Exploring Weak-Supervision and Generative Models for Semantic Segmentation

Prithvijit Chattopadhyay
CS 8803
prithvijit3@gatech.edu

Ramprasaath Selvaraju
CS 8803
ramprs@gatech.edu

Viraj Prabhu
CS 8803
virajp@gatech.edu

1 Introduction

Semantic segmentation is the task of assigning each pixel in an image an object class. Accurately assigning semantic labels requires precisely pinpointing the borders of objects, which requires a much higher localization accuracy than other high-level visual recognition tasks such as image classification or object detection. This constraint naturally suggests the use of probabilistic graphical models to model uncertainty in predictions combined with robust and powerful deep convolutional approaches to obtain accurate and well-localized segmentation maps.

Additionally, given the complexity of this task, it seems reasonable to require a complex method trained on large-scale datasets with pixel-level annotations in a supervised fashion to achieve good results. In this project, we propose to overcome this requirement by operating on surrogate localizations obtained from a trained image classification model, without the need for expensive pixel level annotation. By enforcing simple Markovian constraints on noisy and low resolution pixel-labels, we aim to obtain comparable performance on this task. In addition, our proposed approach for the task also works in data-starved instances of the task at hand.

To explicitly model uncertainty for semantic segmentation, we take the task a step further and frame the problem as one of scene understanding. Scene understanding has traditionally held an important position in computer vision due to applications in perceiving, analyzing, and elaborating an interpretation of a real-time dynamic scene. We restrict our modeling domains - modalities - to that of physical scene understanding, which involves characterizing the kind and position (space) occupied by different objects in an image. We particularly deal with three modalities - image, class presence/absence and segmentation maps. Assuming no causal relation between the modalities, we learn a directed probabilistic model with a continuous latent variable as a parent to all the modalities. Since such models usually have intractable posterior distributions, we use gradient based techniques from the stochastic variational inference literature to learn the parameters of our encoding and decoding processes. In addition to performing joint inference over the latent variable we also support inference conditioned on a subset of modalities - allowing us to perform semantic segmentation and predict a distribution over possible scenes given a scene configuration. In data-starved regimes, such a model could also be used to generate datasets by performing a random walk in the space of the continuous latent variable.

This report is organized as follows. In Sec 2, we provide a broad overview of the supervised segmentation task, some of the recent and popular approaches to solve it. We then focus on some of the relevant works in context of our proposed approach and report corresponding results. Sec 3 describes the datasets and the evaluation metrics we use to elicit the performance of our proposed approaches. In Sec 4, we introduce some of the approaches/models which serve as groundwork for our models. We describe our proposed approaches with associated qualitative and quantitative results in secitons 5 and 6. Finally, in Sec 7 we provide details regarding our implementation.

2 Related Work

In this section, we discuss some of the relevant work that cover semantic segmentation, involved inference techniques in graphical models, interpretable visualizations of deep models, joint generative probabilistic models and associated learning and inference techniques. We shall cover each of the following in the subsequent paragraphs and share associated quantitative and qualitative results wherever applicable.

Fully supervised semantic Segmentation. Prior to the emergence of deep learning, most popular semantic segmentation systems relied on learning classifiers on top of hand-crafted features, often additionally incorporating context Carreira et al. [2012] and structured prediction techniques He et al. [2004], Krähenbühl and Koltun [2011]. In general, approaches such as TextonForest Shotton et al. [2008] and Random forest based classifiers Shotton et al. [2011] were very popular. More recently, popular approaches have employed deep learning, with different classes of approaches emerging for solving the combined segmentation and classification task. One class of approaches performs bottom-up segmentation followed by region classification via a deep convolutional neural network (DCNN) Girshick et al. [2014]. Another class of approaches separately obtains image segmentations and DCNN features for dense image labeling and combines them for the semantic segmentation task Farabet et al. [2013]. Another popular set of approaches directly trains DCNN's to provide dense category-level pixels in a fully convolutional fashion Long et al. [2015].

Inference in Graphical Models. Conditional Random Fields (CRFs) are the most prevalent graphical models used in the context of semantic segmentation. The key idea in most approaches is to assume Markovian constraints on nodes positioned on pixels with a categorical distribution over all the possible class-labels upon which we would like to perform the task of either refinement of segmentations, or segmentation itself, via some maximum likelihood estimate. The Markovian assumption naturally leads to the Gibbs energy formulation involving unary potentials (on nodes) and pairwise potentials (on neighboring nodes). In general, such a structured inference problem is NP-hard due to the sheer number of dimensions in the associated Ising model. As in Payet and Todorovic [2010], performing MCMC inference as in over such structures without any simplifying assumptions usually renders the problem intractable. Although tractability in relatively small non-loopy graphs can be achieved via dynamic programming inspired techniques such as variable elimination, this is not desirable from the perspective of semantic segmentation because of two primary reasons – firstly, sub-sampling the actual input ($\sim 10 \times 10$) and performing inference on that would lead to loss of visual context in terms of structure learning to associated labels that are too noisy to result in anything meaningful. Secondly, this still does not render the approach practical in terms of feasible time-complexity. One major stride made towards efficient inference has been to achieve sublinear time-complexity by performing approximate inference on such CRFs. One such approach that is particularly relevant is that of Krähenbühl and Koltun [2011]; i.e., performing inference on such CRFs by assuming unary potentials encoded by label classifiers and edge potentials encoded in a linear combination of Gaussian kernels, which enables one to perform efficient inference in such models via a mean field approximation. More details on this approach are provided in Sec.4.

Figure 1: The example above denotes the result of inference performed over a CRF for semantic segmentation via standard MCMC ($\sim$ 36 hours; left) Payet and Todorovic [2010] versus efficient inference by assuming pairwise potentials as combination of gaussian kernels and performing message passing ($\sim 0.2s$; right) Krähenbühl and Koltun [2011].

Graphical Models in Semantic Segmentation. Combining the salient aspects of many of the above approaches, Chen et al. [2016] propose combining deep convolutional neural networks with probabilistic graphical models for this task. Specifically, they overcome the issue of the poor localization property of higher level DCNN's responses by adding a Conditional Random Field (CRF) for refinement. While prior work has attempted combining the two, they have primarily employed locally connected CRF models while Chen et al. [2016] propose a fully connected CRF, representing each pixel as a node receiving unary potentials from the DCNN. In addition to this, the paper makes additional modeling contributions such as using atrous convolutions to explicitly control resolution of feature responses and increase receptive fields of pixels, and a novel spatial pyramid pooling scheme to perform segmentation at different scales, which when combined leads to very strong performance on a range of challenging semantic segmentation datasets including PASCAL VOC-2012, PASCAL-Context, PASCAL-Person-Part, and Cityscapes.

