

Bayesian Uncertainty and Expected Gradient Length - Regression: Two Sides Of The Same Coin?

Megh Shukla
Mercedes-Benz Research and Development India
megh.shukla@daimler.com

Abstract

Active learning algorithms select a subset of data for annotation to maximize the model performance on a budget. One such algorithm is Expected Gradient Length, which as the name suggests uses the approximate gradient induced per example in the sampling process. While Expected Gradient Length has been successfully used for classification and regression, the formulation for regression remains intuitively driven. Hence, our theoretical contribution involves deriving this formulation, thereby supporting experimental evidence [4, 5]. Subsequently, we show that expected gradient length in regression is equivalent to Bayesian uncertainty [22]. If certain assumptions are infeasible, our algorithmic contribution (EGL++) approximates the effect of ensembles with a single deterministic network. Instead of computing multiple possible inferences per input, we leverage previously annotated samples to quantify the probability of previous labels being the true label. Such an approach allows us to extend expected gradient length to a new task: human pose estimation. We perform experimental validation on two human pose datasets (MPII and LSP/LSPET), highlighting the interpretability and competitiveness of EGL++ with different active learning algorithms for human pose estimation.

1. Introduction

Imagine that as researchers, we are tasked with improving computer vision models for a client's products, deployed across the globe. The challenge is that training datasets may not reflect all real world use cases. While we could improve real-world model performance by collecting data from end-users, the resultant collection would be humongous and data annotation is arduous. *Instead, can we identify a subset of images for annotation, to maximize the model performance per set of images annotated? Simultaneously, can we improve the interpretability in our sampling process?*

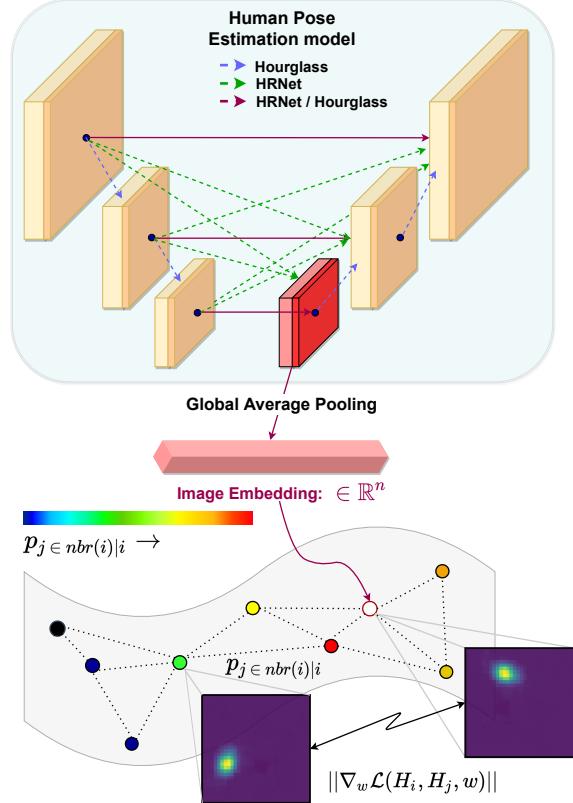


Figure 1. While our theoretical contribution links expected gradient length (EGL) to Bayesian uncertainty, our algorithmic contribution approximates the effect of ensembles with a single deterministic network: The notion of neighbors for an (image, target) pair uses low-dimensional representations from the neural network. These representations allow us to interpret the embedding space into neighborhood distributions [49]. This neighborhood distribution quantifies the probability of various inferences being the target inference for the given sample, approximating an ensemble. Not only do we leverage previously labelled samples within the EGL framework, the approach quantifies the intuition that similar inputs have similar representation. Being label shape agnostic, EGL++ extends EGL to the new task of human pose estimation.

Our search for a solution leads us to Active Learning (AL) [38], a suite of algorithms designed to select a subset of images for annotation when labelling the entire unlabelled pool of data is infeasible. This subset is obtained using the model's feedback, ensuring that the selected images impart new information for the model to learn. AL is a cyclical process of selection-annotation-training, allowing for fast prototyping, lower budget requirements and potentially limiting bias in the data. AL approaches span uncertainty [12], ensemble [30] and diversity [37] to name a few. While diversity based approaches seek to introduce semantic variations, uncertainty quantifies model ambiguity to select images for annotation. Ensemble approaches personify wisdom of the masses, with multiple models used to determine the annotation set.

We focus on Expected Gradient Length (EGL) [38], a classical AL algorithm used in diverse applications such as speech recognition, text representation and biomedical imagery. While EGL has been extensively studied in classification, recent approaches apply the framework for regression [4, 5]. However, the formulation for regression remains intuitively driven and lacks theoretical background. With this motivation, we focus on supporting the empirical results in [4, 5] by developing theory based on previous statistical results involving Fisher information [17, 28, 39, 44]. Additionally, we also explore an alternative interpretation of the framework, allowing us to extend EGL to the new task of human pose estimation. Our contributions are:

1. **Theory:** We theoretically derive the formulation for EGL - regression [4, 5], and further show the equivalency between the resultant closed form solution and predictive variance [22].

2. **Algorithm:** EGL++ is an efficient, alternative interpretation of the EGL framework. We show that the effect of an ensemble can be approximated with a single deterministic network by leveraging previously labelled data explicitly. This interpretation additionally allows us to extend expected gradient length to the completely new task of human pose estimation.

Similar inputs have similar representations since neural networks are almost everywhere differentiable. EGL++ quantifies this intuition, weighing the importance previously labelled data points in the neighborhood of any given sample have. Since this approach depends on neural representations and agnostic of the target inference shape, EGL++ allows us to extend the EGL paradigm to human pose estimation (HPE). Active learning for HPE is tricky ¹; popular architectures for HPE are fully *convolutional* and regress 2D heatmaps, limiting the choice of AL algorithms. We show that EGL++ is competitive in maximizing the performance of human pose models when full annotation is infeasible, with an added advantage of interpretability.

