Active Learning and Its Application in Semantic Segmentation

Sunsheng Gu

Department of Electrical and Computer Engineering, University of Waterloo

***Abstract*— Active learning refers to a scheme where a machine learning model is allowed to select which instances to learn from. Various selection criteria can be applied to select the most informative instances to be labelled, so as to minimize labelling effort while achieving high performance. Semantic segmentation means assigning each pixel in an image to a certain class. Labelling cost is huge for segmentation as pixel-level labels are required for training. This motivates researchers to apply active learning to semantic segmentation. 4 active learning frameworks are reviewed in this survey: CEAL, SA, CEREALS, and VAAL, all of which demonstrated effectiveness in improving model performance for semantic segmentation, some of them also offered significant reduction in labelling effort. CEAL and SA were applied to background/foreground segmentation for medical images, whereas CEREALS and VAAL were applied to multi-class segmentation for road images.**

I. Active Learning

A. Motivation and Scenarios

Active learning (AL) refers to a machine learning algorithm actively selects instances for query. The query is answered by an oracle (usually a human expert) in the form of a label. The key motivation behind active learning is the belief that a machine learning algorithm can achieve higher accuracy and learn faster if it is allowed to select which instances to learn from. There are 3 different scenarios for active learning[1] as described in Figure 1. In membership query synthesis, the model generates instances anew instead of drawing them from some underlying distributions. This is applicable for automatic scientific discoveries where each instance is a specific experiment, and the label is not given by a human, but given by experimental results[2]. In stream-based selective sampling, the model decides to query or discard each instance as instances are being drawn from the input distribution one at a time[3]. Some query strategy needs to be defined to evaluate each instance, more details on that will be provided in the next section. Pool-based sampling is illustrated in Figure 2, it refers to the scenario where there exists a large unlabelled pool U and a small labelled pool L and the model selects the best instances to query from U[4]. After labelling, the newly labelled instances are added to the labelled set and the model is trained again. This process repeats until the model performance plateaus or computational cost exceeds certain thresholds.

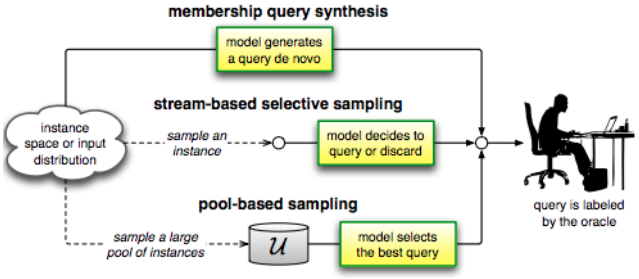

Fig. 1: Active learning schemes[1].

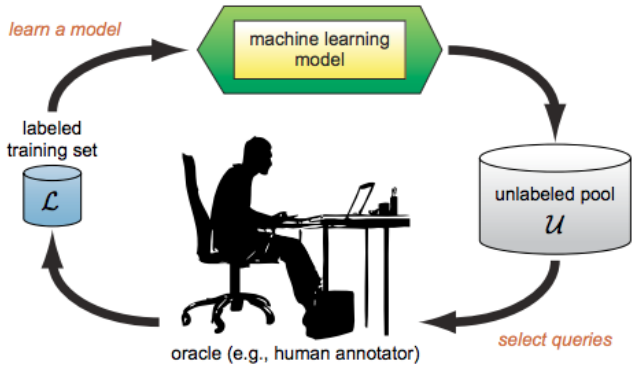

Fig. 2: Pool-based active learning cycle[1].

B. Query Strategies

1) Uncertainty Measures: A machine learning model can use a measure of uncertainty to decide whether to query an instance. One common approach is to select the least confident instance:

$$x^*_{LC} = \arg\max_x 1 - P_\theta(\hat{y}|x) \tag{1}$$

Where $\hat{y}$ is the label that maximizes the posterior probability $P_\theta(\hat{y}|x)$. Since the softmax function is frequently used to obtain $P_\theta(\hat{y}|x)$, this approach is also called max-softmax in some literature. Instances with low max-softmax scores (i.e. $1 - P_\theta(\hat{y}|x)$ is large) indicate that the model is uncertain about them, hence they would provide a lot of information to the model if they are labelled.

The max-softmax measure considers only the highest probability class but ignores all other information. Margin sampling is proposed to partially correct this deficiency:

$$x^*_M = \arg\min_x P_\theta(\hat{y_1}|x) - P_\theta(\hat{y_2}|x) \tag{2}$$

It uses the difference in probability between the top two most probable class labels as a measure of uncertainty. Smaller difference indicates higher uncertainty[1].

So far, the most popular uncertainty measure is entropy:

$$x_H^* = \arg\max_x -\sum_i P_\theta(y_i|x) \log P_\theta(y_i|x) \quad (3)$$

The y_i values span over all possible labels. Hence, entropy takes all class labels into consideration. When all class labels have the same probability values (i.e. uniform distribution), the entropy value peaks. When the model is very certain that the instance has a particular label, entropy value is low[5].