Interpretable Visualizations. Generating visual support for decisions made by neural networks, i.e., *visual explanations* has been a rich line of work in itself in the past couple of years. Recently, Selvaraju et al introduced GradCAM Selvaraju et al. [2016], a technique for producing "visual explanations" for decisions from CNN-based models. They propose to do this by computing the gradient of the score for a particular class with respect to convolutional layer activations. This provides an importance score of every convolutional layer neuron for a particular decision. A weighted sum of network activations provides a rough heatmap of where the model 'looks' in order to predict the class. This resulting heatmap is of a very low resolution (14 x 14) and only concentrates on the most discriminative aspects in the image. These visual heatmaps are relevant in the context of our project goal of semantic segmentation as these can act as noisy, pixel-level `surrogate` annotations that one can utilize in a low-data regime or even in the complete absence of annotated data. Since, these localization(s) are coarse, noisy but class-discriminative, efficient inference on CRFs (used to refine object boundaries via node and edge potentials) is a critical component to generate semantic segmentation predictions. In addition, these can be obtained for networks trained for classification (for which obtaining annotations is cheap in general). Following this, we describe our approach to combine relevant components from each of these to perform weakly supervised image segmentation and discuss baseline results in context.

Generative Models. Advances in generative models for visual understanding in recent years can be broadly categorized into two kinds of approaches - Generative Adversarial Networks (GANs) Goodfellow et al. [2014] and Variational Auto-encoders (VAEs) Kingma and Welling [2013]. We particularly focus on approaches along the lines of VAEs. Introduced in Kingma and Welling [2013], VAEs are directed probabilistic models with continuous latent variables (hence with intractable posteriors over the latent space) that are learned via gradient based approaches by optimizing the variational lower bound (Stochastic Variational Inference Hoffman et al. [2013]) on the marginal likelihood of the observed data. From an optimization perspective, gradient propagation is usually a problem since learning requires propagating them through a sampling procedure. Classical works in this domain have tried to address this problem over the years via different techniques - policy gradients or the log-derivative trick Sutton et al. [2000], the reparameterization trick Kingma and Welling [2013], Kingma et al. [2015], Burda et al. [2015] and so on. The primary concerns while optimizing the evidence lower bound (ELBO) in such models are the nature of gradients (biased or unbiased) and the associated variance(s). Following the success of the Stochastic Gradient Variational Bayes (SGVB) algorithm to optimize the ELBO several other directed graphical models have been proposed. In the presence of multi-modal observed data some conditional and joint generative probabilistic models that are relevant to our approach are - conditional variational auto-encoder Sohn et al. [2015], semi-supervised generative models Kingma et al. [2014], joint generative models Suzuki et al. [2016] and visually grounded generation Vedantam et al. [2017]. Our proposed approach builds directly on top of the underlying principles in joint and conditional generative models. The underlying principles with a relevant subset of work has been covered in Sec 4.

3 Dataset and Evaluation metric

In this section, we describe the datasets used for our experiments and the associated evaluation metric on the downstream task of semantic segmentation.

Figure 2: Example images from the 20 semantic classes in PASCAL VOC'2012

3.1 Dataset

One of the standard datasets on which semantic segmentation results are usually reported is the Pascal Visual Object Categories (VOC) '2012 Everingham et al. [2015], containing 20 object classes. The dataset was collected and annotated by the computer vision community with the primary goal to perform scene understanding in realistic scenes involving several visual object classes. Problems are treated as fundamental supervised learning paradigms involving several unimodal and multimodal inference tasks such as image classification, captioning, object detection and segmentation. Examples of object classes are person, bird, cat, cow, aeroplane, chairs, etc. Depending on the task involved the datasets provide associated coarse as well as fine-grained annotations.

	Approach	mIoU
Full	DeepLab Chen et al. [2016]	**79.7**
	FCN-8s Long et al. [2015]	62.2
	Context_CNN_CRF Lin et al. [2016]	77.8
Semi	GAIN Li et al. [2018]	**56.8**
	SEC Kolesnikov and Lampert [2016]	51.7
	AE-PSL Wei et al. [2017]	55.7

Table 1: mIoU on PASCAL VOC'2012 of state of the art fully supervised and semi-supervised semantic segmentation approaches.

The classes can be broadly divided into four top-level categories: person, animals, vehicles, and indoor objects, each of which in turn contains subcategories to make up the 20 semantic classes. Fig. 2 illustrates the 20 semantic categories. For our experiments, we use the train and validation splits of VOC'2012, containing 1464 and 1449 images, respectively. Table 1 summarizes results of some of the approaches described in 2, on PASCAL VOC'2012. We include results of both state of the art fully-supervised and semi-supervised approaches.

3.2 Evaluation Metric

Performance is reported as mean Intersection over Union (mIoU) between the ground truth annotated pixels and the segmented pixels associated with individual classes averaged across instances and classes. Fig.3 illustrates how the metric is computed. It is evident that IoU is a count based measure, whereas the output predictions of segmentation models are usually distributions over classes per-pixel. While evaluating the mIoU measure for a predicted segmentation output the reported metric is class-normalized mean so as to accurately capture the performance of the segmentation model instead of sub-par performance being reported as something better due to pixel predictions being dominated by a subset of classes present as a majority in the image.

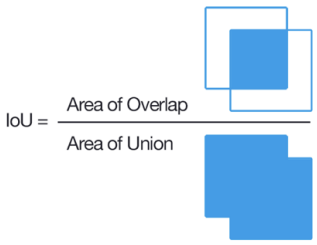

Figure 3: Illustration of the mIoU calculation

4 Preliminaries

In this section, we cover the underlying fundamentals of some approaches which are relevant in the context of our work. Sec 4.1, describes the dense conditional random field used to refine segmentation predictions in our proposed approaches. Sec 4.2 describes the visualization technique - GradCAM Selvaraju et al. [2016] - used to obtain noisy and coarse annotations in a weakly supervised setting. Finally, Sec 4.3 covers some ground on the underlying formulation used in Stochastic Variational Inference (SVI) particularly under the scope of VAEs.

4.1 Conditional Random Field

In this section we review the general Conditional Random Field formulation as applied to semantic segmentation, and further highlight the approach outlined in Krähenbühl and Koltun [2011] to perform efficient inference in dense fully-connected CRF's.

A conditional random field is a discriminative undirected graphical model (UGM) that models the conditional probability of a label sequence (hidden) given an observation sequence. This model does not assume independence among the features on the observations. In the case of dense segmentation, fully connected CRF's with pixel-level connectivity offer strong expressivity to perform segmentation.

