: Representative Sampling and CWE-MCD. To combine both queries, we adopt the Ozdemir et al. (2018)'s Borda count approach. In short, we separately calculate the $\mathcal{R}$ and $\mathcal{E}$ scores for each image in $\mathcal{U}$. Then each unlabeled image is ranked based on how high each score is. The final sampling is based on the 15 highest combined rank.

4. Results and discussion

Initially ($t = 0$), the labeled dataset $\mathcal{L}_0$ has six train images and the unlabeled pool $\mathcal{U}_0$ has 131 images, and at each active learning iteration, a given query $\mathcal{Q}$ will select 15 images ($n = 15$). We repeat this procedure until we run out of images, resulting in nine AL iterations ($t_{max} = \left\lceil \frac{131}{15} \right\rceil = 9$). Regarding the model, at each AL iteration, we train a model from scratch, using the Adam optimizer with a learning rate of $1e-2$ and a batch size of eight. Additionally, we use two callbacks that monitor the validation's data dice score. The first is a callback that reduces the learning rate by a factor of 0.1 if the metric does not improve after eight epochs. Secondly, a callback that halts the training if the metric does not improve by ten epochs (early stopping). Therefore, we set the number of epochs to 500.

Also, we train the U-Net with the entire unlabeled pool combined with the initial training data ($\mathcal{L}_0 \cup \mathcal{U}$), achieving a test dice score of 0.95. We denote this value as the model's Upper Bound (UB). The average dice score, for each AL iteration, for the proposed methods are described in Table 1, and illustrated in Figure 2. Noticeably, every query (Clustering, Representative, CWE-MCD, Representative CWE-MCD) performed better than the baseline (Random) in all iterations. Despite clustering being a simple technique, in the first iterations, it outperforms the baseline. As the size and diversity of the unlabeled pool

TABLE 1 Average dice score at each AL iteration for the proposed methods.

Query $\mathcal{Q}$	AL iteration t									
	0	1	2	3	4	5	6	7	8	9
Random	0.02	0.15	0.24	0.34	0.47	0.66	0.79	0.89	0.91	0.95
Clustering	0.02	0.19	0.30	0.39	0.51	0.68	0.81	0.90	0.91	0.95
Representative	0.02	0.25	0.36	0.45	0.58	0.75	0.89	0.94	0.95	0.95
CWE-MCD	0.02	0.23	0.34	0.49	0.63	0.82	0.92	0.94	0.95	0.95
Representative CWE-MCD	0.02	0.24	0.35	0.51	0.64	0.84	0.93	0.94	0.95	0.95

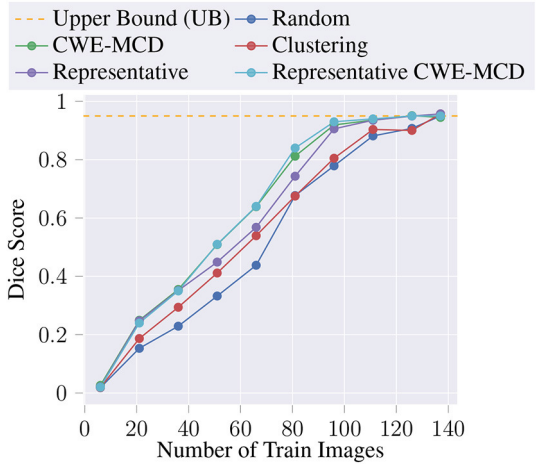

FIGURE 2
Average dice score at each AL iteration for the proposed methods.

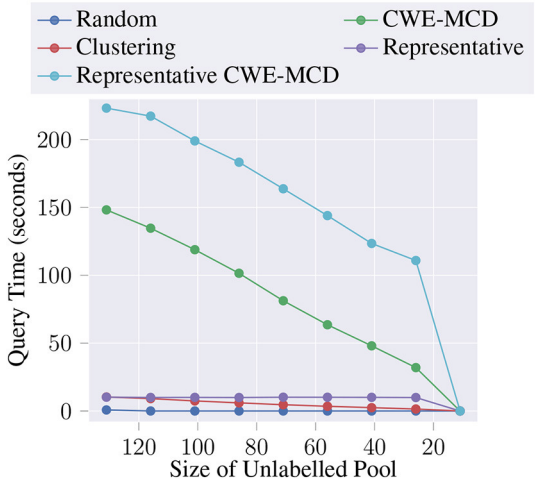

FIGURE 3
Average query time (in seconds) at each AL iteration for the proposed methods.

