Active Learning for Video Description With Cluster-Regularized Ensemble Ranking

David M. Chan[1], Sudheendra Vijayanarasimhan[2], David A. Ross[2], and John F. Canny[1,2]

[1] University of California, Berkeley, USA
{davidchan,canny}@berkeley.edu
[2] Google Research, USA
{svnaras,dross}@google.com

Abstract. Automatic video captioning aims to train models to generate text descriptions for all segments in a video, however, the most effective approaches require large amounts of manual annotation which is slow and expensive. Active learning is a promising way to efficiently build a training set for video captioning tasks while reducing the need to manually label uninformative examples. In this work we both explore various active learning approaches for automatic video captioning and show that a cluster-regularized ensemble strategy provides the best active learning approach to efficiently gather training sets for video captioning. We evaluate our approaches on the MSR-VTT and LSMDC datasets using both transformer and LSTM based captioning models and show that our novel strategy can achieve high performance while using up to 60% fewer training data than the strong state of the art baselines.

Keywords: Active Learning, Video Captioning

1 Introduction

Automatic video captioning is an emerging area in computer vision research that aims to generate textual descriptions of the visual components of a video. This has various applications including improving video accessibility for the blind and visually impaired [1], summarizing video [2], searching and indexing. Unfortunately, training models to do video captioning requires manual descriptions of every second of the video from a large corpus of representative videos. One of the largest current single-clip video captioning datasets, MSR-VTT, has only tens of thousands of unqiue uncorrelated videos whereas solving video captioning will likely require several orders of magnitude more to express the wide diversity of subjects, situations, and relationships possible in video data.

Active learning is a valuable approach in domains where unlabeled and partially labeled examples are readily available but obtaining manual annotations is expensive, such as is the case with automatic video captioning. However, while there has been significant investigation of active learning for computer vision tasks such as object recognition [3], object detection [4], video classification [5]

and video segmentation [6], video captioning has received comparatively little attention. The reason for this is likely rooted in the complexity of the label space. Video captioning requires both sequential input and output, dramatically increasing the complexity of traditional active learning frameworks. To our knowledge, this is one of the first works to define active learning strategies for efficiently collecting training sets for automatic video captioning.

In this paper we explore several active learning strategies for sequence to sequence active learning in video captioning, including uncertainty sampling based on label confidence, sequence entropy and query by committee methods. There are several unique challenges to active learning for deep sequence to sequence models: While traditional active learning methods [7] select one example at a time to label, retraining the model in its entirety after each new example selection, this strategy is impractical for training models such as transformer networks and LSTMs [8, 9], due to increased training time (hours vs. minutes) and increased inference time (seconds vs. milliseconds). Thus, it is far more efficient to select a large batch of examples at a time to label when using a crowd-sourced collection process [10, 11]. Traditional batch-active learning often uses ranking functions which are intractable in deep sequence to sequence learning [12–14], making active learning for video description a challenging problem, with no tractable solutions for deep neural networks.

In this work we conduct a thorough empirical analysis of various active learning strategies on two recent and standard video captioning datasets, MSR-VTT and LSMDC, using both transformer based and LSTM based captioning models, and describe a novel cluster-regularized method which is both tractable to compute, and provides strong performance in our test scenario. Our key contributions are:

1. Demonstrating that traditional uncertainty sampling techniques do not significantly outperform random sampling, likely because of the difficulty of estimating the sequence entropy.
2. A novel ensemble based ranking method (Cluster-Regularized Ensemble Divergence Active Learning, Section 3.1) specifically designed for video description active learning which outperform random sampling by a significant margin.
3. A clustering-based active learning regularization method which can help to increase sample diversity, and when combined with our query-by-committee methods can save as much as 60% of the manual annotation effort while maintaining high performance (Section 3.2).

2   Related Work

In order to reduce human effort when constructing training sets, various active learning strategies have been proposed for computer vision tasks such as object recognition [3, 15], detection [4], video classification [5] and video segmentation [6]. These methods typically select the next example to query for a label based on uncertainty sampling, entropy, or predicting reductions in risk to the underlying

model (see [7] for a comprehensive review). However, active learning for sequence labeling tasks such as automatic video captioning has received litle attention.

In the natural language processing literature, active learning methods have been proposed for actively selecting examples based on uncertainty sampling [16, 17] or query by committee approaches [18]. In [19], the authors provide a thorough analysis of various active learning methods for sequence labeling tasks using conditional random field (CRF) models. Current state-of-the-art video captioning models, however, typically utilize neural network based architectures such as transformer networks [8] or LSTMs [9] and very little research exists on how to successfully apply active learning for complex models — Transformer networks and LSTMs are expensive to train, taking hours to days to converge, compared to shallow linear models or CRFs employed in previous active learning studies (taking only minutes). Therefore querying a single example at a time is inefficient. It is far more efficient to select a large batch of examples at a time to label when using a crowd-sourced collection process as is typically the case [4].