2) Query-by-Committee: The Query-by-Committee (QBC) approach uses a committee of models to vote on which label should be assigned to an instance. The instances which the models disagree the most on will be queried. Vote entropy is a common measure of disagreement:

$$x_{VE}^* = \arg\max_x -\sum_i \frac{V(y_i)}{C} \log \frac{V(y_i)}{C} \quad (4)$$

Where $V(y_i)$ is the number of votes that the label yi receives from the committee, and C is the total number of models in the committee. This can be seen as the QBC version of entropy uncertainty sampling[1].

3) Density-Weighted Methods: There is one problem with selecting queries only according to uncertainty or disagreement: some of these samples might be outliers and if introduced to the labelled training set, would degrade rather than improve model performance. Hence the information density approach was proposed so that the selected queries are both uncertain and representative of the sample distribution:

$$x_{ID}^* = \arg\max_x \phi_A(x) * (\frac{1}{U} \sum_{u=1}^{U} sim(x, x^{(U)})^\beta \quad (5)$$

Where $\phi_A(x)$ is the informativeness of the instance x according to some strategy such as uncertainty measure, and U is the number of samples in the unlabelled pool. The second term is called the density term. The "sim" in the density term is a similarity measure. The density term as a whole evaluates x based on its average similarity to all samples in the unlabelled pool. β is adjusts the importance of the density term[1].

II. Semantic Segmentation

A. Background Knowledge

Semantic segmentation refers to the task of classifying every pixel in an image. Two examples are provided in Figure 3. This is especially important for autonomous driving, where the vehicle needs to understand the scene around it.

For simpler tasks such as image classification, accuracy would be a good measure of performance. But applying pixel-wise accuracy to semantic segmentation would lead to skewed results, as there are over-represented classes (e.g. sky or road) and there are under-represented classes (e.g. pedestrians or cars) in the same image. Pixel-wise accuracy would place too much emphasis on the over-represented classes and does not reflect performance on under-represented classes. Therefore, mean intersection over union (mIOU) is proposed as an alternative performance indicator. Figure 4 demonstrates the concept of IOU. Using a person as an example, ground truth pixels are those that actually represent the person, and the model also predicts certain pixels to represent the person. The count of pixels in the intersection of the ground truth pixels and the prediction is computed, so is the count of pixels of the union of these two regions. Then IOU = (intersection pixel count)/(union pixel count). Ideally, the IOU should be 1. mIOU is the mean of IOU across all classes, so that performance on under-represented classes is not overlooked[8].

Fig. 3: Semantic segmentation example. Each colored patch in the output image represents a type of object in the input image[6].

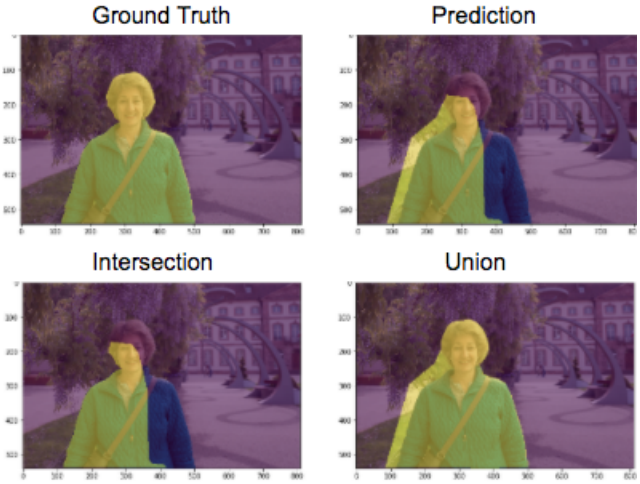

Fig. 4: IOU example with a human as the object[7].

B. Recent Development

Recent interests in multi-class semantic segmentation was sparked by a study on fully convolutional neural networks (FCN)[8]. Convolution layers are usually followed by pooling layers to reduce feature map dimensions and to extract higher level features. Such a process can be called "downsampling". Typical convolutional neural networks (CNN) designed for classification would have dense layers and a softmax at the end, in order to produce class

predictions for the image as a whole. But spatial information is lost in dense layers, hence they are not suitable for semantic segmentation where such information is necessary for assigning predictions for each pixel. In FCN, the last convolutional layer is followed by deconvolutional layers to upsample the feature map to the same size as the input image. To improve output resolution, deconvolutional layers can be combined with outputs from previous pooling layers before further upsampling. The novel FCN approach achieved 62.2% mIOU on PASCAL VOC2012 test set, which far exceeds the previous state-of-the-art model both in mIOU and inference time[8].