As briefly described in 2, Krähenbühl and Koltun [2011] introduce an efficient inference algorithm for such dense fully-connected CRF's. With dense pixel-level connectivity, traditional inference is naturally impractical. Consider X as a random field defined over a set of variables $X_1, ..., X_N$. The domain of each variable in a set of labels $L = l_1, l_2, ..., l_k$. Consider also a random field I defined over variables $I_1, ..., I_N$. I ranges over possible input images of size N and X ranges over possible pixel-level image labelings. I_j is the color vector of pixel j and X_j is the label assigned to pixel j. In the fully connected pairwise CRF model, G is the complete graph on X and the corresponding Gibbs energy can be expressed as the set of all unary and pairwise cliques, with

$$E(X) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j),$$

The unary potentials are typically confidences from a learned classifier. The pairwise potentials can be further simplified as a weighted sum of Gaussian kernels in a feature space as

$$\psi_p(x_i, x_j) = \mu(x_i, x_j) \sum_{m=1}^{K} w^{(m)} k^{(m)}(\boldsymbol{f_i}, \boldsymbol{f_j}),$$

Here each kernel $k^{(m)}(\boldsymbol{f_i}, \boldsymbol{f_j})$ is a combination of an appearance kernel and a smoothness kernel, and vectors $(\boldsymbol{f_i}, \boldsymbol{f_j})$ are feature vectors for pixels i and j in a feature space. The appearance kernel tries to enforce that nearby pixels with similar color should belong to the same class. The smoothness kernel removes small isolated regions. Moreover, μ is a label compatibility function. The hyperparameters are learned from data, or via grid search on a validation set.

$$k(\boldsymbol{f_i}, \boldsymbol{f_j}) = w^{(1)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\alpha^2} - \frac{|I_i - I_j|^2}{2\theta_\beta^2}) + w^{(2)} \exp(-\frac{|p_i - p_j|^2}{2\theta_\gamma^2}).$$

In the above formulation, using a mean field-approximation to the conditional random field distribution helps alleviate the problem of intractable inference, as it affords the possibility of performing message passing updates while minimizing the KL-divergence between the target and the proposal distributions. While this still results in quadratic complexity, this message passing update can be interpreted as a convolution of Gaussian kernels in the pixel feature space.

$$\tilde{Q}_i^{(m)}(l) = \sum_{j \in \nu} k^{(m)}(\boldsymbol{f_i}, \boldsymbol{f_j}) Q_j(l) - Q_i(l) = [G_{\Lambda^{(m)}} \times Q(l)](\boldsymbol{f_i}) - Q_i(l),$$

Here $\tilde{Q}_i^{(m)}$ represents the approximation to the marginal $Q_i^{(m)}$, and $G_{\Lambda^{(m)}}$ represents a Gaussian kernel in feature space. Drawing implications from the sampling theorem, this essentially implies

efficient function reconstruction is possible in practice from this low-pass band-limiting filter by ignoring samples beyond 2 standard deviations, resulting in $\mathcal{O}(N)$ complexity in the number of variables. The kernel parameters are then approximated by approximating the gradients for each training image and performing L-BFGS optimization in this low-data regime.

4.2 Grad-CAM

Deep convolutional networks have been the cornerstone of visual understanding tasks in the recent years with performances surpassing those of simpler models by a significant margin by building a hierarchical representation of the data. However, unlike low-capacity and sample efficient model classes such as Support Vector Machines (SVMs), deep networks suffer from the problem of interpretability - the ability to characterize the reasoning behind success and failure cases from the perspective of generalization. As such one major effort in the deep learning community has been to develop explanation techniques or interpretable models which help us understand *why the model made a particular decision when provided with an input*. Some notable works in this sub-domain are along the lines of - explainable AI Samek et al. [2017], Grad-CAM Selvaraju et al. [2016], understanding the utility of explanations from downstream tasks Chandrasekaran et al. [2017] and so on. Often, implicit (or explicit) attention based works in this domain provide us a way to obtain visual explanations for decisions made by a classifier. We particularly focus on Grad-CAM Selvaraju et al. [2016], which is provably efficient and is a low cost mechanism to obtain per-class decisions which we subsequently use as weak localization cues for semantic segmentation. Essentially, Selvaraju et al. [2016] present a technique for producing "visual explanations" for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. Using the gradients of any target concept flowing into the final convolutional layer, a coarse localization map highlighting the important regions in the image for predicting a particular concept is produced as a visual explanation. Further, Grad-CAM can be combined with existing fine-grained visualizations to create a high-resolution class-discriminative visualization which can be readily applied to a range of CNN-model families, including tasks such as image classification, image captioning, and visual question answering. Such visualizations lend insights into failure modes of these models while maintaining fidelity to the underlying model. Fig. 4 demonstrates one such map for a deep classifier trained on the ImageNet Deng et al. [2009] dataset.

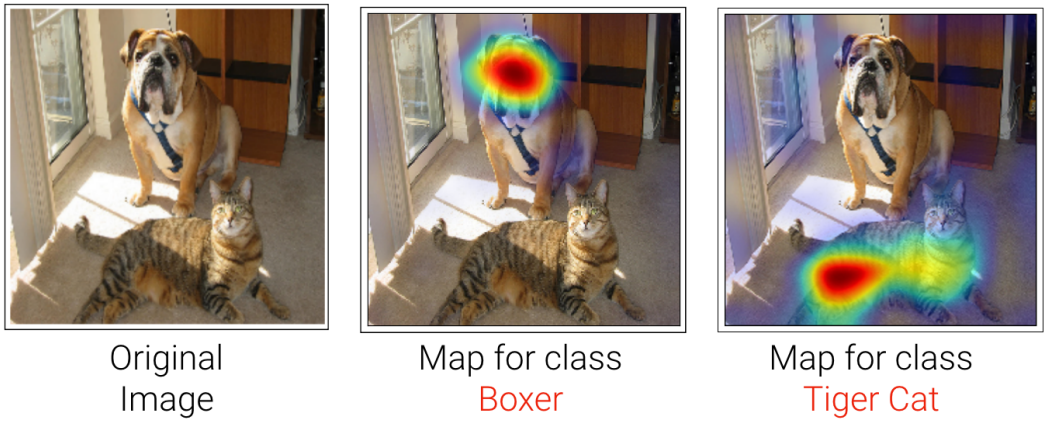

Figure 4: Example Grad-CAM visualizations for existent categories, Boxer and Tiger-Cat

The underlying idea is demonstrated as follows. In order to obtain a class-discriminative localization map, Grad-CAM $L^c_{\text{Grad-CAM}} \in \mathbb{R}^{u \times v}$ of width u and height v for any class c (from a pre-trained classification network), we first compute the gradient of the score for class c, y^c (before a softmax and this unnormalized scores), with respect to feature maps activations A^k of a convolutional layer in the deep network.These gradients are subsequently global-average-pooled to obtain the an empirical

point estimates of the neuron importance weights α_k^c associated with the class c:

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}} \tag{1}$$

This weight α_k^c represents a *partial linearization* - a first order Taylor approximation - of the network downstream from A^k, and captures the 'importance' of feature map at channel k for a target class c. A weighted combination of forward activation maps by the neuron importance weights provides us a loose approximation of the expected saliency map which is then refined by a ReLU operation ($f(x) = \max(0, x)$) to obtain,

$$L^c_{\text{Grad-CAM}} = \texttt{ReLU}\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}} \propto \texttt{ReLU}(\mathbb{E}_{p_\theta(\alpha)}[A^k]) \tag{2}$$

Notice that this results in a coarse heatmap of the same size as the convolutional feature maps.