Batch-mode active learning methods have been proposed for vision and other tasks in [12–14]. Batch selection requires more than selecting the N-best queries at a given iteration because such a strategy does not account for possible overlap in information. Therefore, the selection functions typically try to balance informativeness with a diversity term to avoid querying overlapping examples [13]. In this work, we take cues from [13], and develop a batch active-learning method for sequence learning, that is regularized using a measure of information diversity (an idea from [13]), but is tuned to be computed efficiently over sequence learning tasks, such as those in [7].

In addition to moving to batch sampling, automated video description is unique in that it has multiple possible correct sequence labels. Recent methods are usually based on expected gradient updates [20] or the entropy of a sample distribution [19], and are unable to account for scenarios where there are multiple correct labels, or there is dynamic underlying label entropy. In addition, these methods often require computing an estimate of expected model updates over the space of possible labels. This estimate can be extremely expensive for sequence learning (which has exponential label space growth), and there's no clear way of sampling from caption spaces without learning a complex language model.

Among recent methods, Coreset active learning [21], uses an integer linear program (or a greedy optimization) to find a lambda-cover over the feature set. By operating at a feature level, Coreset takes advantage of the semantic compression of the model to find sets of unlabeled samples that are useful to the model's prediction. We discuss our method compared to Coreset in Section 3.3.

Some recent methods including VAAL [22] and ALISE [23] have approached active learning from an adversarial perspective. These methods train a discriminator which attempts to determine which samples are labeled and unlabeled, then select the likely unlabeled samples for training. However, they typically require large number of samples to reliably train the discriminator which is unavailable in the beginning of the active learning process. Nonetheless, it would be an interesting future direction to explore adversarial models for active learn-

ing on complex latent spaces. Deep Bayesian active learning [24] shows some promise, however strong Bayesian networks for multi-modal vision and language problems are still out out of reach for large scale complex datasets.

3   Methods

In this work we introduce a new method for sequence active learning for video description, and compare against several baseline algorithms (Those listed below, along with Coreset [21] active learning and ALISE [23]). Throughout this section, we refer to a video v_i, and its associated set of descriptions $\mathcal{D} = \{c_1(v_i) \ldots c_n(v_i)\}$. A set of descriptions generated by a model m_j is referred to by $\{c_{m_j,1}(v_i) \ldots c_{m_j,n}(v_i)\}$. Videos may have multiple descriptions either through multiple-sampling of the model generative distribution, or through multiple ground-truth labels of the same video. The probability distribution $P_{m_j}(c_i)$ is the likelihood of a description c_i under the model m_j, and the distribution $\mathcal{P}^{cond}(m_j, c_i^k) = P_{m_j}(c_i^k(v_i)|c_i^{k-1}(v_i), \ldots, c_i^0(v_i))$ is the conditional distribution of the next word k under the model given the previous words in the description.

3.1   Active Learning Methods

Random Selection Baseline: Our baseline method is to sample new data points from the training set uniformly at random. Random selection is a strong baseline. It directly models the data distribution through sampling, placing emphasis on representative data, but not "novel" data. Trying to sample outside the random distribution is more likely to cause over-sampling of parts of the data (demonstrated in Figures 3), leading to poorer overall validation performance.

Maximum Entropy Active Learning: Traditional methods for active learning [7] are often entropy based. As a second strong baseline, we present a maximum entropy active learning method in which we rank samples based on a sample of the entropy of the dataset. Unfortunately, given the exponential number of computations that have to be made in the sequence length, the entropy of the entire output distribution is intractable to compute directly. Thus, to approximate the entropy of the description distribution we compute the mean entropy of the word output distributions at each new word along the generation process of a new description of a sample using our current model. Thus, using a candidate model m, we sample K candidate sentences for each video, and we select samples which maximize the ranking function:

$$R(v_i) = \frac{1}{K} \sum_{k=1}^{K} \sum_{w=1}^{|c_{m,k}(v_i)|} -P_m(c_{m,k}^w(i)) \log P_m(c_{m,k}^w(v_i)) \qquad (1)$$

where $R(v_i)$ is our approximate estimate of the entropy of any given sample's distribution.

Minimum Likelihood Active Learning: In the minimum likelihood active learning scenario, we select samples where the descriptions that the model generates have the lowest log likelihood under the model distribution. Thus, using a candidate model m, we sample K candidate sentences for each video, and then choose samples which minimize the ranking function:

$$R(v_i) = \frac{1}{K} \sum_{k=1}^{K} \sum_{w=1}^{|c_{m,k}(v_i)|} \log P_m(c_{m,k}^w(v_i)|c_{m,k}^{w-1}(v_i) \ldots c_{m,k}^0(v_i)) \qquad (2)$$

Empirically, we find that the minimum likelihood active learning method is a stronger method than the entropy for use in video captioning (See Figure 2), however this measure of uncertainty suffers from the fact that the model may be very confident about its wrong answers, and will be unable to learn effectively when this is the case. Because these very confident wrong answers are never sampled (or are sampled later in the training process), the model is unable to correct for the initially introduced bias.