The work on FCN[8] inspired the design of SegNet[6]. SegNet is also called the "encoder-decoder network", with the encoder performing downsampling and the decoder performing upsampling, as shown in Figure 5. Compared to FCN, SegNet contains more upsampling layers to match each downsampling (i.e. pooling) step. SegNet decoder utilizes the corresponding max pooling indices in the encoder in upsampling, so that fine grained details can be used to define more precise boundaries. On the CamVid dataset, SegNet performed slightly better on mIOU compared to FCN (47.7% vs 47%), but SegNet had slower inferencing time[6].

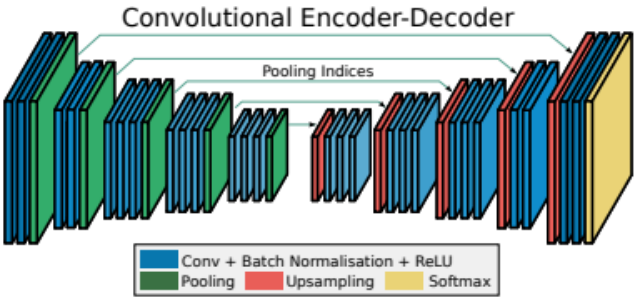

Fig. 5: The encoder-decoder architecture of SegNet[6].

Several new concepts were introduced into semantic segmentation, creating the current state-of-the-art model DeepLabV3+[9]. One concept is atrous convolution as demonstrated in Figure 6. The main advantage provided by this technique is varying the receptive fields (thus capturing features at various scales) of the convolutional kernel without increasing computational cost. Another concept is spatial pyramid pooling, referring to the process of applying filters or pooling operations with different receptive fields on an input feature map. When this is achieved through atrous convolution, the process is called atrous spatial pyramid pooling, and it is the key idea behind DeepLabV3[10]. As shown in Figure 7, inspired by SegNet (i.e. encoder-decoder), DeepLabV3+ further improves DeepLabV3 by adding one more encoding and decoding steps. DeepLabV3+ achieved 82.1% mIOU on test set of the Cityscapes dataset, setting a new state-of-the-art performance[9].

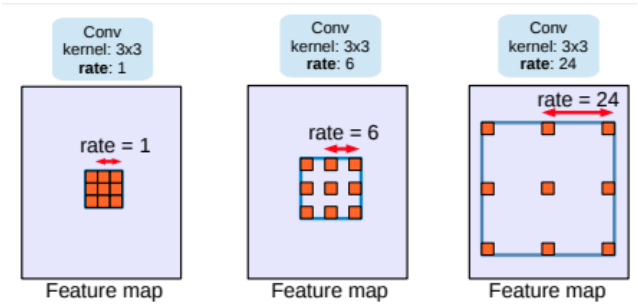

Fig. 6: Atrous convolution kernels with various rates[10].

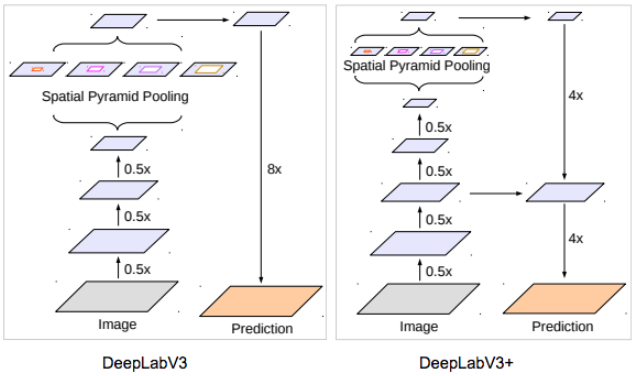

Fig. 7: Overall architecture of DeepLabV3 and DeepLabV3+[9].

III. Application of AL in Semantic Segmentation

A. Medical Images

In the field of medical imaging, the problem of labelling cost is especially acute as high quality pixel-wise labels are very difficult to obtain. Annotators are usually hand-picked experienced medical professionals, hence the number of qualified annotators is small. Creating large fully annotated training datasets is unrealistic. Hence, it would be necessary to select the most informative samples to be labelled, so that the machine learning model can perform well with a relatively small labelled pool[11].

1) CEAL for Melanoma Segmentation: In 2016, a cost-effective active learning (CEAL) method was proposed to further reduce human labour in pool-based sampling AL for image classification[12]. Instead of considering only uncertain or representative samples from the unlabelled pool, they proposed a complementary query selection strategy. At each AL iteration, they select the top K uncertain samples for labelling, as well as high confidence samples exceeding some confidence threshold δ for pseudo-labelling, i.e. assigning those samples with the label predicted by the model. The newly labelled and pseudo-labelled samples are added to the labelled set for the next iteration. They evaluated their models on the CACD dataset with 160,000 face images of 2,000 celebrities[13] and on the Caltech-256 dataset with 30,607 images containing 256 types of objects[14]. On both dataset, using CEAL with margin sampling was able to achieve the same accuracy as the previous AL state-of-the-art (triple criteria AL, which jointly evaluates uncertainty, density, and diversity) with far less labelled samples.