¹Elaborated in *Relation to Prior Art*, Related Work section

2. Related Work

Active Learning: Settles' active learning survey [38] remains a gold mine of information, detailing various classical methods. Classical uncertainty algorithms such as [20, 26] have a rich history, employing variations of entropy and maximum margin using softmax outputs of the model. Recent works include [29] which use GANs to generate high entropy samples for multi-class problems, and [21] which employs uncertainty techniques for region based segmentation.

Ensembles [2, 23, 30] build upon the Query by Committee paradigm, using multiple models to select the annotation set. [30] builds ensembles with artificial training data, [2] compares ensemble learning with recent active learning techniques and [23] uses cues from the softmax outputs in classification to show the use of maximum margin in a multi-class setting. Diversity algorithms [9, 37, 50] as the name suggests seek to incorporate a high degree of variation in the sampled set of unlabelled images. Core-set [37] remains a popular approach that utilizes the linearly separable embedding space in the penultimate layer of classification networks. Uncertainties in deep learning is an exciting area of research, summarized in Gawlikowski *et al.* [12]. Bayesian techniques [7, 10, 11, 14, 22, 25] provide theoretical results to quantify uncertainty with the network predictions. These approaches rely on estimating the aleatoric and epistemic uncertainties to quantify the ambiguity associated with data and network inferences. Amersfoort *et al.* [47] borrows ideas from RBF networks and computes uncertainty in a single forward pass. Other approaches include Learning Loss [42, 51] which use an auxiliary network to predict the 'loss' for an image.

Expected Gradient Length: Expected gradient length (EGL) [38] utilizes the gradient norm in determining the most informative samples. Early works [39, 40, 52, 53] successfully leveraged EGL in classification and text. Huang *et al.* [17] show that expected gradient length is a consequence of reducing the variance of the estimator over the testing set. Zhang *et al.* [54] use expected gradient length for sentence and document classification using CNNs. While previous methods worked with discrete outputs, [4, 5] demonstrates the use of expected gradient length in regression.

Human Pose Estimation: Single person human pose estimation (HPE) has been widely studied [3, 31, 45, 46] in literature. Popular architectures (Fig: 1) draw inspiration from U-Net [35] extracting features from multiple scales using a top-down approach. Human pose models take as input 3-channel RGB images and regresses a 2-D heatmap (one for each joint), denoting the location for the joint. Active learning for human pose estimation was first discussed in [27], which proposed multi-peak entropy by computing, normalizing, and performing softmax over the local maxima present in the heatmaps.

Relation to Prior Art: Our work is geared towards expected gradient length for regression and succeeds Cai *et al.* [4, 5]. We theoretically derive the formulation, supporting the experimental results in Cai *et al.* We go one step further and show that expected gradient length in regression is equivalent to prediction uncertainty. Our algorithmic contribution (EGL++) approximates ensembles with a single deterministic network by utilizing previously labelled samples within the EGL framework. Both DUQ [47] and EGL++ use distances as a measure of uncertainty. However, DUQ learns class specific weight matrices to encode features in classification, whereas EGL++ uses learned representations directly in t-SNE. Gradient penalty in DUQ serves as a regularizer to prevent feature collapse, whereas gradients form the core of EGL++. (*Details in the supplementary material*)

Challenges with Human Pose: While previous literature explores EGL for classification and regression, we extend the EGL framework to include human pose estimation (HPE). Our contribution is non-trivial since the challenges posed by HPE, namely a fully convolutional architecture regressing 2D heatmaps limits the choice of active learning algorithms. Bayesian uncertainty uses dropouts, causing convolutional architectures to suffer from strong regularization [6]. Entropy based approaches are shown to be less effective [51]. HPE models are not deployed as ensembles and do not use voting. Core-set uses the fact that classification enforces linear separability in the penultimate layer [37], which does not hold true for HPE. Learning loss [42, 51] specializes in detecting *lossy* images and is not suitable for generalized active learning. Some close works include bayesian hand pose estimation [6], however the network architecture and inputs differ significantly from human pose estimation. Aleatoric uncertainty has been effectively used in body joint occlusions [15], however the algorithm does not cater to active learning. Concerns [37, 51] also remain on scalability of bayesian methods to large datasets.

2.1. Revisiting Expected Gradient Length

The intuition supporting Expected Gradient Length (EGL) is simple: using gradient as a measure of change imparted in the model by a given sample. While a converged network incurs negligible gradient across the training set, a sample from a different distribution incurs a large gradient since it imparts new information to the model.

Given the lack of labels for the unlabelled pool, EGL algorithms define a distribution over the labels to compute the expected gradient. The formulation for EGL [38, 39] is shown in Eq: 1:

$$x^* = \arg \max_x \sum_i P(y_i|x; \theta) \|\nabla_{\theta} l(\mathcal{L} \cup \{x, y_i\}; \theta)\| \quad (1)$$

Settles and Craven [38] use the softmax outputs as a distribution over the labels \bar{y} , with the gradient computed assum-

ing $y_i \in \bar{y}$ is the correct label for x . However, this approach is computationally expensive, and is also not representative of tasks where discrete probabilistic outputs are not available such as in regression.

To overcome the lack of labels, Cai *et al.* [5] extended expected gradient length to regression by building an ensemble of models. Let $\mathcal{F}^K = \{f_1, f_2 \dots f_K\}$ represent a set of hypothesis obtained by training on subsets of labelled data $z = \{(x_1, y_1), (x_2, y_2) \dots (x_N, y_N)\}$ and f_z obtained by training on the entire pool of labelled data. Then, the sampling formulation can be represented as:

$$x^* = \arg \max_x \frac{1}{K} \sum_{k=1}^K \|(f_z(x) - f_k(x))x\| \quad (2)$$

Eq: 2 is intuitively defined and uses a committee of hypothesis to approximate the change induced by the sample x in the original model f_z . Therefore, the samples picked for annotation are those where there is higher level of disagreement between the model f_z and various weak learners f_k .