Query By Committee Ensemble Agreement Active Learning: To help alleviate the issues with single model uncertainty, we introduce the notion of an ensemble agreement active learning ranking based on query by committee methods for traditional active learning [18]. With this method, we sample a set of likely captions from each member of an ensemble of models (using beam search), and compute the mean pairwise likelihood. For an ensemble of L models $\{m_1, \ldots, m_L\}$, from each model m_l we sample captions $\{c_{m_i,1} \ldots c_{m_i,K}\}$ for each available unlabeled video. Our ranking criterion is then to minimize:

$$R(v_i) = \frac{1}{L(L-1)} \sum_{p=1}^{L} \sum_{\substack{q=1 \\ q \neq p}}^{L} \sum_{k=1}^{K} \sum_{w=1}^{|c_{m_p,k}(v_i)|} \frac{\log \mathcal{P}^{cond}(m_q, c_{m_p,k}^w(i))}{K|c_{m_p,k}(v_i)|} \qquad (3)$$

The idea here is to select samples for labeling which have low agreement, as these are the samples have higher uncertainty under our model/training process. In this scenario, we alleviate many of the concerns with models having high confidence in wrong answers, as this phenomenon tends to be local to particular models, and these highly incorrect answers will have low likelihood under the learned distributions of the other members of the ensemble.

Query By Committee Ensemble Divergence Active Learning (Proposed Method): While entropy/perplexity measures for active learning have been well explored in the literature [7], it is unclear if these measures are correct for the captioning task. Even if the caption distribution for a video has high entropy, meaning there are many possible likely captions (or even many possible correct captions), this high entropy does not mean that the model is unsure of the outcome. Samples that have many possible captions will thus be over-sampled, since any of the generated captions will have fundamentally lower likelihood than a sample with fewer possible captions. In order to avoid this, we

present a method, which computes the KL-divergence between the conditional distributions of the ensemble members. Thus, if the models would choose similar words, given similar inputs - we consider the models to be in agreement. Similarly to the above, for an ensemble of L models $\{m_1...m_L\}$, from each model m_l we sample captions $\{c_{m_i,1}...c_{m_i,K}\}$ for each available unlabeled video. We then choose samples which maximize:

$$R(v_i) = \frac{1}{L(L-1)zhou} \sum_{p=1}^{L} \sum_{\substack{q=1 \\ q \neq p}}^{L} \sum_{k=1}^{K} \frac{D_{KL}\left(P_{m_p}(c_{m_p,k}(v_i))||P_{m_q}(c_{m_p,k}(v_i))\right)}{K} \quad (4)$$

Unfortunately, computing the full joint distribution is prohibitively expensive. Thus, instead we restrict the computation of the divergence to the sum of per-word divergences:

$$R(v_i) = \frac{1}{L(L-1)} \sum_{p=1}^{L} \sum_{\substack{q=1 \\ q \neq p}}^{L} \sum_{k=1}^{K} \frac{D(m_p, m_q, c_{m_p,k}(v_i))}{K} \quad (5)$$

where

$$D(m_p, m_q, c_{m_p,k}(v_i)) =$$

$$\frac{\displaystyle\sum_{w=1}^{|c_{m_p,k}(i)|} D_{KL}\left(\mathcal{P}^{cond}(m_p, c_{m_p,k}^w(v_i))||\mathcal{P}^{cond}(m_q, c_{m_p,k}^w(v_i))\right)}{|c_{m_p,k}(v_i)|}$$

$$(6)$$

is the per-word KL-divergence along the generation of the description $c_{m_p,k}(v_i)$ in each of the models. Compared to the likelihood method, this model gives a better estimate of the divergence of the distributions learned by the models of the ensemble. This measure is also independent of the sample length, and distribution perplexity, confounding factors when looking only at the likelihood of the samples.

3.2   Improving Diversity With Clustering

During the training of the initial active learning models, we noticed through a qualitative investigation that models seemed to be over-sampling parts of the training feature space. This was confirmed by running the experiments shown in Figure 3. To help combat this, we enforced a clustering-based diversity criterion. We first performed a k-means clustering of the training data using the mean (across the temporal dimension) of our visual features. We chose $K = N/20$ clusters, where N is the number of training samples in the dataset. See section 4 for a justification for this number of clusters. We then force the active learning algorithm to select at most ϕ samples from each cluster, which notably increases diversity. For the experiments in this paper, we found $\phi = 3$ to be the best hyper-parameter value, out of $\phi = 1, 2, 3, \ldots 10$.