4.3 Probabilistic Generative Models

In this subsection, we will cover some ground in terms of the formulation associated with a variational autoencoder and discuss some existing extensions in the context of the same.

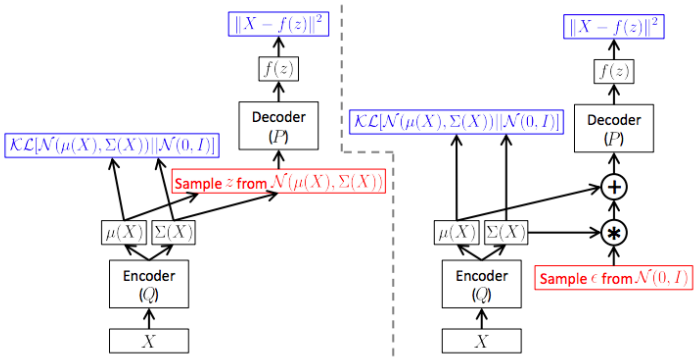

Figure 5: The general architectural pipeline used to implement a vanilla variational auto-encoder. Notice the reparameterization trick on the right half.

Variational Auto-encoder. Consider some dataset $\mathbf{X} = \{\mathbf{x}^i\}_1^N$ of i.i.d. samples of some continuous or discrete variable $\mathbf{x}$. The primary assumption involved here is that the data is generated by some random process involving an unobserved continuous random variable $\mathbf{z}$. Data is generated in an i.i.d. fashion from this (hypothetical) process (see Fig. 6) in hierarchical manner by first sampling the latent variable ($\mathbf{z} \sim p(\mathbf{z})$) and then sampling the datapoint ($\mathbf{x} \sim p_\beta(\mathbf{x}|\mathbf{z})$) from the conditional distribution. Since the *true* posterior under this generative process is intractable due to the presence of the marginal likelihood term ($p(\mathbf{x}) = \int p(\mathbf{z})p_\beta(\mathbf{x}|\mathbf{z})d\mathbf{z}$) efficient MLE or MAP estimate of the parameters of the generative process is a problem. Stochastic Variational Inference frames this problem under the purview of optimization. Essentially, we approximate the true posterior by a variational family $q_\theta(\mathbf{z}|\mathbf{x})$ and optimize a lower bound (Evidence Lower Bound) on the marginal likelihood of the data. This is summarized as follows:

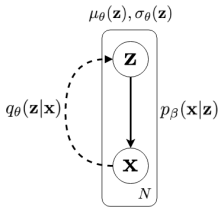

Figure 6: The generative process of a VAE depicted as a plate diagram. The solid arrows represent the generative processes while the dotted ones represent the inference networks.

$$\log p(\mathbf{x}^i) = \texttt{KL}[q_\theta(\mathbf{z}|\mathbf{x}^i)||p(\mathbf{z}|\mathbf{x}^i)] + \texttt{ELBO} \tag{3}$$

$$\texttt{ELBO} = \mathbb{E}_{q_\theta(\mathbf{z}|\mathbf{x}^i)}[\log p_\beta(\mathbf{x}^i|\mathbf{z})] - \texttt{KL}[q_\theta(\mathbf{z}|\mathbf{x}^i)||p(\mathbf{z})] \tag{4}$$

Note that such a modeling assumption allows us to tractably infer the posterior over the latent variable given the data and allows us to model the generative process as well. Optimizing the ELBO over an expectation over the observed data is tricky. The usual Monte Carlo estimator for this type of problem using the log derivative trick (see REINFORCE Sutton et al. [2000]) yields a high variance gradient estimate resulting in an unstable and tedious learning process and is not useful for practical purposes. Kingma and Welling [2013] proposed an alternate gradient estimator based on the reparameterization trick. Essentially, the process of sampling the latent variable $\mathbf{z}$ can be interpreted as follows when the prior over the same is assumed to be a zero-centered unit-gaussian ($p(\mathbf{z}) = \mathcal{N}(0, I)$).

$$\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{x}^i) \tag{5}$$

is equivalent to

$$\epsilon \sim \mathcal{N}(0, I) \text{ and } \mathbf{z} = \mu_\phi(\mathbf{z}) + \epsilon\sigma_\phi(\mathbf{z}) \tag{6}$$

Separating the noise component from the deterministic prediction of the sufficient statistics of the approximate posterior allows us to propagate gradients backward through the whole stochastic process. The process described here is only for a gaussian but similar reparameterization tricks can also be applied for distributions which satisfy certain conjugacy conditions (including the recently proposed Gumbel-Softmax trick for reparameterizing a categorical distribution on the simplex Jang et al. [2016]). The usual architectural pipeline for training such a model is demonstrated in Fig. 5.

Extensions to VAE. Since Kingma and Welling [2013] proposed the Stochastic Gradient Variational Bayes estimator, there has been a significant amount of work in the domain of probabilistic models in terms of modeling more complex generative processes and observed data regimes due to the inherent smooth differentiability aspect of the reparameterization trick. Some notable work in this domain that are applicable in the context of our proposed approach are:

- **Conditional Variational Autoencoder.** Utilizing the Evidence Lower Bound formulation Sohn et al. [2015] proposed a conditional generative process where ELBO is derived as a lower bound on the conditional likelihood of one modality given the other, i.e., $\log p(\mathbf{y}|\mathbf{x})$ allowing us to learn distributions over structured outputs based on given observations. This naturally inspires a semantic segmentation baseline that we study in Sec. 6.
- **Joint Multimodal Variational Autoencoder.** While the cVAE allows us to perform model conditional generative and inference processes; JMVAE Suzuki et al. [2016] takes it one step further by assuming to causal relationship between the modalities present. Having a structure where observations for multiple data modalities are modeled as being generated from a single latent parent random variable traditionally requires one to perform joint-inference - estimating the posterior over the latent variable conditioned on all the data modalities. JMVAE does exactly this except it adds regularization terms that learn to model unimodal inference under missing data modalities. We study a simplified version of the JMVAE model (see Sec. 6 in order to limit model capacity and still optimize the ELBO efficiently.
- **Joint Models with retro-fitted Unimodal Inference Networks.** Vedantam et al. [2017] recently proposed a joint model where the unimodal inference networks are retro-fitted in the regular ELBO associated with a joint model giving rise to the TELBO objective in a principled manner. Such a formulation allows the authors to even perform inference under missing modes in a data-modality via a product of (gaussian) experts formulation. Our proposed model for scene understanding is an extension to the TELBO objective without any arbitrary regularizing terms as in JMVAE.