3. Theory

While Cai *et al.* [4, 5] provides experimental evidence using Eq: 2, we derive and show that Eq: 2 is a special case of the closed form solution obtained for EGL - regression. Further exploration reveals that the EGL framework too unifies aleatoric and epistemic uncertainty as done in [22]. Perhaps Bayesian uncertainty and EGL are two sides of the same coin?

3.1. The No Free Lunch Theorem

Let $\mathcal{Z} \in \{\mathcal{X} \times \mathcal{Y}\}$ represent the domain and z_{obs} denote the observed samples. We model the true distribution p over \mathcal{Z} , parameterized by θ_0 as $p(x, y|\theta_0) = p(y|x, \theta_0)p(x)$. Subsequently, [17, 28, 44] define a new distribution q for the observed values z_{obs} as $q(x, y|\theta_0) = p(y|x, \theta_0)q(x)$ where $q(x)$ reflects the observed samples x in z_{obs} . However, is it correct to assume that θ_0 alone describes z_{obs} ?

We argue that this assumption violates the No Free Lunch theorem [41]: Let $\mathcal{A}(z_{obs}) \rightarrow \theta$ represent the empirical risk minimizer (ERM) on the observed samples $z_{obs} \sim p_z$ which yields a hypothesis θ from the set of *unbiased* hypotheses Θ . Then for any hypothesis θ returned by the learner \mathcal{A} , there exists a distribution p_z on which it fails.

A simpler, equivalent statement is: *there is no universal learner, no learner can succeed on all learning tasks* [41]. This is based on two factors: 1) The ambiguity associated with the unknown distribution p_z and 2) The set of unbiased hypotheses Θ not reflecting prior knowledge. Any learning algorithm \mathcal{A} requires prior knowledge in the form of bias in the hypotheses Θ . The lack of a bias indicates an infinite number of unique hypothesis that perfectly agree on z_{obs} , but disagree to varying degrees on the remaining sample

space. Since we can only model z_{obs} and not p_z , this implies any of the ERM hypotheses can be the true hypothesis, which forms the crux of the theorem.

The No Free Lunch theorem holds significance in the Active Learning paradigm. Active Learning assumes a small set of labelled samples $z_{obs} \sim p_z$, an initial trained model θ and a hypotheses class Θ (such as a neural network) with low bias. Since z_{obs} does not adequately represent p_z and the hypotheses set Θ has low bias relative to z_{obs} , the ERM learner potentially returns multiple θ with a low empirical risk as the initial trained model. Hence, while we can safely assume that θ_0 parameterizes the true distribution p , the assumption that θ_0 alone represents the observed samples is incorrect. Therefore, our distribution over the observed samples is in fact an expectation over all the multiple plausible hypotheses returned by an ERM learner that explain the observed samples z_{obs} .

$$\begin{aligned} q(x, y) &= \mathbb{E}_\theta [q(y|x, \theta)q(x)] \\ &= \int_\theta q(y|x, \theta)q(x)\pi(\theta|z_{obs})d\theta \end{aligned} \quad (3)$$

Although representing the parameters as a distribution is standard practice in Bayesian analysis, the No Free Lunch theorem within the scope of active learning explains the existence of multiple hypothesis which can be quantified using bayesian analysis.

3.2. Fisher Information

The proof proceeds further using a well known result on the asymptotic convergence of model parameters to its true value: $\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{D} \mathcal{N}(0, I_q^{-1}(\theta^*))$. Related works [17, 39] argue that one way of converging to the true parameters (LHS) is by maximizing the Fisher Information with respect to the training distribution q . (minimizing the inverse). This is equivalent to maximizing $\mathbb{E}_{q(x,y)}[\nabla_\theta l(x, y, \hat{\theta})\nabla_\theta^T l(x, y, \hat{\theta})]$. Using Eq: 3, we show that this is equivalent to maximizing:

$$q^* = \arg \max_q \int_x q(x) \int_y \int_\theta q(y|x, \theta) \pi(\theta|z_{obs}) \|\nabla_{\theta_0} l(x, y, \theta_0)\|^2 d\theta dy dx \quad (4)$$

Since integrating over the parameter space is intractable, we approximate $\pi(\theta|z_{obs})$ by bootstrapping and building an ensemble of learners f_i parameterized by θ_i , where $z_i \subset z_{obs}$. Next, we define f_z to be trained on all z_{obs} , with θ_z our best estimate of the true parameters θ_0 . We note that finding a training distribution q that maximizes the expected gradient is same as including the sample x having the highest expected gradient in the training set [17]:

$$x^* = \arg \max_x \frac{1}{K} \sum_{k=1}^K \int_y q(y|x, \theta_k) \|\nabla_{\theta_z} l(x, y, \theta_z)\|^2 \quad (5)$$

3.3. Linear Regression

Statistical linear regression results [33] show that $q(y|x, \theta_k)$ is a gaussian distribution $\mathcal{N}(\mu_k(x), \sigma_k(x))$.

$$x^* = \arg \max_x \frac{1}{K} \sum_{k=1}^K \int_{-\infty}^{\infty} \mathcal{N}(y, \mu_k, \sigma_k)[(f_z(x) - y)x]^2 dy \quad (6)$$

Note that we have substituted $l(x, y, \theta_z)$ in Eq: 5 with the regression mean square error for the model f_z . On comparing with Cai *et al.* (Eq: 2), we note that the gradient term in our formulation is squared as a result of maximizing the Fisher Information. Also Eq: 2 is equivalent to Eq: 6, if exactly one sample y_i is drawn from $q(y|x, \theta_i)$ for each $i \in K$ corresponding to the mode of the distribution.