3.3   Comparison with Coreset Active Learning

While our method shares some significant similarities at a glance to Coreset [21] (i.e. we both use delta-covers of a space to regularize the sampling), they have some notable differences. The Coreset method uses the distribution of the feature space, combined with k-centers over the unlabeled data to select a set of samples which should be annotated. This is equivalent to finding a delta cover over the distribution of the data in the unlabeled space. Our proposed method (Ensemble Divergence + Cluster Regularization) uses the uncertainty of the underlying model to compute a score, and then attempts to regularize this score across the data space by enforcing that no two samples are too close together. Our method not only achieves better performance on our sequence learning tasks, but also runs notably quicker than Coreset, which can fail to solve the Integer Linear Program efficiently. It is interesting future work, however outside the scope of this exploratory paper, to explore selecting among Coresets using our uncertainty metric. Figure 1 directly compares Coreset and Greedy Coreset with our proposed model on the video description problem.

3.4   Models

The goal of this paper is to explore active learning methods across multiple different model structures. In our experiments we use both a transformer-based model based on Zhou *et. al.* [25], and the popular S2VT RNN architecture [9] (See supplementary materials for details). Our models are able to achieve performance comparable to state-of-the-art models using vision-only features and without using policy gradients to optimize a downstream metric [9, 26]. By adding multi-modal features, and direct REINFORCE optimization, you can gain 7-10 CIDEr points over our implementations [26]. However, while there are more complex model pipelines, we chose two very simple architectures to demonstrate the efficacy of active learning, improve iteration time, and decrease the chance of confounding visual effects. We expect the presented methods to transfer to more complex optimization schemes, and powerful architectures given the flexibility of the formulation and our ablation results.

3.5   Datasets

We demonstrate the performance of our model on two common video description datasets, MSR-VTT [10] and the LSMDC [27]. While these methods may apply to video datasets generated using Automated Speech Recognition (HowTo-100M [28]) or dense captioning tasks (ActivityNet Captions [29]), we focus on pre-clipped videos with high quality descriptive annotations. We refer the reader to the supplementary materials for a description of the datasets in use.

3.6   Experimental Setup

Feature Extraction and Pre-processing: To avoid conflating the visual representations of the data with the performance of the captioning model, we

follow video-captioning convention and pre-extract features from the videos using a Distill-3D (D3D) [30] model pre-trained on the Kinetics-600 dataset for activity recognition. The videos are resized on the short-edge to 256px, then center-cropped to 256x256. They are down-sampled to 64 frames at 3 frames per second (with cyclic repetition for videos that are too short), and then passed through the D3D model to generate a 7x1024 representational tensor for the video which is used in the captioning process. The text data is tokenized using a sub-word encoding [31] with a vocabulary size of 8192.

Training: Each model is trained in PyTorch [32] with a batch-size of 512 for 80 epochs. We use the ADAM [33] optimizer with a warm-up learning rate schedule with 1000 steps of warm-up, ranging from $1e^{-6}$ to $1e^{-3}$, then decaying over 250,000 steps to 0. We run our experiments using 8 Tesla T4 accelerators on Google Cloud Platform, making use of NVIDIA Apex[3] for mixed-precision fp-16 training.

3.7   Evaluation:

In all active learning methods, we begin by seeding the method with 5% of the data, chosen randomly. For a fair comparison, this random slice is shared across each of the active learning methods. We then train an initial classifier for use with our active learning method. When the classifier has converged (achieved minimum loss on a validation dataset, or trained for 80 epochs, whichever comes first), we use the classifier, and the proposed ranking methods (Using a cluster-limit $\phi = 3$, and a set of 8 sampled captions) to select an additional 5% of the training data. This happens 19 additional times (20 total evaluation points), until all of the training data has been selected. At each step, we run an evaluation of the model to determine the performance. Exploring the active learning process when using larger and smaller batches is interesting future work — when selecting very few examples, there is more potential benefit, but more computation required, selecting more samples requires less computation, but can be a more difficult task. Exploring ideas in continual learning, where the classifiers are re-initialized with weights from the previous active learning step is also interesting future work, however we found in practice that this does not heavily influence the training process.

During evaluation, we sample 8 candidate sentences with a temperature of 0.8, which are then evaluated using the COCO Captions Evaluation Tools [34]. For tuning, we use a validation dataset sub-sampled from the training dataset, and we report the results on the official validation dataset (unseen during training/tuning) below. For ensemble-based metrics, we use the mean performance of the ensemble members. For non-ensemble based metrics, we perform multiple runs of active learning, and report the error as a 95% bootstrapped confidence interval. While the 95% is somewhat arbitrary, we present the full trajectories, for readers to explore.
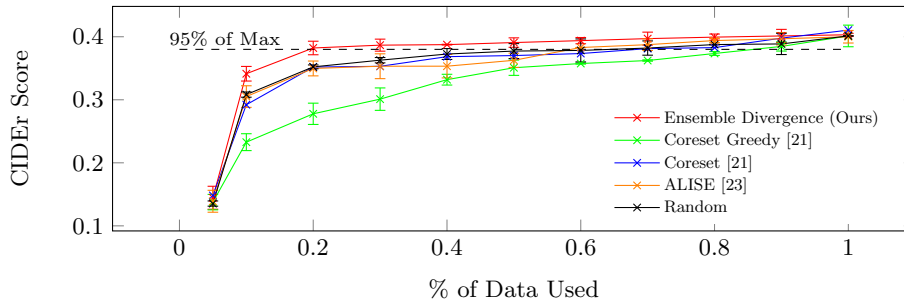
---