Inspired by the success of CEAL[12], two researchers applied CEAL to the problem of segmenting melanoma from background in medical images[11]. For segmentation, they used the U-net architecture. To measure uncertainty, they used Monte Carlo dropout to get several different models and use QBC method to determine uncertainty. They used the ISIC 2017 Challenge dataset for Skin Lesion Analysis towards melanoma detection with 2000 images[15]. They started training with 600 images, then in each iteration, they added 10 images where melanoma was not detected, 10 images with the highest uncertainty, and 15 randomly selected images from the unlabelled pool to be annotated by ground truth. They also added the images with high certainty predictions to be pseudo-labelled by the model. To evaluate the segmentation results, they used the dice index:

$$DICE(A,B) = \frac{2|A \bigcap B|}{|A| + |B|} \tag{6}$$

Where A are regions predicted to represent melanoma, and B are ground truth regions of melanoma. They ran 9 AL training cycles, with 2 epochs in each cycle. They were able to significantly improve certainty and prediction goodness over the cycles[11].

2) Suggestive Annotation (SA): The suggestive annotation method[16] was able to establish a new state-of-the-art performance on the 2015 MICCAI Gland Challenge dataset with just 50% of the available training set[17]. This dataset has 85 training images and 80 test images (60 in part A and 20 in part B). Suggestive annotation is one variation of the information density method mentioned in Section I where both uncertainty and representation are considered. The query selection is now a two step process: first, the top K most uncertain samples were chosen; then, within the K samples (K $>$ k), the top k samples (call this set S_a) which represent the unlabelled pool the best are selected for querying. In each AL training cycle, 4 bootstrapped (sampling a subset of the unlabelled pool with replacement) sample sets were used to train 4 FCNs. Then the pixel-wise uncertainty is the disagreement between the 4 FCNs' predictions on each pixel, just like the QBC approach in Section I. The overall uncertainty of a sample image is the mean of the uncertainties of the pixels. The K images with highest mean uncertainty are then selected, call this set S_c. Then the selection algorithm evaluates pairwise similarities between each sample in S_c and all other samples in the unlabelled pool S_u. The output of the last convolution layer of the segmentation network is selected as the image descriptor I. Then the representativeness of the query set S_a for an image in S_u is $f(S_a, I_x) = \max_{I_x \in S_u} sim(I_i, I_x)$, where I_x and I_i are image descriptors, one from S_u and the other from S_a, and $sim(\cdot,\cdot)$ is cosine similarity. Then the representativeness of S_a for S_u is $F(S_a, S_u) = \sum_{I_j \in S_u} f(I_i, I_x)$. This problem is then reformulated to be the maximum set cover problem (i.e. selecting a fix-sized S_a to cover as much as S_u) by forcing $sim(\cdot,\cdot)$ to take on a value of either 0 or 1. This problem is the solved in an iterative manner using a greedy approach[16].

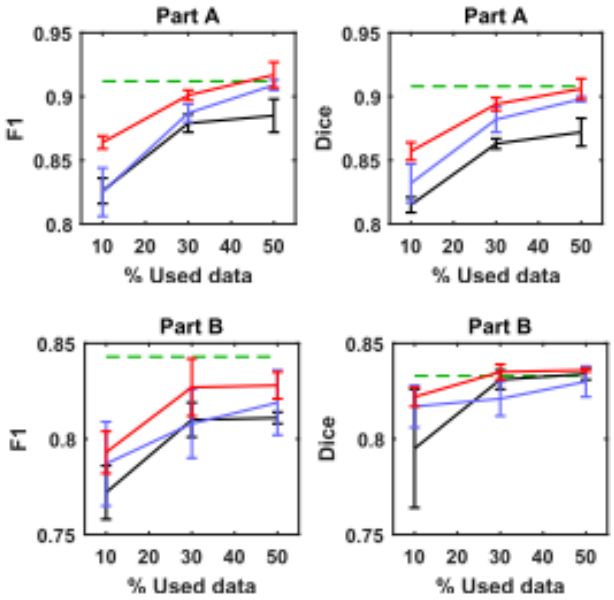

Fig. 8: Results comparison on the MICCAI dataset. Red line corresponds to suggestive annotation, blue line corresponds to using uncertainty sampling only, black line corresponds to random sampling, and dotted green line corresponds to the previous state-of-the-art performance[16].

The model used for segmentation is adopted from DCAN[18] with the addition of batch normalization and residual blocks, which speeds up convergence in training, and bottleneck structure, which reduces number of parameters without degrading performance much[19]. K is set to 16 and k is set to 8. The results of suggestive annotation seems promising, with just using 50% of the training data, it was able to achieve state-of-the-art performance on both F1 score and Dice score as shown in Figure 8.