Eq: 6 has a closed form solution. Let $(y - f_z(x))^2 = [(y - \mu_k) + (\mu_k - f_z)]^2$. The integral can be reduced to: $\int \mathcal{N}(y, \mu_k, \sigma_k)(y - \mu_k)^2 dy + 2 \int \mathcal{N}(y, \mu_k, \sigma_k)(y - \mu_k)(\mu_k - f_z)dy + \int \mathcal{N}(y, \mu_k, \sigma_k)(\mu_k - f_z)^2 dy$. The first term corresponds to the variance of the normal distribution, the second term results in a zero as the means cancel out and the last term coefficient is independent of y , resulting in $(\mu_k - f_z)^2$. The final closed form solution completing our derivation is:

$$x^* = \arg \max_{x \in \mathcal{U}} \frac{\|x\|^2}{K} \sum_{k=1}^K \sigma_k^2(x) + (\mu_k(x) - f_z(x))^2 \quad (7)$$

3.4. Non-Linear Regression

Bayesian linear regression allows us to model $q(y|x, \theta_i)$ as a Normal distribution, simplifying our analysis. Can we follow a similar approach to extend our analysis to non-linear regression?

Aleatoric Uncertainty

Regression assumes that the residuals are normally distributed $\epsilon \sim \mathcal{N}(0, \sigma)$. [22, 32] show that if σ is made learnable for each input, this reduces to computing the heteroscedastic aleatoric uncertainty for the input x .

$$l(x, y, \theta) = \frac{1}{2\sigma(x)^2} \|y - \hat{y}\|^2 + \frac{1}{2} \log \sigma(x)^2 \quad (8)$$

If we let $\hat{y}, \sigma = f_\theta(x)$, Eq: 8 is a consequence of minimizing the negative log-likelihood over the normal distribution.

Gaussian Process

Gaussian Process is a widely used statistical technique that imposes a distribution over the functions which can describe the observed data. The essence of Gaussian process is to represent each data point as a sample from a unique random

variable, with the covariance between random variables being denoted by a kernel matrix \mathcal{K} . While a complete discussion is beyond the scope of the paper, Gaussian Process regression allows us to model: $q(y|x, z_{obs}) \sim \mathcal{N}(\mu, \sigma)$ where μ, σ have a closed form solution [8]:

$$\begin{aligned}\mu &= \mathcal{K}(x, x_{obs})\mathcal{K}(x_{obs}, x_{obs})^{-1}\bar{y}_{obs} \\ \sigma &= \mathcal{K}(x, x) - \mathcal{K}(x, x_{obs})\mathcal{K}(x_{obs}, x_{obs})^{-1}\mathcal{K}(x_{obs}, x)\end{aligned}\quad (9)$$

Integrating Gaussian Process into the EGL framework is trivial. Instead of using parametric models θ_i to form an ensemble (Eq: 5), we can use an ensemble of Gaussian Process q_i fitted on $z_i \subset z_{obs}$.

Formulation

Both Aleatoric as well as Gaussian Process allow us to model $q(y|x, z_{obs}) \sim \mathcal{N}(\mu, \sigma)$. Also, the gradient for non-linear regression (from Eq: 5) can be split as: $\nabla_{\theta_z} l(x, y, \theta_z) = (y - f_z(x))\nabla_{\theta_z} f_z(x)$, where $\nabla_{\theta_z} f_z(x)$ is independent of the ensemble. Hence, our analysis for linear holds true for non-linear regression too:

$$x^* = \arg \max_{x \in \mathcal{U}} \frac{\|\nabla_{\theta_z} f_z(x)\|^2}{K} \sum_{k=1}^K \hat{\sigma}_k^2(x) + (\mu_k(x) - f_z(x))^2 \quad (10)$$

3.5. Comparison with Predictive Uncertainty

Kendall and Gal [22] define predictive variance as:

$$Var(y) \approx \frac{1}{K} \sum_{k=1}^K \hat{\sigma}_k^2 + \frac{1}{K} \sum_{k=1}^K f_k^2(x) - \left(\frac{1}{K} \sum_{k=1}^K f_k(x) \right)^2 \quad (11)$$

We highlight that both predictive uncertainty (Eq: 11) and EGL (Eq: 7, 10) use two measures of uncertainty. The first corresponds to local uncertainty (σ_k): what an individual model believes the uncertainty is. The second corresponds to global/epistemic uncertainty ($\mu - f$): measure of disagreement between models (from an ensemble/dropout).

If we consider Aleatoric - EGL, then approximating $\|\nabla_{\theta_z} l(x, y, \theta_z)\| \approx c\|(y - f_z(x))\|$ allows us to derive predictive uncertainty from EGL. The basis of this approximation is that *chain rule* ensures that the gradient is highly correlated with the residual.

3.6. Significance

Although Cai *et al.* [4, 5] provided experimental evidence for expected length in regression, their formulation was intuitively driven. Our theory supports this experimental evidence, highlighting that the original formulation (Eq: 2) is a special case of our derived formulation. We then proceed to show that Expected Gradient Length for regression can unify both aleatoric and epistemic uncertainty, proving equivalence with [22].

4. Algorithm: EGL++

Our theory shares the following assumptions with bayesian uncertainty: 1) The underlying task supports ensembles/dropouts 2) $q(y|x, \theta)$ can be determined analytically. While our assumptions hold true for most regression tasks, can we adopt EGL for active learning to tasks where our assumptions are violated? Our discussion in Sec: 2 highlights our inability to use dropouts or ensembles with human pose estimation (HPE). Additionally, modelling $q(y|x, \theta)$ as aleatoric uncertainty is tedious since HPE models regress 2-D heatmaps.

To circumvent these challenges, we propose EGL++, an alternative interpretation of the EGL framework. EGL++ approximates the effect of ensembles by leveraging previously observed labels y_j as potential labels for unlabelled images x_i . We facilitate this by representing each sample as an *image-embedding-(label/prediction)* triplet, and use t-SNE to quantify the neighborhood in the embedding space. This approach makes EGL++ highly interpretable, since it quantifies our intuition that similar inputs have similar representations and thus predictions.