[3] `https://github.com/NVIDIA/apex`

Fig. 1: Validation performance of active learning methods on the MSR-VTT dataset using the CIDEr metric [35]. Each run represents the mean of a bootstrap sample of ten runs. Our proposed method significantly outperforms all other methods, achieving 95% of the max performance while using only 25% of the data. This figure is measured 10 intervals instead of 20, due to the cost of Coreset's ILP solver.

4   Results & Discussion

Our key results using the transformer architecture on the MSR-VTT dataset are presented in Figure 1. Clearly, we can see that the clustered-divergence outperforms the benchmark models by a wide margin, using about 25% of the data to achieve a CIDEr score of 0.38 (95% of max). A full set of results is shown in Figure 2 for the methods from Section 3.1. Some additional qualitative results are presented in the supplementary materials.

Our method is highly prone to over-sampling parts of the distribution. To demonstrate over-sampling by examining the performance of our models across multiple clusters. Figure 3 shows that enforcing diversity is key to our approach: If we use no clustering, we actually fail to outperform random performance while adding a few clusters allows us to mitigate this effect and adding sufficient clustering allows for significant performance benefits. We can also see the effect of clustering by examining the mean distance to the validation set over the active learning iterations. We can also see from Figure 3 that the agreement method alone is unable to efficiently distribute across the validation set, however random and clustered methods achieve similar distribution effects. It's interesting to note, however, that even without the cluster enforcement the agreement metrics select from more visual diversity than the entropy/likelihood methods - leading to better performance (Table 1). The results for a cluster-regularized random selection method are given in Figure 2, however it is not significantly different from random alone, since the random method already samples uniformly from the set of input samples. Figure 4 shows that as we increase the number of ensemble members, the performance increased, however there are diminishing returns, as the models begin to capture the same data.

We can also see from Figure 2 that the ordering of methods can be dependent on the metrics chosen. While our proposed method outperforms all of the baseline methods, it is most helpful under the CIDEr and ROUGE metrics which
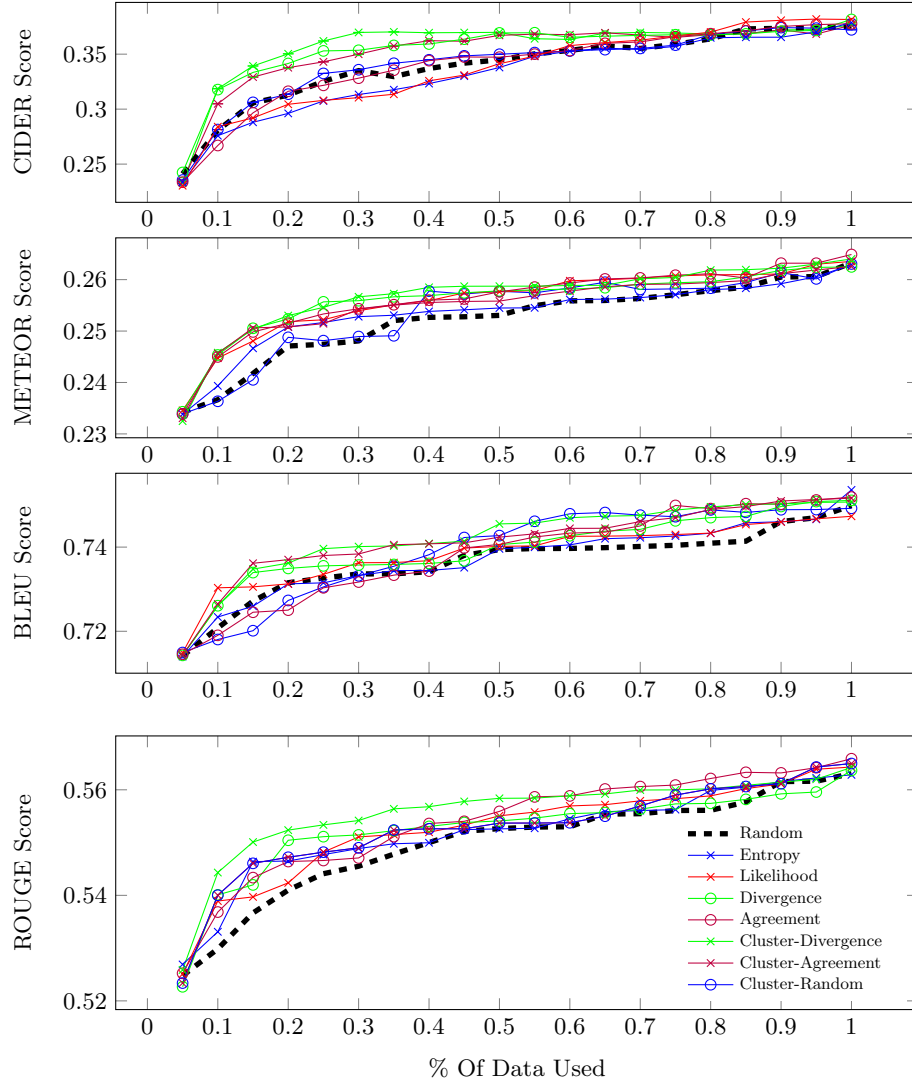
Fig. 2: Validation performance across many potential active learning methods on the MSR-VTT dataset using the transformer model structure with respect to CIDEr Score [35], METEOR Score [36], BLEU Score [37] and ROUGE-L Score [38]. The curves presented are the means of 3 individual experiments using each method. Error bars are omitted for clarity. ALISE and Coreset are omitted due to computation time costs (However see Figure 1 for a comparison on CIDEr).