B. Road Images

Other than medical imaging, autonomous vehicles (AVs) is another field where AL for semantic segmentation would be very valuable. For AVs, the tolerance for safety critical mistakes must be very low. Deep convolutional neural networks (CNN) have demonstrated promising results for multi-class road image segmentation, as explained in Section II. However, it was found that performance of CNN improves linearly as amount of training data increases exponentially[21]. The amount of labelling effort is further increase by the fact that for AV, unlike in medical imaging, having a binary background/foreground segmentation is far less than enough. Each class of road object needs to be segmented properly to ensure safety. This huge demand for labelled data lead to several recent works aiming to use AL to select the best query samples and therefore to reduce annotation effort.

1) CEREALS: With the goal of minimizing labelling cost, a group of researchers in Germany developed CEREALS, Cost-Effective REgion-based Active Learning for Semantic Segmentation in 2019, which achieved 95% of the maximum mIOU when spending only 17% labelling effort on

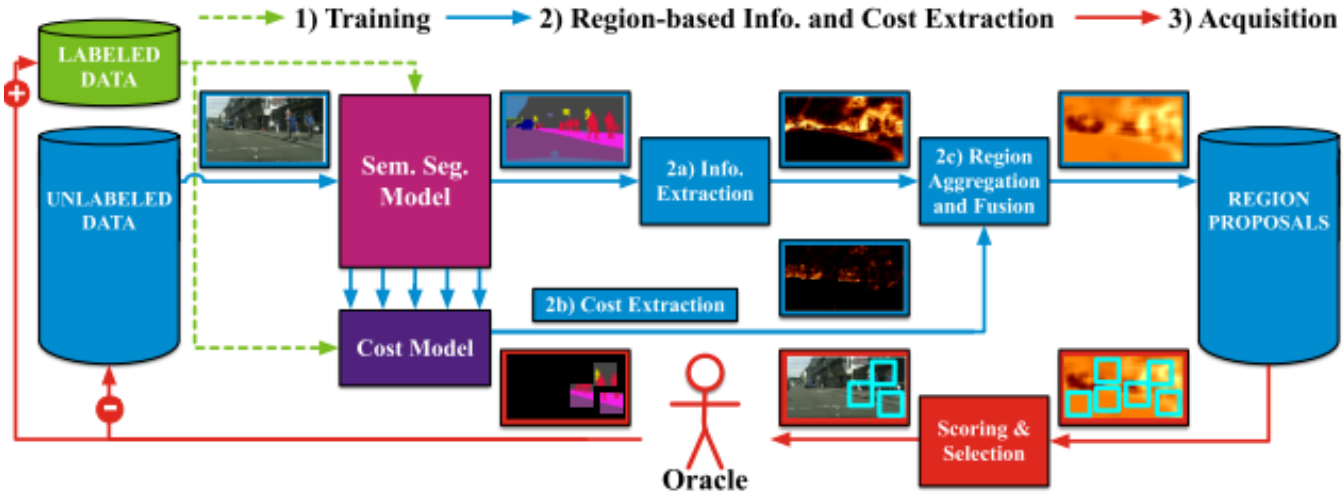

Fig. 9: AL training cycle of CEREALS. Note that there are two machine learning models involved: the segmentation model and the cost estimation model[20].

Cityscapes dataset[20]. The Cityscapes dataset is specifically created for semantic segmentation for AV research. It contains 5,000 fine resolution and 20,000 more coarsely annotated images, covering 30 classes of various objects and is collected from 50 cities over different seasons[22]. CEREALS is innovative in two ways: 1) it selects queries based on both informativeness and labelling cost, 2) instead of querying for whole images, it queries for regions to be labelled, which is shown to be more cost-effective in terms of labelling effort. To measure informativeness, two uncertainty measures were experimented: entropy and vote entropy, the equations are provided in Section I. To obtain vote entropy, in each AL cycle, an ensemble of models is obtained by applying dropout to the segmentation model. After pixel-wise uncertainties are calculated, they are accumulated over the region and normalized to [0,1]. Cost is quantified by the number of clicks executed by the annotator. The Cityscapes dataset contains ground truth information about location and frequency of clicks in each training images. To predict cost information, this ground truth click information and semantic information are fed into a cost estimation model to learn which regions in the image are likely to incur high labelling cost. The regional cost value is normalized to [0,1] as well. The training scheme is illustrated in Figure 9.

After the regional informativeness I and regional cost C are calculated, they need to be fused together to produce an overall score. The most effect scoring appeared to be $g = (1 - C) \cdot I$ [20]. Initially, the segmentation model trains on 50 randomly selected images. In each AL cycle, the regional g scores are obtained in a sliding-window fashion. Non-maximum suppression is applied to eliminate overlaps. Then the regions with highest g scores are queried. The total area of regions selected is equivalent to 50 images, i.e. if each image is 2048*1024 pixels, 50 images need to be queried in each cycle, and region size is 512*512, then 400 regions are queried in each cycle, as the image is 8 times larger than the region. The segmentation networks is FCN8s[8] pretrained on ImageNet, a width multiplier[23] of 0.25 is applied to reduce number of parameters and speed up training.