Therefore, in the absence of a closed form solution, EGL++ solves:

$$x^* = \arg \max_{x \in \mathcal{U}} \sum_{n=1}^{n=N} q_{tsne}(y_n|x, \theta) \|\nabla_{\theta} l(x, y_n, \theta)\|^2 \quad (12)$$

We discuss each of the components: $q_{tsne}(y|x, \theta)$ and $\|\nabla_{\theta} l(x, y_n, \theta)\|^2$ in greater detail:

Conditional Probability

Computing $q(y|x, \theta)$ can be further split into computing network representations h for the set of images x , and modelling the relation between these representations h . Introducing network representations has two major benefits. Not only do smaller representations improve compute time, network embeddings allow us to model semantic relations established by the neural network.

Dimensionality reduction algorithms are useful in modelling the spatial representation between high dimensional samples. While classical algorithms such as Locally Linear Embedding [36] beautifully describe the sample space, they have high time complexity $O(n^3)$ due to eigen-decomposition. Hence, we turn our attention to t-SNE, a popular visualization algorithm with $O(n^2)$ time complexity and optimized for GPU usage.

t-SNE: t-SNE [16, 43, 48, 49] provides a convenient method to model $q(y|x, \theta)$ in high-dimensional space. Let h_i represent the embedding for the i^{th} sample. Then the probability of h_j being a neighbor of h_i is:

$$q(h_j|h_i, \theta) = \frac{\exp(-\|h_i - h_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|h_i - h_k\|^2/2\sigma_i^2)} \quad (13)$$

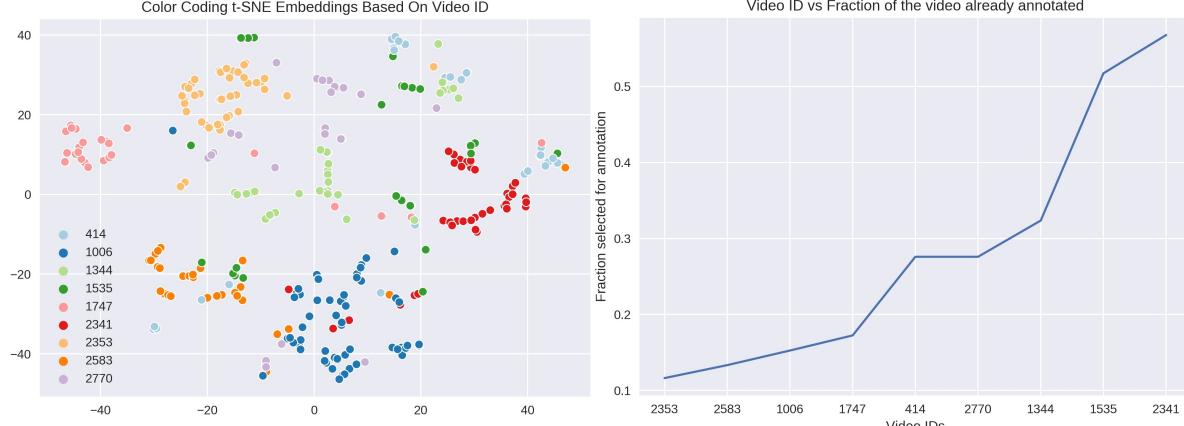


Figure 2. (Left) t-SNE visualization of MPII, color-coded by video. (Right) Fraction of frames chosen per video. Heavily sampled videos (1344, 1535, 2341) either have multiple clusters across the embedding space or have overlapping representations with other videos. In contrast, lightly sampled videos (2353, 2583) have embeddings concentrated in a small space, needing fewer labelled samples.

The neighborhood for every sample i is quantified using a gaussian distribution centered at h_i with variance σ_i^2 . Perplexity [49] tunes the variance to account for variable density in the embedding space, thereby determining the spatial extent of the neighborhood for h_i .

Since we know the ground truth y_j for each embedding h_j when image x_j is from the labelled set, our heuristic lies in approximating $q(h_j|h_i, \theta)$ with $\hat{q}(y_j|h_i, \theta)$. This heuristic allows us to leverage previously annotated data and quantify the probability of y_j being the true label for x_i . We divide our embeddings into labelled $h_{\mathcal{L}}$ and unlabelled sets $h_{\mathcal{U}}$ where we identify the neighborhood for each $h \in \mathcal{U}$ in terms of all $h \in \mathcal{L}$. This results in a conditional probability matrix $q(h_{\mathcal{L}}|h_{\mathcal{U}}, \theta) = \hat{q}(y_{\mathcal{L}}|h_{\mathcal{U}}, \theta)$ of shape $(|\mathcal{U}| \times |\mathcal{L}|)$, allowing us to compute the expectation for $\|\nabla_{\theta} l(x_{\mathcal{U}}, y_{\mathcal{L}}, \theta)\|^2$.

Gradient Computation

In a utopian world where deep learning models train instantaneously, gradient computations would nary have been an issue. Alas, that is not the case, and EGL++ needs fast yet efficient heuristics for gradient computation. We use two approximations to speed up EGL++. The first approximation is in limiting the number of neighbours per sample. Previously we had computed a conditional probability matrix $\hat{q}(y_j|h_i, \theta)$, which is used in calculating the expected gradient length $\sum_j \hat{q}(y_j|h_i, \theta) \|\nabla_{\theta} l(x_i, y_j, \theta)\|^2$ for the unlabelled sample x_i . Computing this value for all labels y_j is an expensive task, especially when most values of $\hat{q}(y_j|h_i, \theta) \approx 0$. Instead, we restrict y_j to be among the top-K most probable values given by $\hat{q}(y_j|h_i, \theta)$, limiting the number of gradient computations needed.

The second approximation uses Goodfellow's technique for efficient per example gradient computation [13]. Computing gradients per example is difficult since popular deep

learning frameworks aggregate gradients across the mini-batch. Goodfellow's solution uses intermediate representations to compute the gradient in convolutional and linear layers for each sample x_i in the mini-batch . While this is not the true gradient for x , (for e.g., no batch normalization gradients) this heuristic allows for fast computations of gradients in a single forward pass.