Method	Mean Number of Clusters Selected/Iteration
Random Selection	195.47 ± 21.2
Cluster-Regularized Divergence	212.5 ± 14.4
Cluster-Regularized Agreement	202.3 ± 17.6
Cluster-Regularized Entropy	215.7 ± 12.8
Agreement Only	181.00 ± 16.9
Entropy	160.31 ± 16.4
Likelihood	169.25 ± 13.7

Table 1: Average number of clusters selected per iteration. The random and cluster-normalized methods select from a wider visual variety of samples, while the non-normalized samples select very few clusters on average.
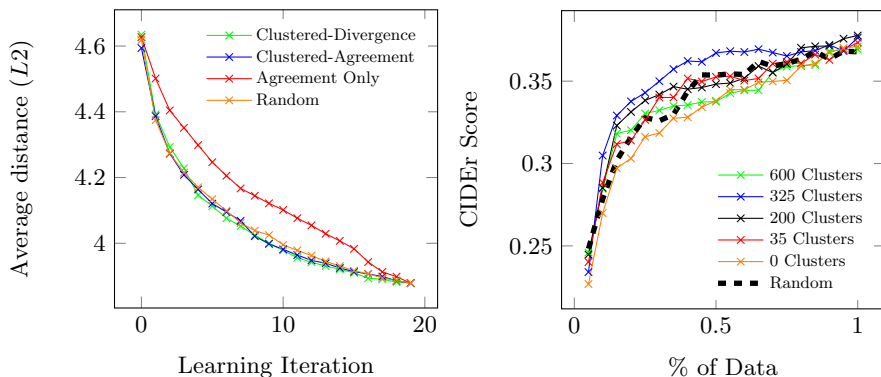
Fig. 3: (Left) Average distance of validation samples to the nearest training sample over the active learning process. Models with improved diversity improve the distance to the training set more rapidly. We suspect this diversity is why random methods work well vs. non-diversity enforced methods as random methods contain a built-in coverage of the dataset. (Right) Performance of the cluster-divergence active learning method across different numbers of clusters. Performance is greater with greater numbers of clusters, until saturation, where performance regresses to random.

prefer higher-level descriptions of the scenes. The method helps less for improving metrics that depend on lower-level semantics, such as BLEU and METEOR. We suspect that this is due to the influence of the active learning method on sampling a diverse set of samples - as increasing the sample diversity can help to improve high-level understanding of a scene, while perhaps having detrimental impacts on the language modeling process.

While we have made the case that a strong diversity of samples is required, it is also interesting to look at exactly which samples were selected. Figure 5 demonstrates some of the diversity of samples selected by our methods in comparison to the samples selected by the random method. We can see that the active learning method is sampling from a diverse set of elements from each cluster, while the random method is sampling a representative sample, but not necessarily the most relevant or useful videos.
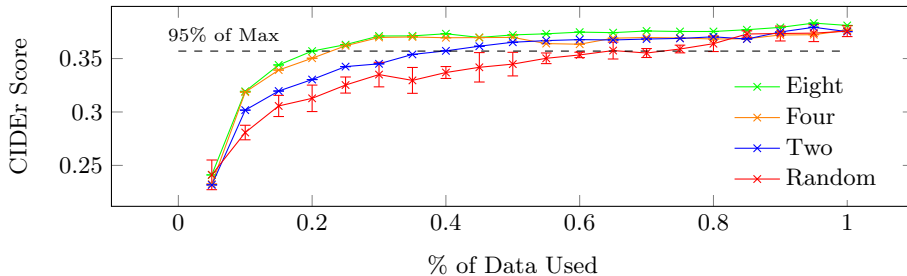
Fig. 4: Validation performance with differing numbers of ensemble members on the MSR-VTT dataset. We see increasing the number of ensemble members leads to increased performance. We speculate that the diminishing returns are caused by independent models capturing similar information.