5 Weakly-Supervised segmentation

Motivation: As we have seen before, the task of semantic segmentation involves labeling each pixel in the image with its associated label (category name). Training data for this task usually consists of per pixel annotations which is extremely hard and expensive to obtain. Also training models with full supervision take tremendous amount of time. In this project, we propose to overcome this requirement by operating on surrogate localizations obtained from a trained image classification model, without

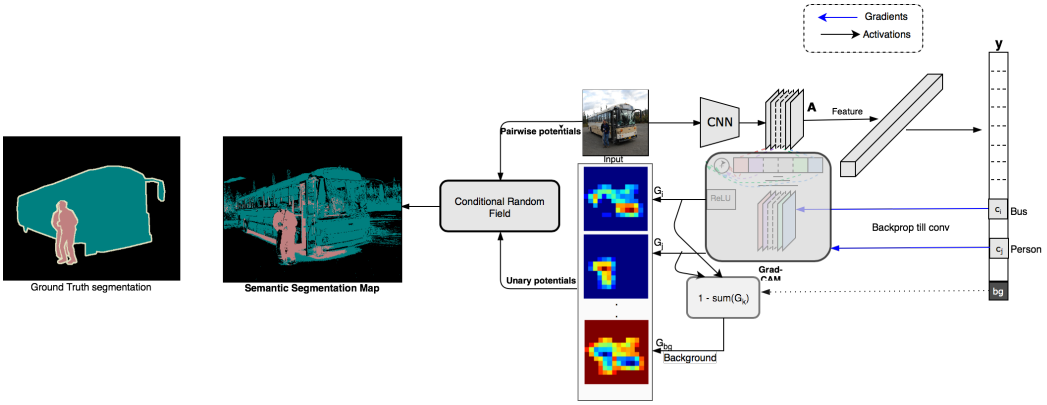

Figure 7: Our pipeline for weakly supervised image segmentation.

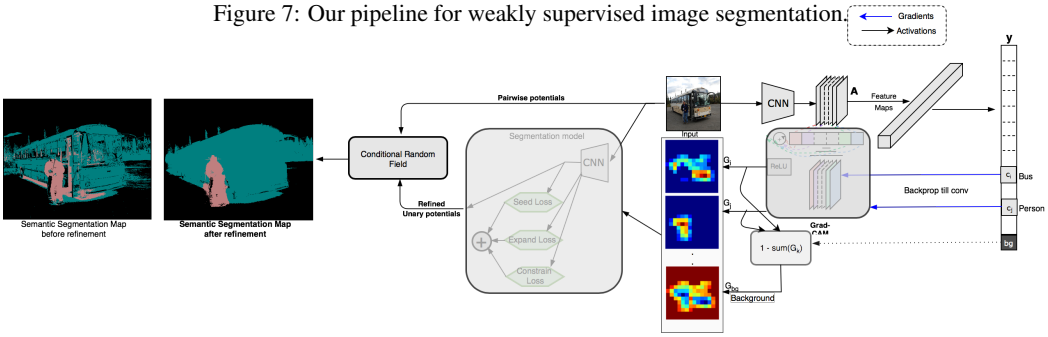

Figure 8: Our pipeline for weakly supervised image segmentation by refining unary potentials of CRF using SEC

the need for expensive pixel level annotation. By enforcing simple Markovian constraints on noisy and low resolution pixel-labels, we aim to obtain comparable performance on this task.

Our proposed approach for this task involves utilizing interpretable visualizations obtained from deep models as weak supervision for training semantic segmentation models. Our proposed approach involves three primary steps. Firstly, we train a deep classification model on the dataset concerned via cheap image-level label annotations provided in the dataset. Secondly, we obtain class-level saliency maps through Grad-CAM which provides low resolution noisy surrogate maps. We then add a CRF on top to obtain better results. Thirdly, we refine the unary potentials obtained through Grad-CAM using the approach in SEC Kolesnikov and Lampert [2016], by training a segmentation model. Following this, we replace the unary potentials in step 2, with refined maps, and thus obtain significant qualitative and quantitative improvements. We perform several ablation studies and show results of our experiments on the standard PASCAL VOC 2012 segmentation dataset.

5.1 Training a classification model with image level labels

PASCAL VOC 2012 dataset has images which have multiple categories present. We finetune a simple image classification model, VGG-16, pretrained on ImageNet. VGG-16 which has 5 convolutional layers and 2 fully connected layers. We trained this architecture with a sigmoid cross entropy loss. The output of this network is a probability distribution for each of the 20 classes in the dataset. We pick a threshold of 0.5 to estimate which classes are present in the image. In the next subsection we see how we use this classification model to obtain weak surrogate maps which can be used as a seed for obtaining segmentation masks.

5.2 Obtaining class-saliency maps

We compute the Grad-CAM map for each of the classes obtained from step (1). PASCAL requires us to get segmentation masks for the background category as well. In order to get an associated saliency map for the background class, we sum the maps for the present categories, and subtract it from 1.

Approach	PASCAL VOC mIoU
Grad-CAM	26.9
Grad-CAM + CRF	31.2
Grad-CAM + SEC Kolesnikov and Lampert [2016]	44.8
Grad-CAM + SEC Kolesnikov and Lampert [2016] + CRF	**50.6**

Table 2: mIoU on PASCAL VOC'2012 of our proposed approaches which use weak supervision.

Approach	mIoU
Grad-CAM	26.9
Grad-CAM + CRF (only smoothness kernel)	26.9
Grad-CAM + CRF (only appearance kernel)	30.7
Grad-CAM + CRF (smoothness and appearance kernel)	**31.2**

Table 3: Results with ablations of CRF.

As can be seen in the figure, the Grad-CAM maps obtained are of low resolution (14×14) compared to original image which is typically of size (300×500). Also the obtained maps are highly noisy. These maps still contain important information which can help localize the class in the image.

5.3 CRF for segmentation

We use the CRF from Krahenbuhl Krähenbühl and Koltun [2011] with unary potentials from step (2) and pairwise potential from image. The downstream CRF is responsible for refining the segmentation predictions based on an underlying smoothness and appearance variation kernel that helps us prune out extensive modifications along visual structures such as edges, corners and other such contextually relevant visual constraints. This results in segmentations like the one shown in Fig 10 and Fig 11. Quantitative results can be found in Table 2.

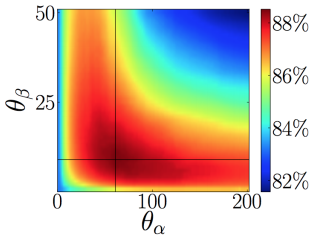

Figure 9: Effect of varying number of long range connections in the appearance and smoothness kernel in the CRF. We find a sweet-spot which gives us the best results. Red values indicate higher mIoU score and blue indicates lower mIoU scores. Plot from Krähenbühl and Koltun [2011]

CRF Ablations: We performed the following ablations of the CRF from Krähenbühl and Koltun [2011].