Algorithm 1: EGL++

```

Input: Human pose estimator ( $f$ ), Number of
       neighbors ( $K$ ), Budget ( $B$ )
Output: Unlabelled images ( $x_{\mathcal{U}}^*$ ) for annotation
Data: Labelled ( $x_{\mathcal{L}} - y_{\mathcal{L}}$ ) and Unlabelled ( $x_{\mathcal{U}}$ )
       images
1  $h_{\mathcal{U}} = f_{enc}(x_{\mathcal{U}})$            // encodings for unlabelled
2  $h_{\mathcal{L}} = f_{enc}(x_{\mathcal{L}})$            // encodings for labelled
3 Compute matrix  $q(h_j|h_i, \theta)$  using Eq: 13, where
    $i \in \mathcal{U}$  and  $j \in \mathcal{L}$            // Size  $(|\mathcal{U}| \times |\mathcal{L}|)$ 
4  $\hat{q}(y_j|h_i, \theta) = q(h_j|h_i, \theta)$  // Heuristic,  $h_j \rightarrow y_j$ 
5 Initialize  $\{\theta_{fast} : \theta_0\}$  layers  $\in$  linear, conv}
6  $(\hat{q}(y_j|h_i, \theta), y_j) = \text{top - } K(\hat{q}(y_j|h_i, \theta), y_j)$ 
   /* New  $\hat{q}$ ,  $\nabla$  matrix size:  $|\mathcal{U}| \times |\mathcal{K}|$  */
   /* CPU multiprocessing - parallelize for */
7 for  $x_i$  in  $\mathcal{U}$  do
   /* GPU parallel across  $j$  for each  $i$  */
8   for  $y_j$  in  $\mathcal{L}$  do
     /* Fast Gradient Approximation */
9     Compute gradient norm  $\|\nabla_{\theta_{fast}} f(x_i), y_j\|^2$ 
   /* Expectation is hadamard product  $\hat{q} \circ \nabla$ ,
      reduce summed along  $j$ . Shape:  $|\mathcal{U}|$  */
10   $\bar{\mathbb{E}}_y[\nabla_{\theta}(y, x_{\mathcal{U}}, \theta)] =$ 
     $\sum_j \hat{q}(y_j|h_i, \theta) \circ \|\nabla_{\theta_{fast}} f(x_i), y_j\|^2$ 
11 return  $x_{\mathcal{U}}^*$ : Return samples corresponding to top -  $B$ 
       expected gradients from  $\bar{\mathbb{E}}_y[\nabla_{\theta}(y, x_{\mathcal{U}}, \theta)]$ 

```

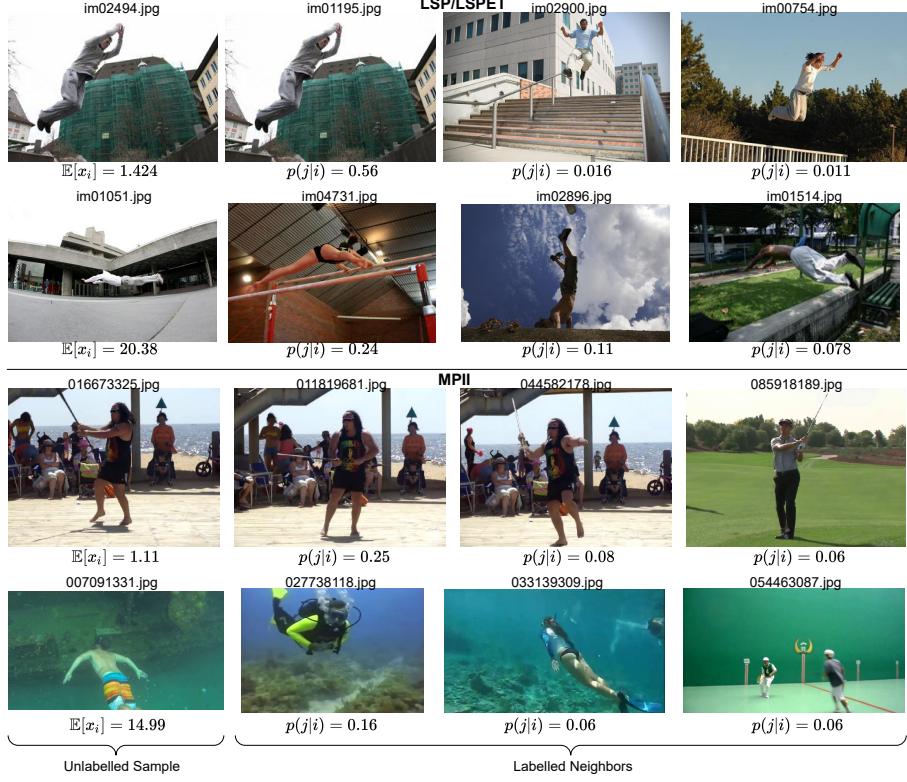


Figure 3. *Interpretability*: EGL++ quantifies the intuition that similar inputs have similar representations. Unlabelled images (first column) with the smallest expected gradient already have similar annotated images. In contrast, unlabelled images with the highest expected gradient share limited semantic similarity with its nearest neighbours. (*More figures in the supplementary material*)

5. Experiments and Results

Goal: EGL and Bayesian uncertainty provide theoretical results, but what happens when the underlying assumptions aren't feasible? We turn the spotlight on active learning for human pose estimation (HPE), with our previous discussion summarizing the associated challenges with the architecture (Sec: 2). While EGL++ extends the EGL paradigm to human pose, can we achieve competitive performance in comparison to popular active learning algorithms for HPE? *Keeping in mind practical applications such as an incoming data stream, can EGL++ maximize the model performance per set of images annotated?*