One important thing to note is that because we are sampling from data that is in the initial training data for the two datasets, the results presented in this paper may be an optimistic upper bound for the performance of an active learning tool. There is a significant amount of cleaning and curating that goes into these datasets which may or may not impact the final results of the classification, and the effort may be higher when annotating video in the wild. Future techniques may need to be developed for efficiently cleaning data, or curating samples that are relevant to captioning as a whole.

One downside to our experimental method is that our models do not achieve optimal performance in each training step, as the optimal hyper-parameters of the model change as more data is added. To ease this issue we use an adaptive training scheme which trains for more iterations than necessary with early stopping, however it is an interesting direction of future work to explore auto-tuning during the learning process to improve performance.

Our proposed method is not limited to the dataset or model. Figure 7 demonstrates the performance of our best method, clustered divergence, on the LSMDC dataset. We can see here that we achieve a CIDEr score of 0.121 (95% of max) with only 50% of data required by random sampling. Thus, we can see that the performance of the active learning method is not just limited to the MSR-VTT dataset. In addition, Figure 6 demonstrates that the performance is not limited only to our transformer based model. The S2VT model also improves, achieving a CIDEr score of 0.3219 with only 60% of data required by random selection.

In addition to requiring fewer data, our method can be significantly more efficient than the current state of the art methods. On our test-bench machine, we saw the following ranking times using the MSR-VTT dataset (Samples / Sec): Random: 2012.04, Entropy: 12.41, Cluster-Regularized Ensemble Ranking: 11.08, ALISE: 6.89, Coreset-Optimal: 0.11, and Coreset-Greedy: 11.89.

5   Conclusion & Future Work

In this paper, we have presented an initial set of methods aiming to tackle the active learning problem for video description, a challenging task requiring

Fig. 5: Visualization of four clusters of videos from the training dataset. Highlighted elements were selected by the cluster-divergence learning method (red), or the random method (yellow) in the first two iterations. In clusters with low visual diversity active learning selects fewer samples (top-left, bottom-left, bottom-right), while selecting more samples in clusters with high visual diversity (top-right), suggesting that the active method is choosing more informative samples.

complex modeling where due to the complexity of the output distribution, many active learning methods are unable to function efficiently, or at all. We have shown that we can achieve 95% of the full performance of a trained model with between 25% and 60% of the training data (and thus, manual labeling effort), across varying models and datasets.

While pairwise measures among ensemble members may be a good model of uncertainty, there are many such measures. Expected gradient variance methods such as [20, 19] are good candidates for future exploration. While such methods now do not account for the complexity of multiple correct labels, and dynamic entropy distributions, we may be able to compute high quality estimates. Such gradient methods may work in scenarios where the KL divergence between the final distributions of the models may be relatively low, but the evaluated sample has useful second-order gradient information.

It is also interesting, and likely fruitful, future work to explore different methods for clustering the elements of the training dataset. In many cases, we would like to enforce a subject-level diversity among the different inputs (as show by
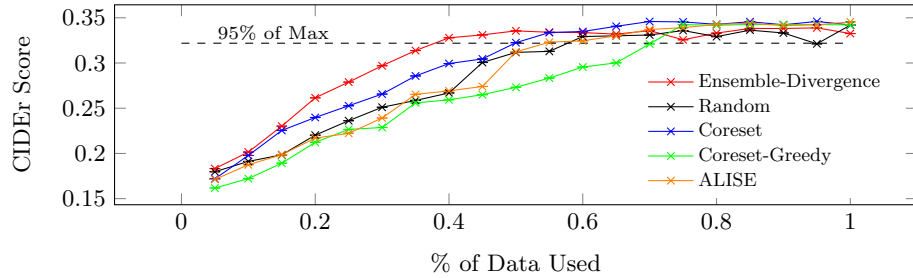
Fig. 6: Performance using the LSTM model. While overall performance is lower, the clustered-divergence learning method can save more than 20% percent of the data.
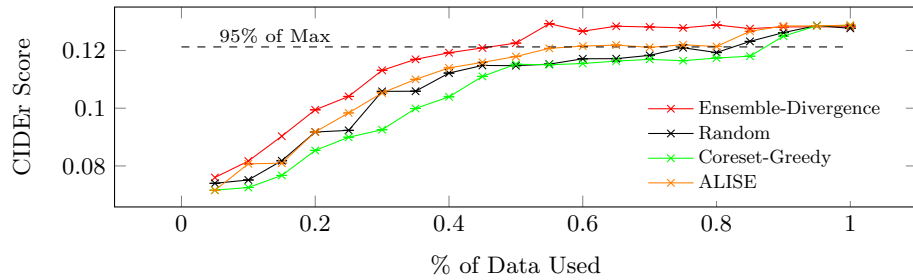

Fig. 7: Validation performance for the LSMDC dataset. We achieve strong performance using almost 35% less data. We do not include Coreset, as it took > 24 hours per active-learning step to compute.