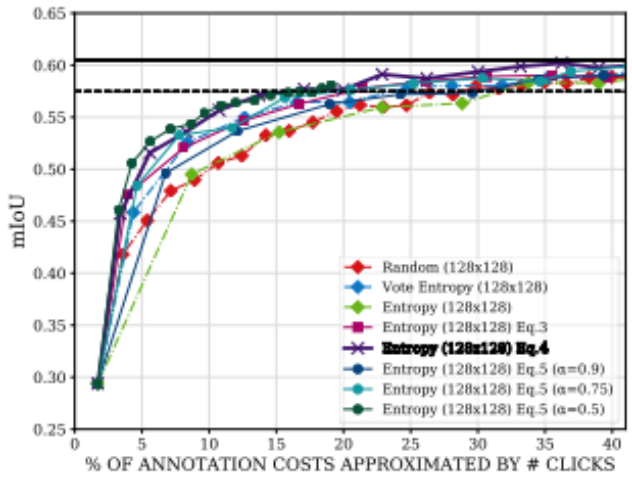

Fig. 10: Comparison of performance of different query strategies with respect to annotation cost. The black line is the maximum mIOU achieved using the modified FCN8s model when training with the fully annotated training set of Cityscapes. The dashed line is 95% of that maximum mIOU[20].

Two experiments were conducted. In the first experiment, only the informativeness score was considered, cost was not. Various region sizes were tested to find the most optimal region size. Ground truth cost data demonstrated that annotation cost indeed varies between different regions, hence the assumption that cost is uniform in all samples, which some other paper use, does not hold. It was discovered that using 128*128 region size allows CEREALS to improve mIOU with less effort than using other region sizes. It was also discovered that number of total pixels labelled is not an accurate representation of labelling effort, as highly informative regions (therefore more uncertain and confusing) tend to take more clicks to annotate than other regions. In the second experiment, region size is fixed to be 128*128

pixels, and different sampling schemes were considered, including entropy, vote entropy, several different ways of fusing entropy with cost (the Eq.3, Eq.4, Eq.5 etc in Figure 10). Using entropy with this equation $g = (1 - C) \cdot I$ (i.e. Eq.4 in Figure 10) produced the best result: achieving 95% of the best possible mIOU using only 17% total labelling effort. Compared to roughly 27% effort required to achieve roughly the same mIOU when random sampling was applied, CEREALS does offer significant human labour reduction[20].

2) VAAL: There is another work on active learning for computer vision using an innovative sampling strategy: variational adversarial active learning (VAAL)[24]. VAAL is not specifically designed for just semantic segmentation, it is a general framework that can be applied to any image related machine learning tasks, be it classification or segmentation. VAAL is impressive as it was claimed to have established a new state-of-the-art active learning performance on CIFAR10/100, Caltech-256, ImageNet, Cityscapes, and BDD100K. The VAAL framework contains 3 different networks as shown in Figure 11: a variational autoencoder (VAE) that learns a latent space encoding of the images, an adversarial network that selects samples to be queried, and a task-specific network that performs classification or segmentation. The authors reasoned that uncertainty sampling is vulnerable to outliers, hence, instead of calculating uncertainty explicitly, they learn it implicitly using the adversarial network by learning a similarity measure, more details will be covered.

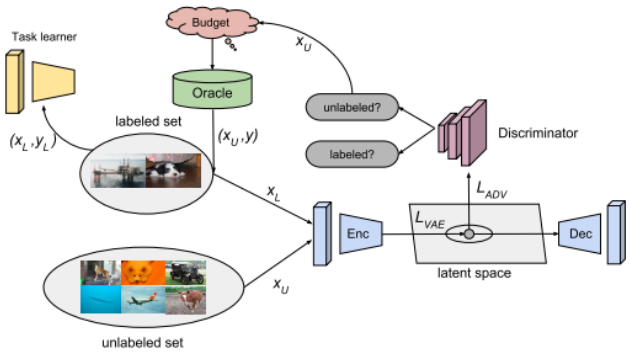

Fig. 11: Structure of the VAAL model[24].

To understand VAAL, a brief introduction on VAE would be beneficial. An autoencoder (AE) is a model that first encodes the input into some lower dimensional latent space and produce an encoding, then decodes the encoding to reconstruct the input. It can be used for de-noising, dimensionality reduction, or content generation. One problem with AE is that the ecoder is prone to overfitting. A VAE can be seem as a regularized version of an AE: instead of encoding an input as a single point in the latent space as AE, VAE learns each input as a distribution in the latent space. When using normal distribution, that would mean to learn the mean and covariance matrix of the distribution[25].