Experimental Design: Our code (in PyTorch [34]) is available on <https://github.com/meghshukla/ActiveLearningForHumanPose>. We compare EGL++ with three state-of-the-art algorithms in active learning for human pose: Coreset [37], Multi-peak entropy [27] and Learning Loss [51]. We simulate multiple cycles of active learning, with each cycle selecting 1000 new images for annotation. Base models trained on an initial random 1000 images are shared by all algorithms. Following [27, 51], we use MPII [1] and LSP/LSPET [18, 19] human pose datasets. The MPII dataset consists of images corre-

sponding to everyday human activity, whereas LSP emphasizes on sporting activities. For MPII, we report our results on the Newell validation split [31], as done in [27, 51]. We use the LSP authors' data split; the first 1000 images used for training the base models, and the last 1000 images acting as the testing set. The entire LSPET dataset is used as the unlabelled pool for active learning sampling. We use the smaller two-stacked Hourglass [31] with PCKh of 88% (following [51]) as the human pose estimator. Following Multi-Peak Entropy, we extract single persons into separate images if the underlying image contains multiple persons. We also use standard evaluation metrics: PCKh@0.5 for MPII and PCK@0.2 for LSP-LSPET.

5.1. Qualitative Assessment

Fig: 2 confirms our hypothesis that similar inputs have similar representations. Videos with multiple clusters across the embedding space usually exhibit high variation in content, therefore requiring a higher number of selected frames to represent the different clusters. Similarly, videos with overlapping representations have a higher rate of sampling. Videos with concentrated grouping of embeddings indicate content with limited movements, and can be represented with fewer labelled samples.

MPII Newell Validation Split: Mean+Sigma (5 runs), one-tailed paired t-test (vs EGL++) at 0.1 significance value															
#images →	2000			3000			4000			5000			6000		
Methods	μ	σ	p-value	μ	σ	p									
Random	75.95	0.55	0.003	78.33	0.65	0.012	80.31	0.91	0.006	81.35	0.41	0.001	82.23	0.74	0.007
Core-set [37]	76.61	0.6	0.0047	79.24	0.7	0.245	81.25	0.67	0.072	82.23	1.14	0.123	82.97	1.11	0.064
Multi-peak [27]	76.74	0.61	0.0054	79.56	0.46	0.462	81.19	0.31	0.063	82.61	0.5	0.093	83.11	0.71	0.062
Learning Loss [51]	76.28	0.76	0.0276	79.27	0.52	0.185	81.35	0.35	0.152	82.94	0.44	0.319	83.79	0.46	0.053
EGL++ (ours)	77.28	0.63	-	79.58	0.33	-	81.53	0.51	-	83.07	0.25	-	84	0.38	-

LSP Test Split: Mean+Sigma (5 runs), one-tailed paired t-test (vs EGL++) at 0.1 significance value															
#images →	2000			3000			4000			5000			6000		
Methods	μ	σ	p-value	μ	σ	p									
Random	80.34	0.31	0.285	81.81	0.24	0.297	82.68	0.32	0.138	83.35	0.36	0.095	84.13	0.16	0.029
Core-set [37]	79.69	0.82	0.043	81.41	0.45	0.096	82.25	0.39	0.021	83.11	0.38	0.032	83.73	0.31	0.006
Multi-peak [27]	80.36	0.4	0.225	81.48	0.53	0.125	82.63	0.23	0.119	83.29	0.2	0.036	84.3	0.44	0.063
Learning Loss [51]	79.58	0.39	0.002	81.39	0.34	0.071	82.31	0.42	0.038	83.31	0.25	0.029	84.2	0.53	0.098
EGL++ (ours)	80.49	0.45	-	81.91	0.27	-	83.03	0.43	-	83.91	0.51	-	84.68	0.36	-

Table 1. (Active Learning Simulation) While [27] establishes active learning for human pose, we focus on a more practical challenge: maximizing model performance per set of images annotated. The quantitative results and p-value from t-test for significance indicate that EGL++ is competitive, performing equivalent to if not better than its peers across all active learning cycles.

Fig: 3 highlights the interpretability associated with EGL++. Images with the lowest EGL score are well represented in the training set, leaving little ambiguity in the potential label. In contrast, images with the highest EGL score do not share semantic similarities with its neighbours, leaving considerable ambiguity over the potential label. We also note that EGL++ fairs well in interpretability compared to its peers. Learning Loss and its variants predict the 'loss' for an image but lacks explanation. Core-set encourages diversity in the sampled set, but fails to explain the selection of individual images in the set. Multi-peak entropy has higher explainability since it captures the ambiguity within the heatmap, however the approach does not reason for the ambiguity given an image.

5.2. Quantitative Assessment

Test for statistical significance: We report the mean (μ), std. dev. (σ) and **p-value**, the latter obtained from the paired t-test to measure the significance of our results at a 90% confidence level [24]. The paired t-test is advantageous since: 1) It accounts for random good/bad initializations of the base models (mean by definition averages it) 2) It is independent of the number of samples selected per cycle. Since single person datasets have limited training examples (MPII: ~20K, LSP-LSPET: ~11K), the number of samples selected per cycle is small. This directly translates into smaller jumps in mean across all methods, as is evident from Tab: 1.

Comparison: MPII and LSP-LSPET allow us to examine the behavior of the algorithms under two different conditions: when the labelled and unlabelled pool follow the same distribution (MPII), and when the pools differ (LSP - initial labelled, LSPET - unlabelled). While active learning algorithms perform significantly better (based on p-value) than random sampling on MPII, these algorithms struggle in the initial cycles of LSP-LSPET due to the change from LSP to LSPET. Also, learning loss being a deep learning based method improves in performance as the number of

labelled instances increase. We also observe that Core-Set does not extend well to human pose, since linear separability of embeddings from the penultimate layer is not enforced by human pose estimation. Multi-peak entropy fares well in the initial phases of active learning, with its performance declining in the subsequent stages. Entropy for active learning has been well studied in [51], and while multi-peak entropy fares better than vanilla entropy, it shares the same weaknesses as its predecessor. Like its peers, EGL++ too struggles to outperform random sampling when the model shifts from LSP to LSPET. However, EGL++ is the fastest to recover, performing better than its peers based on the p-value. In general, we observe that at the minimum, EGL performs as good as the best active learning technique.