decreases, the performance of the clustering query becomes identical to the baseline. Both the baseline and the clustering queries need the entire unlabeled pool as training data to reach the UB. Nonetheless, clustering still proves suitable in early AL iterations. The three more advanced methods we build (Representative, CWE-MCD, Representative CWE-MCD) show significant dice score gains compared to the baseline. In the first two iterations, these present closely the same performance. From the second, CWE-MCD and Representative CWE-MCD provide superior performance. This fact can corroborate (Ozdemir et al., 2021) statement that uncertainty may not be a sufficiently calibrated metric until the training data size is adequately large. Additionally, Representative CWE-MCD provides slightly better results over CWE-MCD and Representative until the sixth iteration. This is expected, as Representative CWE-MCD combines the best aspects of the diversity and uncertainty sampling, thus being a more powerful technique (Munro, 2021). Despite this slight performance superiority, these three queries can reach the UB at around 111 training images compared to the required 137 when using the baseline or the clustering methods. This means that using these queries allowed the same level of performance with $\approx$ 81.02% of the data, a saving of $\approx$ 18.98% (26 images).

Additionally, we decided to measure the average query time in seconds, at each AL iteration, for each of the proposed queries. Figure 3 depicts the results. As expected, the baseline is almost instant. The clustering technique is also pretty fast, decreasing times as the iterations increase. One of the best performing queries, CWE-MCD, presents a significantly higher time complexity $\mathcal{O}(n)$, meaning that the required time scales linearly to the size of the unlabeled pool. In practice, this query might be unsuitable for large unlabeled pools. Thus, to offset the computation times, it might be required to tune the number of Monte Carlo samples (i.e., number of inferences). Notwithstanding, representative query presents a linear time complexity $\mathcal{O}(1)$ while still delivering noteworthy dice score gains, thus being suitable for large unlabeled pools. Lastly, Representative CWE-MCD presents the same behavior as CWE-MCD but with a significant additional time overhead due to representativeness and Borda count computations. In practice,

the use of this query will be even more limited than the CWE-MCD due to the increased times. However, its use can still be advantageous in situations that demand maximum performance, situations where time is not a constraint, or when applied to a small subset of the unlabeled pool.

5. Conclusions

In this work, we study and compare the effectiveness of different AL query strategies in the ambit of the segmentation of dogs' femur and acetabulum bones in X-ray images. In detail, we suggest measuring the uncertainty by calculating class-wise entropy using Monte Carlo Dropout. Furthermore, we propose to combine this uncertainty metric with a representativeness method, inspired by the works of Ozdemir et al. (2018). This method is superior to the others, allowing an 18.98% reduction in the amount of required annotated data. Despite this method being $\mathcal{O}(n)$ in time complexity, representative sampling is $\mathcal{O}(1)$ time complex, with comparable performance levels, thus suitable for large unlabeled datasets. For future research, we intend to study other possible combinations of the entropy and representativeness methods presented in this paper. In addition, the creation of annotation software unified with an AL framework, equipped with automatic pre-annotation capabilities, would allow further time savings for veterinary professionals.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

DM, LG, and VF contributed to conception and design of the study. PF-G, BC, SA-P, and MF organized the database. DM, LG, MF, and VF defined the methodology. DM, LG, MG, and VF performed validation and data analysis. DM wrote the first draft of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was financed by project Dys4Vet (POCI-01-0247-FEDER-046914), co-financed by the European Regional Development Fund (ERDF) through COMPETE2020—the Operational Programme for Competitiveness and Internationalization (OPCI). The authors are also grateful for all the conditions made available by FCT—Portuguese Foundation for Science and Technology, under the projects UIDB/04033/2020, UIDB/CVT/00772/2020, and Scientific Employment Stimulus—Institutional Call—CEECINST/00127/2018 UTAD.

Acknowledgments

The authors acknowledge Prof. Fintan McEvoy, University of Copenhagen who is a collaborator on this project and the Danish Kennel Club for allowing access to images from their data archive.

Conflict of interest

Author MF was employed by Neadvance Machine Vision SA.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Arthur, D., and Vassilvitskii, S. (2007). "K-means++: the advantages of careful seeding," in *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* (New Orleans, LA), 1027–1035.

Gal, Y., and Ghahramani, Z. (2016). "Dropout as a Bayesian approximation: representing model uncertainty in deep learning," in *33rd International Conference on Machine Learning, ICML 2016, Vol. 3* (New York, NY: International Machine Learning Society (IMLS)), 1651–1660.

Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV). doi: 10.1109/CVPR.2016.265

Jin, Q., Yuan, M., Qiao, Q., and Song, Z. (2022). One-shot active learning for image segmentation *via* contrastive learning and diversity-based sampling. *Knowledge-Based Syst.* 241:108278. doi: 10.1016/j.knosys.2022.108278

Kendall, A., and Gal, Y. (2017). "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems* (Long Beach, CA), 5575–5585.

Kim, M., Yun, J., Cho, Y., Shin, K., Jang, R., Bae, H. J., et al. (2019). Deep learning in medical imaging. *Neurospine* 16, 657–668. doi: 10.14245/ns.1938396.198

Lianghui, F., Gang, H. J., Yang, J., and Bin, Y. (2019). "Femur segmentation in X-ray image based on improved U-Net," in *IOP Conference Series: Materials Science and Engineering, Vol. 533* (Guangzhou). doi: 10.1088/1757-899X/533/1/012061

Mahapatra, D., Bozorgtabar, B., Thiran, J. P., and Reyes, M. (2018). "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11071* (Granada), 580–588. doi: 10.1007/978-3-030-00934-2_65

Moreira da Silva, D. E., Alves-Pimenta, S., Gonçalo, P. F., Colaço, B., Santana, A., Ferreira, M., et al. (2021). "Relationship between the hip congruence index and hip FCI categories in dogs," in *46th Word Small Animal Veterinary Association World Congress, WSAVA 2021*. Available online at: https://wsava2021.com/wp-content/uploads/sites/48/2021/11/WSAVA21-Abstracts-Nov-15-by-topic.pdf

Moreira da Silva, D. E., Filipe, V., Franco-Gonçalo, P., Colaço, B., Alves-Pimenta, S., Ginja, M., et al. (2022). "Semantic segmentation of dog's femur and acetabulum bones with deep transfer learning in X-ray images," in *Intelligent Systems Design and Applications. ISDA 2021. Lecture Notes in Networks and Systems,* eds A. Abraham, N. Gandhi, T. Hanne, T. P. Hong, T. Nogueira Rios, and W. Ding (Cham: Springer), 461–475. doi: 10.1007/978-3-030-96308-8_43

Munro, R. (2021). *Human-in-the-Loop Machine Learning, 6th Edn*. Shelter Island: Manning.

Nguyen, C. D. T., Huynh, M.-T., Quan, T. M., Nguyen, N. H., Jain, M., Ngo, V. D., et al. (2021). GOAL: gist-set online active learning for efficient chest X-ray image annotation. *Proc. Mach. Learn. Res.* 143, 545–553.

Ozdemir, F., Peng, Z., Fuernstahl, P., Tanner, C., and Goksel, O. (2021). Active learning for segmentation based on Bayesian sample queries. *Knowl. Based Syst*. 214:106531. doi: 10.1016/j.knosys.2020.106531

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., and Goksel, O. (2018). "Active learning for segmentation by optimizing content information for maximal entropy," in *Lecture Notesin Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11045* (Granada), 183–191. doi: 10.1007/978-3-030-00889-5_21

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* 2, 559–572. doi: 10.1080/14786440109462720

Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., et al. (2020). A survey of deep active learning. *ACM Comput. Surveys* 54, 1–40. doi: 10.1145/3472291

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015, Vol.* 9351 (Cham: Springer) 234–241. doi: 10.1007/978-3-319-24574-4_28

Shah, R., and Sharma, P. (2018). "Bone segmentation from X-ray images: challenges and techniques," in *Advances in Intelligent Systems and Computing, Vol. 672* (Singapore: Springer), 853–862. doi: 10.1007/978-981-10-7512-4_84

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Techn. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Siddique, N., Paheding, S., Elkin, C. P., and Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: a review of theory and applications. *IEEE Access* 9, 82031–82057. doi: 10.1109/ACCESS.2021.3086020

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*. 15, 1929–1958. doi: 10.5555/2627435

Tan, M., and Le, Q. V. (2019). "EfficientNet: rethinking model scaling for convolutional neural networks," in *36th International Conference on Machine Learning, ICML 2019* (Long Beach, CA), 10691–10700.

Zhao, Z., Yang, X., Veeravalli, B., and Zeng, Z. (2020). "Deeply supervised active learning for finger bones segmentation," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* (Montreal: Institute of Electrical and Electronics Engineers Inc.), 1620–1623. doi: 10.1109/EMBC44109.2020.9176662

Zhao, Z., Zeng, Z., Xu, K., Chen, C., and Guan, C. (2021). DSAL: deeply supervised active learning from strong and weak labelers for biomedical image segmentation. *IEEE J. Biomed. Health Inform*. 25, 3744–3751. doi: 10.1109/JBHI.2021.3052320