For VAAL, there are two preprocessing steps required prior to query selection: transductive representation learning (call it "TRD") and adversarial representation learning (call it "ADV"). In TRD, the VAE learns a low-dimensional latent space representation of all data from both the labelled and unlabelled set. The reasoning is that some features might be under-represented in the relatively small labelled set, the representation learned from labelled set would not capture all characteristics of the data. Therefore it would be beneficial to apply a transductive algorithm and learn from unlabelled data as well. The goal of TRD is to find latent space distributions that well-represent the all samples of the labelled and unlabelled pools. The loss function for this task is termed as $\mathcal{L}_{VAE}^{trd}$ [24]. For ADV, there are two opposing objectives. The objective for the VAE is to encode samples from both the labelled and unlabelled pools with similar distributions in the latent space, call the loss function for this task $\mathcal{L}_{VAE}^{adv}$. Whereas the objective for the adversarial network, which acts as a discriminator, is to perform the binary classification task of distinguishing labelled and unlabelled samples, call the loss function of this taks $\mathcal{L}_D$. Then the overall loss term for the VAE can be written as:

$$\mathcal{L}_{VAE} = \lambda_1 \mathcal{L}_{VAE}^{trd} + \lambda_2 \mathcal{L}_{VAE}^{adv} \tag{7}$$

Where λ_1 and λ_2 are tunable hyperparameters determining the effect of each of the two objectivs for VAE. Then the VAAL algorithm as a whole can be described below, where T denotes the task-specific module. The specific loss equations are provided in [24].

Algorithm 1 Variational Adversarial Active Learning

Input: Labeled pool (X_L, Y_L), Unlabeled pool (X_U), Initialized models for θ_T, θ_{VAE}, and θ_D
Input: Hyperparameters: epochs, λ_1, λ_2, α_1, α_2, α_3
for $e = 1$ to epochs **do**
  sample $(x_L, y_L) \sim (X_L, Y_L)$
  sample $x_U \sim X_U$
  Compute $\mathcal{L}_{\text{VAE}}^{trd}$ by using Eq. 1
  Compute $\mathcal{L}_{\text{VAE}}^{adv}$ by using Eq. 2
  $\mathcal{L}_{\text{VAE}} \leftarrow \lambda_1 \mathcal{L}_{\text{VAE}}^{trd} + \lambda_2 \mathcal{L}_{\text{VAE}}^{adv}$
  Update VAE by descending stochastic gradients:
  $\theta'_{VAE} \leftarrow \theta_{VAE} - \alpha_1 \nabla \mathcal{L}_{\text{VAE}}$
  Compute $\mathcal{L}_{\text{D}}$ by using Eq. 3
  Update D by descending its stochastic gradient:
  $\theta'_D \leftarrow \theta_D - \alpha_2 \nabla \mathcal{L}_{\text{D}}$
  Train and update T:
  $\theta'_T \leftarrow \theta_T - \alpha_3 \nabla \mathcal{L}_{\text{T}}$
end for
return Trained $\theta_T, \theta_{VAE}, \theta_D$

For segmentation task, the dilated residual network[26] was chosen. At the beginning of the AL cycle, 10% of the training data is used as the initial labelled pool. Then 5% more training data is queried and introduced to the

labelled pool each cycle. This batch of data selected from the unlabelled pool to be queried consists of the samples which the discriminator reckons as most unlikely to be from the labelled pool, i.e. the discriminator sees them as most different from the labelled samples. This is how the discriminator evaluates uncertainty implicitly, as the unlabelled samples which differ the most from the labelled ones would confuse the task network the most, hence would help the task network the most if labelled.

The authors benchmarked VAAL against many other methods mentioned in this survey, such as QBC, suggestive annotation (SA)[16], core-set (which is the the maximum set cover problem in SA), Monte-Carlo dropout (used in [20] and [11]), and random sampling. Amazingly, VAAL outperforms all these query schemes by a clear margin on both Cityscapes and BDD100K as shown in Figure 12. Cityscapes is already introduced when discussing CEREALS. BDD100K is a driving video dataset containing 120 million road images sampled all across the U.S. and covering different weather conditions. 10,000 of those images are fully annotated for semantic segmentation[27]. What is more amazing about VAAL is that not only did it beat other query strategies on Cityscapes and BDD100K for segmentation, but it also did so on CIFAR10/100, Caltech-256, and ImageNet, which boosts confidence in the superior performance of VAAL.