1. **Varying the parameters of long range connections in the appearance kernel.** As described in the preliminary section, the appearance kernel used by Krähenbühl and Koltun [2011] is defined by, $\exp(-\frac{|p_i-p_j|^2}{2\theta_\alpha^2} - \frac{|I_i-I_j|^2}{2\theta_\beta^2})$ where the first term enforces that nearby pixels should have similar assignments, and the second term enforces that pixels with similar intensities should have similar label assignments. θ_α and θ_β denote the weight of long-range connections in the CRF for pixel location and intensities respectively. In this ablation we vary the weights for the long range connections for the spatial and the intensity terms. A higher value of θ gives uniform weighting for nearby and far-away pixels/intensities, and low values of θ gives higher weights to closer pixels/intensities. We perform a grid search over θ_α and θ_β and observe similar performance to the plot shown in Fig. 9.

2. **Effect of appearance kernel and smoothness kernel.** Here, we study the importance of the appearance and smoothness kernel. We performed 3 variants - only using appearance kernel, only using smoothness kernel and using both kernels. We show results in table 3. We find that the appearance kernel is more important than smoothness kernel and including both kernel gives the best segmentation performance.

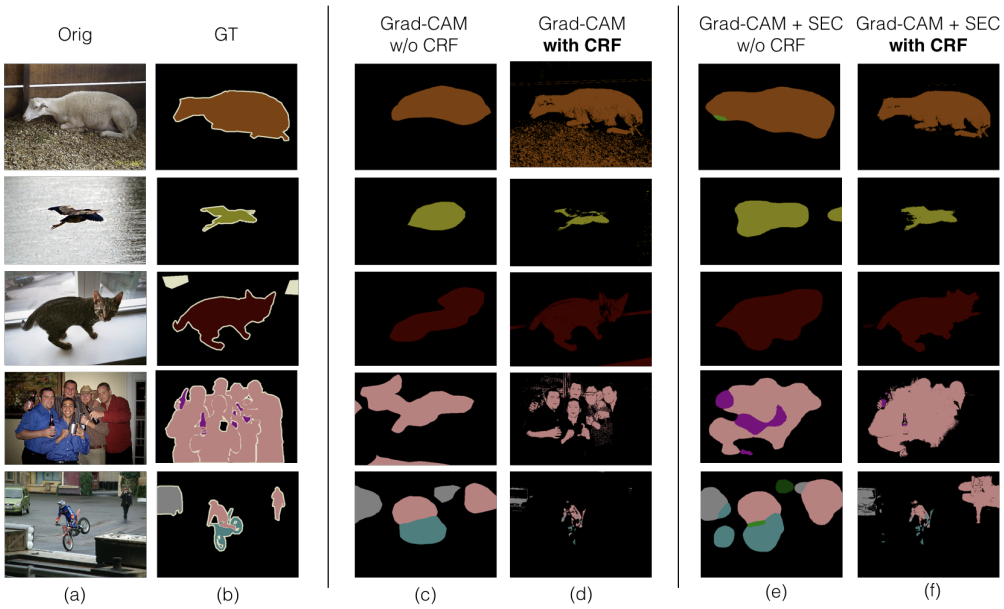

Figure 10: Weakly supervised image segmentation results

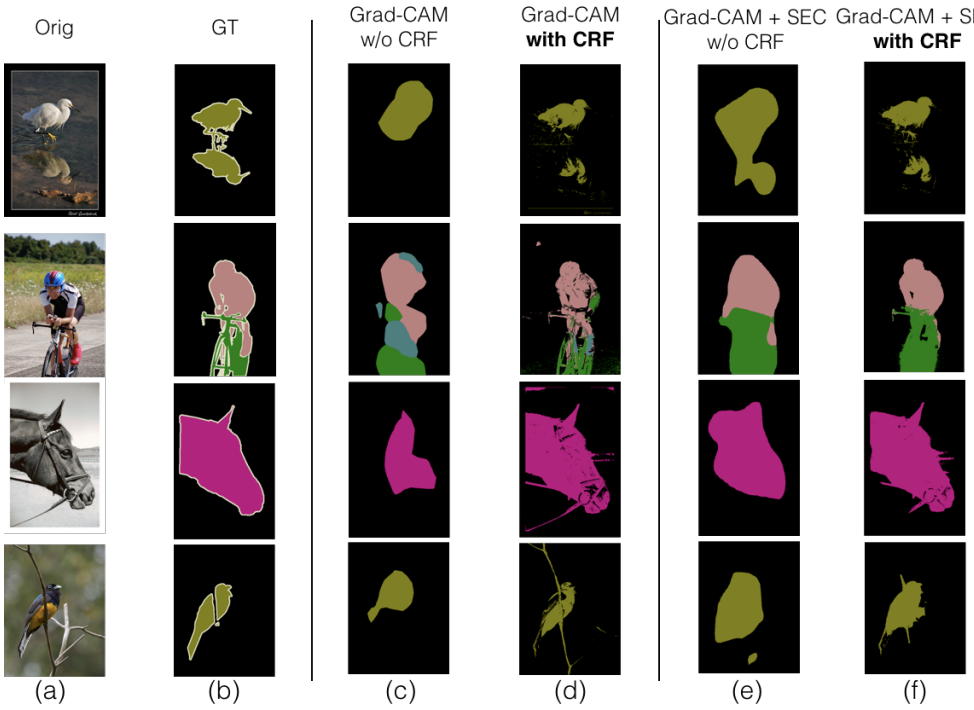

Figure 11: Weakly supervised image segmentation results

5.4 Refining unary potentials using SEC

Seed Expand and Constrain (SEC) In recent work, Kolesnikov and Lampert [2016] introduce a new loss function for training weakly-supervised image segmentation models. Their loss function is based on three principles – 1) to seed with weak localization cues, encouraging the segmentation network to match these cues, 2) to expand object seeds to regions of reasonable size based on information about which classes can occur in an image, 3) to constrain segmentations to object boundaries that alleviates the problem of imprecise boundaries already at training time. They showed that their proposed loss function, consisting of these three losses, leads to better segmentation.

As can be seen in Fig 7 (left), the labels for foreground classes propagate to background classes, and in some cases to other foreground classes. This can be attributed to the artifacts introduced when resizing the low resolution Grad-CAM saliency maps to original image size. We use the segmentation model from Kolesnikov and Lampert [2016] to refine the unary potentials (Grad-CAM saliency maps). We then feed the refined unary potentials and the original image to the CRF, to obtain final segmentation maps. This results in better qualitative and quantitative segmentations as shown in Fig 10 and Fig 11.