Figure 5), however visual similarity may not necessarily be the best metric to use for clustering. Using additional features to rank the diversity of the samples may provide better results, by increasing the individual diversity of each cohort more than k-means clustering in the visual space.

By exploring the applications of our work in practice, we can build robust active learning methods and collect large and effective datasets for video description. We hope these datasets will be used to improve the performance of downstream description tools in this complex and challenging labeling domain.

6   Acknowledgements

References

1. Yoon, I., Mathur, U., Gibson, B., Fazli, T.P., Miele, J.: Video accessibility for the visually impaired. In: International Conference on Machine Learning AI for Social Good Workshop. (2019)
2. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: ECCV. (2016) 766–782
3. Collins, B., Deng, J., Li, K., Fei-Fei, L.: Towards scalable dataset construction: An active learning approach. In Forsyth, D., Torr, P., Zisserman, A., eds.: Computer Vision – ECCV 2008, Berlin, Heidelberg, Springer Berlin Heidelberg (2008) 86–98
4. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: Training object detectors with crawled data and crowds. Int. J. Comput. Vision **108** (2014) 97–114
5. Yan, R., Yang, J., Hauptmann, A.: Automatically labeling video data using multi-class active learning. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. ICCV '03, USA, IEEE Computer Society (2003) 516
6. Vijayanarasimhan, S., Grauman, K.: Active frame selection for label propagation in videos. In Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C., eds.: Computer Vision – ECCV 2012, Berlin, Heidelberg, Springer Berlin Heidelberg (2012) 496–509
7. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
8. Zhou, L., Zhou, Y., Corso, J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. (2018)
9. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence – video to text. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
10. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 5288–5296
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, Ieee (2009) 248–255
12. Hoi, S.C.H., Jin, R., Zhu, J., Lyu, M.R.: Semisupervised svm batch mode active learning with applications to image retrieval. ACM Trans. Inf. Syst. **27** (2009)
13. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning. ICML'03, AAAI Press (2003) 59–66
14. Vijayanarasimhan, S., Jain, P., Grauman, K.: Far-sighted active learning on a budget for image and video recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2010)
15. Vijayanarasimhan, S., Grauman, K.: What's it going to cost you?: Predicting effort vs. informativeness for multi-label image annotations. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2009)
16. Culotta, A., McCallum, A.: Reducing labeling effort for structured prediction tasks. In: Proceedings of the 20th National Conference on Artificial Intelligence - Volume 2. AAAI'05, AAAI Press (2005) 746–751
17. Scheffer, T., Decomain, C., Wrobel, S.: Active hidden markov models for information extraction. In: Proceedings of the 4th International Conference on Advances

in Intelligent Data Analysis. IDA '01, Berlin, Heidelberg, Springer-Verlag (2001) 309–318

18. Dagan, I., Engelson, S.P.: Committee-based sampling for training probabilistic classifiers. In: Proceedings of the Twelfth International Conference on International Conference on Machine Learning. ICML'95, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1995) 150–157

19. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics (2008) 1070–1079

20. Huang, J., Child, R., Rao, V., Liu, H., Satheesh, S., Coates, A.: Active learning for speech recognition: the power of gradients. arXiv preprint arXiv:1612.03226 (2016)

21. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A coreset approach. arXiv preprint arXiv:1708.00489 (2017)

22. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 5972–5981

23. Deng, Y., Chen, K., Shen, Y., Jin, H.: Adversarial active learning for sequences labeling and generation. In: IJCAI. (2018) 4012–4018

24. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org (2017) 1183–1192

25. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 8739–8748

26. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR) **52** (2019) 1–37

27. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123** (2017) 94–120

28. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2630–2640

29. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. (2017) 706–715

30. Stroud, J., Ross, D., Sun, C., Deng, J., Sukthankar, R.: D3d: Distilled 3d networks for video action recognition. In: The IEEE Winter Conference on Applications of Computer Vision. (2020) 625–634

31. Kudo, T., Richardson, J.: Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. arXiv preprint arXiv:1808.06226 (2018)

32. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. (2017)

33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

34. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
35. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015) 4566–4575
36. Agarwal, A., Lavie, A.: Meteor, m-bleu and m-ter: Evaluation metrics for high-correlation with human rankings of machine translation output. In: Proceedings of the Third Workshop on Statistical Machine Translation. (2008) 115–118
37. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318
38. Lin, C.Y., Hovy, E.: Manual and automatic evaluation of summaries. In: Proceedings of the ACL-02 Workshop on Automatic Summarization-Volume 4, Association for Computational Linguistics (2002) 45–51