6. Conclusion

This paper explores expected gradient length (EGL) in regression from a theoretical perspective, making two key contributions. The first contribution derives the closed form solution for EGL-regression, thereby supporting the experimental evidence in literature. Specifically, we show that the intuitive formulation in literature is in fact a special case of our derived solution. *Our second contribution lies in highlighting that EGL-regression unifies both: aleatoric and epistemic uncertainty, allowing us to draw parallels with predictive uncertainty.* With EGL++, we adopt the EGL framework to tasks where creating ensembles or computing a distribution over the labels is infeasible. EGL++ approximates the effect of ensembles by quantifying the probability of existing labels being the true label for an unlabelled image. This approximation allows us to extend the EGL framework to human pose estimation. Our experiments validate that EGL++ is interpretable and competitive in comparison to popular active learning algorithms for human pose estimation.

Acknowledgement: I thank Brijesh Pillai and Partha Bhattacharya, MBRDI, for providing funding for this work.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [2] William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.
- [3] Adrian Bulat, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. Toward fast and accurate human pose estimation via soft-gated skip connections. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 101–108. IEEE Computer Society.
- [4] Wenbin Cai, Muhan Zhang, and Ya Zhang. Batch mode active learning for regression with expected model change. *IEEE Transactions on Neural Networks and Learning Systems*, 28(7):1668–1681, 2017.
- [5] W. Cai, Y. Zhang, and J. Zhou. Maximizing expected model change for active learning in regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60, 2013.
- [6] Razvan Caramalau, Binod Bhattacharai, and Tae-Kyun Kim. Active learning for bayesian 3d hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3419–3428, 2020.
- [7] Stefan Depeweg, Jose Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, Stockholm, Sweden, 2018.
- [8] Chuong Do and Honglak Lee. cs229-gaussian_processes.pdf. http://cs229.stanford.edu/section/cs229-gaussian_processes.pdf.
- [9] Ehsan Elhamifar, Guillermo Sapiro, Allen Yang, and S Shankar Sasrt. A convex optimization framework for active learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 209–216, 2013.
- [10] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.
- [11] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*, 2017.
- [12] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks, 2021.
- [13] Ian Goodfellow. Efficient per-example gradient computations, 2015.
- [14] Théo Guénais, Dimitris Vamvourellis, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Bacoun: Bayesian classifiers with out-of-distribution uncertainty. *arXiv preprint arXiv:2007.06096*, 2020.
- [15] Nitesh B Gundavarapu, Divyansh Srivastava, Rahul Mitra, Abhishek Sharma, and Arjun Jain. Structured aleatoric uncertainty in human pose estimation. In *CVPR Workshops*, volume 2, 2019.
- [16] Geoffrey Hinton and Sam T Roweis. Stochastic neighbor embedding. In *NIPS*, volume 15, pages 833–840. Citeseer, 2002.
- [17] Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients, 2016.
- [18] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, 2010. doi:10.5244/C.24.12.
- [19] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.
- [20] Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2372–2379. IEEE, 2009.
- [21] Tejaswi Kasarla, Gattigorla Nagendar, Guruprasad M Hegde, Vineeth Balasubramanian, and CV Jawahar. Region-based active learning for efficient labeling in semantic segmentation. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1109–1117. IEEE, 2019.
- [22] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017.
- [23] Christine Körner and Stefan Wrobel. Multi-class ensemble-based active learning. In *European conference on machine learning*, pages 687–694. Springer, 2006.
- [24] Michelle Lacey. Tests of significance. https://bit.ly/statistical_testing_yale.
- [25] Quoc V Le, Alex J Smola, and Stéphane Canu. Heteroscedastic gaussian process regression. In *Proceedings of the 22nd international conference on Machine learning*, pages 489–496, 2005.
- [26] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR’94*, pages 3–12. Springer, 1994.
- [27] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4363–4372, 2017.
- [28] Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.
- [29] Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3071–3079, 2020.

- [30] Prem Melville and Raymond J Mooney. Diverse ensembles for active learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 74, 2004.
- [31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hour-glass networks for human pose estimation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 483–499, Cham, 2016. Springer International Publishing.
- [32] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.
- [33] Dmitry Panchenko. Mit ocw - 18.650:lec31.pdf. http://bit.ly/stats_18-443_lec31, 2003.
- [34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
- [36] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
- [37] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.
- [38] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.
- [39] Burr Settles and Mark Craven. An analysis of active learning strategies for sequence labeling tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1070–1079, 2008.
- [40] Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.
- [41] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [42] Megh Shukla and Shuaib Ahmed. A mathematical analysis of learning loss for active learning in regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3320–3328, June 2021.
- [43] Megh Shukla, Biplab Banerjee, and Krishna Mohan Buddhi-raju. LEt-SNE: A hybrid approach to data embedding and visualization of hyperspectral imagery. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3722–3726. IEEE, 2020.
- [44] Jamshid Sourati, Murat Akcakaya, Todd K Leen, Deniz Erdogmus, and Jennifer G Dy. Asymptotic analysis of objectives based on fisher information in active learning. *The Journal of Machine Learning Research*, 18(1):1123–1163, 2017.
- [45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [46] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.
- [47] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pages 9690–9700. PMLR, 2020.
- [48] Laurens van der Maaten. Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research*, 15(93):3221–3245, 2014.
- [49] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [50] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.
- [51] D. Yoo and I. S. Kweon. Learning loss for active learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 93–102, 2019.
- [52] Xinge You, Ruxin Wang, and Dacheng Tao. Diverse expected gradient active learning for relative attributes. *IEEE transactions on image processing*, 23(7):3203–3217, 2014.
- [53] Y. Yuan, S. Chung, and H. Kang. Gradient-based active learning query strategy for end-to-end speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2832–2836, 2019.
- [54] Ye Zhang, Matthew Lease, and Byron Wallace. Active discriminative text representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.