6 Towards Scene Understanding via Generative Models

As motivated in the introduction, in this section we try to take the task of semantic segmentation one step further by using probabilistic models to model the general task of scene-understanding. Defined loosely, scene understanding is a computer vision task involving perceiving, analyzing and elaborating an interpretation of a real-time dynamic scene. As mentioned earlier, we restrict our modeling domains - data modalities - to that of physical scene understanding which involves characterizing the kind and position (space) occupied by different objects in an image. The data modalities that we experiment with are - the image i, the classes present (or absent) c, and the segmentation masks (predicted or otherwise) i_S (see Fig. 12). To avoid notational clutter, in all our formulations of ELBO, we omit the expectation over the dataset ($\mathbb{E}_{x\sim p_{data}}[.]$). In our approaches and baselines, the image is encoded by a small convolutional neural network; the latent variables via two layered multi-layer perceptrons and the images and segmentation outputs are decoded via small de-convolutional neural networks.

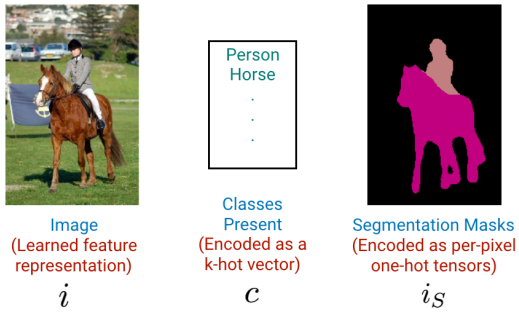

Figure 12: Different data-modalities used in our scene understanding pipeline.

Conditional Variational Auto-encoder (cVAE). An obvious but naive natural extension of the VAE that can used to model structured output representations like segmentation masks given data is the conditional variational autoencoder (cVAE). We treat this as a baseline approach to judge the complexity of modeling stochastic generative processes for semantic segmentation. We reason only over the modalities i and i_S. As a representation of i, we use features extracted from a pretrained deep convolutional network (VGG-16 Simonyan and Zisserman [2014]). The specific generative process that we are interested in modeling in this case is shown in Fig. 13. The dotted lines represent the joint inference network $q_\theta(z|i, i_S)$ and the solid lines represent the joint generator network $p_\beta(i_S|i, z)$, where z is the unobserved latent variable over which we have an appropriate Gaussian prior $p(z)$. The generative process that has been assumed in this modeling choice can be summarized as:

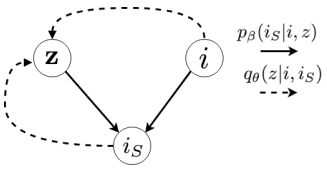

Figure 13: The generative process associated with a conditional VAE depicted via a plate diagram.

$$z \sim p(z) \text{ and } i_S \sim p_\beta(i_S|i, z) \tag{7}$$

Under such a process the ELBO can be written as:

$$\texttt{ELBO} = \mathbb{E}_{q_\theta(z|i,i_S)}[\log p_\beta(i_S|i,z)] - \texttt{KL}[q_\theta(z|i,i_S)||p(z)] \tag{8}$$

The ELBO is optimized via the SGVB as discussed earlier. We use Adam Kingma and Ba [2014] as our optimizer with hyper-parameters chosen via grid-search. The first term in the objective aims to maximize the conditional likelihood of the generated segmentations under the chosen variational family while the second term acts as a regularizer and forces the approximate posterior family to stay close to the chosen prior.

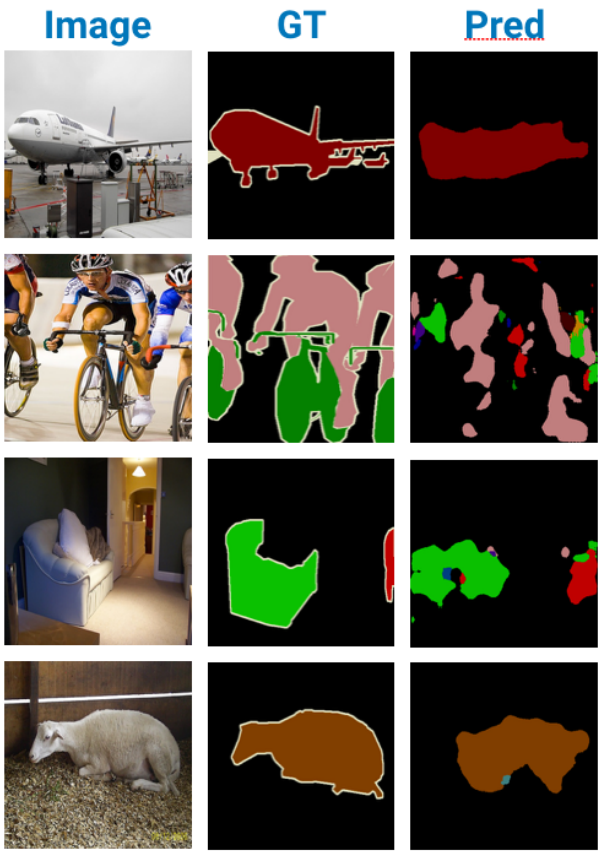

Figure 14: Qualitative results on segmentation using cVAE.

In terms of results, the cVAE architecture achieves a mIoU of 37.75. If we carefully observe the qualitative results presented in Fig. 14, we notice that in addition to poorly modeling the conditional distribution, the cVAE also spits out spurious class predictions. We realized that this is indicative of a pathological problem (see Fig. 15) that the cVAE is likely to suffer from. While optimizing the ELBO, if the dynamics associated over gradient updates occur in a manner such that we are bad at optimizing the KL[.] term in the objective, the modeled posterior will be unable to cover the prior family properly. This results in the variational family occupying only a certain region in the space of prior family and hence spurious samples of the latent variable z are drawn at test time giving us poor performance and segmentation predictions.

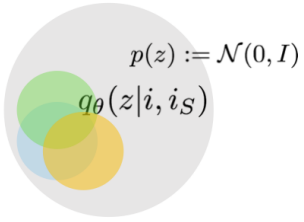

Figure 15: Poor optimization of the KL[.] resulting in small coverage of the prior by the variational family.

Joint Generative Models. In order to move towards some notion of completeness in scene-understanding, we need to be able to perform inference across modalities as well - we should be able to infer one modality conditioned on the others. While a cVAE is an obvious baseline to

attempt this, it only supports this inference in one direction. This boils down to having a joint model that in addition to allowing us to perform efficient joint inference over the latent variable, should also allow us to perform inference conditioned on a subset of modalities. While existing approaches to joint models like JMVAE allow us to do that, the inference networks characterizing inference over a subset of modalities are usually added as regularizers to the ELBO. Thus, there is a need to reason about unimodal (for instance) inference while performing joint inference itself.