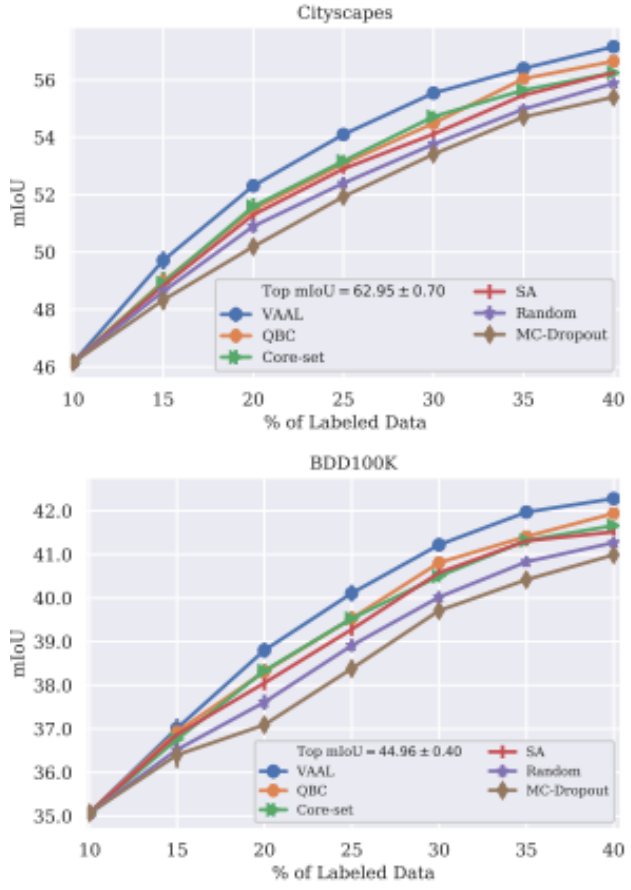

Fig. 12: Structure of the VAAL model[24].

IV. Conclusions

Recent works on applying AL to semantic segmentation showed promising results. The CEAL algorithm, which adds pseudo-labelled high-confidence samples along with uncertainty queried samples to the labelled set, was able to reduce model uncertainty and improve prediction goodness when applied to melanoma segmentation. The suggestive annotation method, which combines uncertainty with representation, was able to achieve a new state-of-the-art on the MICCAI data set with only 50% training data. The CEREALS algorithm combines uncertainty with annotation cost to select regional queries, and was able to achieve relatively high mIOU on Cityscapes with only 17% of the total labelling effort. VAAL is a particularly novel approach as it does not explicitly evaluate uncertainty from the segmentation network. It uses a VAE and a discriminator network jointly to select queries. VAAL was benchmarked against other AL methods on many different datasets, giving high confidence in its effectiveness.

References

[1] B. Settles, "Active learning literature survey," tech. rep., University of Wisconsin–Madison, 2010.

[2] R. King, K. Whelan, F. Jones, P. Reiser, C. Bryant, S. Muggleton, D. Kell, and S. Oliver, "Functional genomic hypothesis generation and experimentation by a robot scientist," *Nature*, vol. 427, no. 6971, pp. 247–252, 2004.

[3] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, no. 2, pp. 201–221, 1994.

[4] D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proceedings of the International Conference on Machine Learning(ICML)*, pp. 148–156, 1994.

[5] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–420, 1948.

[6] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2016.

[7] J. Jordan, "Evaluating image segmentation models," May 2018.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *arXiv preprint arXiv:1411.4038*, 2015.

[9] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *arXiv preprint arXiv:1802.02611*, 2018.

[10] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[11] M. Gorriz, X. Giro-i Nieto, A. Carlier, and E. Faure, "Cost-effective active learning for melanoma segmentation," *arXiv preprint arXiv:1711.09168*, 2017.

[12] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *arXiv preprint arXiv:1701.03551*, 2017.

[13] B. Chun, C. Song, and W. H. Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *Computer Vision – ECCV 2014*, pp. 768–783, 2014.

[14] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," tech. rep., California Institute of Technology, 2007.

[15] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, and A. Halpern, "Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging," *arXiv preprint arXiv:1605.01397*, 2016.

[16] L. Yang, Y. Zhang, J. Chen, S. Zhang, and D. Z. Chen, "Suggestive annotation: A deep active learning framework for biomedical image segmentation," *arXiv preprint arXiv:1706.04737*, 2017.

[17] K. Sirinukunwattana, J. Pluim, H. Chen, and X. Qi, "Gland segmentation in colon histology images: The glas challenge contest," *Medical Image Analysis*, vol. 35, pp. 489–502, 2017.

[18] H. Chen, X. Qi, L. Yu, and P. Heng, "Dcan: Deep contour-aware networks for accurate gland segmentation," in *CVPR*, pp. 2487–2496, 2016.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
[20] R. Mackowiak, P. Lenz, O. Ghori, F. Diego, O. Lange, and C. Rother, "Cereals-cost-effective region-based active learning for semantic segmentation," *arXiv preprint arXiv:1810.0972*, 2018.
[21] S. Chen, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision, ICCV 2017*, pp. 843–852, 2017.
[22] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Beneson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," *arXiv preprint arXiv:1604.01685*, 2016.
[23] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and A. Hartwig, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.
[24] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning samarth," 2019.
[25] C. Doersch, "Tutorial on variational autoencoders," tech. rep., Carnegie Mellon / UC Berkeley, 2016.
[26] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," *CVPR*, vol. 2, p. 3, 2017.
[27] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan, and T. Darrell, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," *arXiv preprint arXiv:1805.04687*, 2020.