To counter this, we propose the generative process in Fig. 16 (inspired from Vedantam et al. [2017]) to model a scene. Arrows colored blue represent the joint inference network while ones colored yellow represent the inference networks when conditioned on a subset of modalities. Assuming no causal relation, we have one universal parent continuous unobserved random variable z that generates the data modalities. Under such a process, while it is trivial to derive the ELBO corresponding to the joint inference network ($q_\theta(z|i, i_S, c)$), to ensure we learn appropriate parameters for the inference networks characterized as $q_\nu(z|i_S)$ and $q_\rho(z|i, c)$ we need to retrofit these networks under the existing generative process. Overall this leads to our objective to be a combination of three ELBOs as shown below:

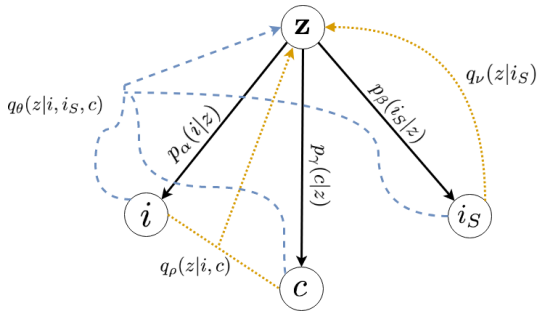

Figure 16: Generative process associated with our proposed probabilistic model characterized by the plate diagram.

$$\texttt{ELBO}_{joint} = \mathbb{E}_{q_\theta}[\log p_\alpha(i|z)p_\beta(i_S|z)p_\gamma(c|z)] - \texttt{KL}[q_\theta(z|i, i_S, c)||p(z)] \tag{9}$$

$$\texttt{ELBO}_1 = \mathbb{E}_{q_\rho}[\log p_\alpha(i|z)p_\gamma(c|z)] - \texttt{KL}[q_\rho(z|i, c)||p(z)] \tag{10}$$

$$\texttt{ELBO}_2 = \mathbb{E}_{q_\nu}[\log p_\beta(i_S|z)] - \texttt{KL}[q_\nu(z|i_S)||p(z)] \tag{11}$$

Note that since the generator parameters are shared across the three ELBOs, the subset-inference networks are implicitly conditioned to operate within the generative processes. This in itself is a very complicated objective to optimize for but if done efficiently allows us to generate segmentation outputs conditioned on the inputs, construct scenes conditioned on the segmentation masks and also to generate datasets in data-starved regime by performing a random walk in the space of z.

In addition to modeling a complicated joint model, we also compare this with a low-capacity model (*j-VAE) without any networks allowing us to perform inference over a subset of modalities and where we ignore one modality while performing joint inference. Upon optimizing, loose *j-VAE achieves an mIoU of 62.25 while our proposed model achieves an mIoU of only 47.23. The qualitative results depicted in Fig. 17 make for some interesting observations emerging from the sub-par performance of our retrofitted model relative to *j-VAE. Clearly, unimodal inference comes at a cost. One obvious conclusion here is that our proposed objective is indeed harder to fit compared to a loose version of the same under amortized inference settings. The variance in gradients associated with the generator parameters are significantly magnified due to those terms appearing thrice in the overall objective under different variational families. We hypothesize that a combination of multi-sample ELBOs for the retrofitting objectives along with the regular characteristic one for the joint inference network might be a candidate solution. Another hidden culprit here is the optimizer. It is widely known that the choice of optimizer under the class of Stochastic Gradient Descent approaches can lead to vastly different (flat) minimas. Given the complexity of the objective, it is reasonable to assume that saddle points existing in one of the ELBO terms could play a significant role in deviating optimization from its ideal course.

7 Logistics

In this section we outline the logistics of our implementation, in particular, the open source libraries and frameworks that we used, and descriptions of specific aspects of our implementation to enable reproducibility. As described, we proposed two different approaches for the semantic segmentation task, and we describe implementation details of each separately. Our source code for each subsection is provided along with this report.

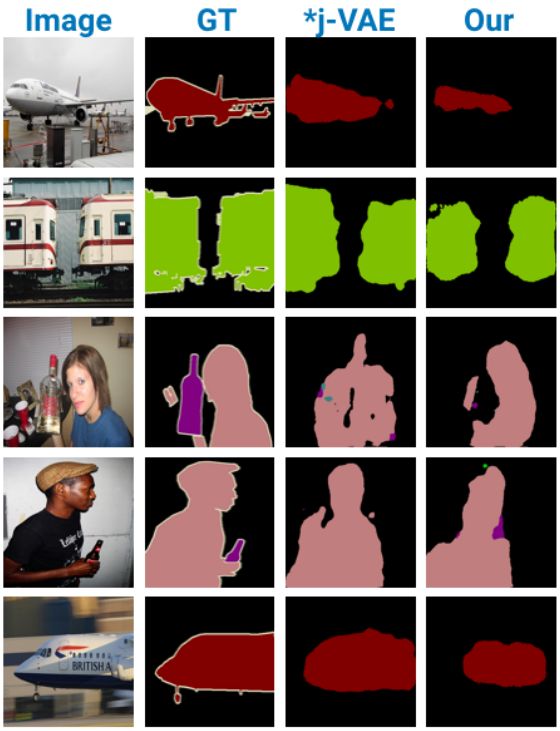

Figure 17: Qualitative results of our proposed approach in comparison with the loose approximation involved in variational family of *j-VAE.

7.1 Weakly Supervised Semantic Segmentation

To implement the weakly supervised semantic segmentation pipeline, we used both various open source libraries as well as implemented certain components from scratch. As described previously, we finetuned a VGG-16 model pretrained on ImageNet which was implemented in the Caffe(Jia et al. [2014]) framework. The produced probability distribution for each of the 20 different classes was thresholded at 0.5 (as determined by cross validation), and Grad-CAM was run on this network for each of the 20 classes using our own Caffe implementation, which was coded from scratch.

For the dense CRF we employed to refine the GradCAM maps (5.3), we used the publicly available code by Krähenbühl and Koltun [2011] – in particular, we used the Python wrapper provided over the original C++ code. To run the CRF ablations, we modified some of the original wrapper APIs for our convenience. Finally, for refining the unary potentials obtained by Grad-CAM (5.4), we used the publicly available code by the authors `https://github.com/kolesman/SEC`, which also internally uses the Caffe(Jia et al. [2014]) framework.

7.2 Scene Understanding via Generative Models

To implement the probabilistic models, we used the open-source auto-differentiator framework called PyTorch Paszke et al. [2017] which has source code written in the programming languages Python and C++. The model was coded from scratch including `ELBO`, optimization schemes, hyper-parameter search, etc. We wrote multi-threaded dataloaders which support loading datapoints over multiple cores in a machine in parallel. In addition, we also parallelized our model architecture over multi-GPUs to support faster training.

Overall code structure was written in terms of model classes, dataloader classes and training or evaluation execution scripts. While the dataloader classes maintain a pointer over the current instances being loaded over a batch, the model class allows us to maintain an object characterizing the architecture of the model. Computing loss, generating samples, etc. were written as methods in the associated model class. The evidence lower bound was implemented based on the closed form solution associated under gaussian parameterizations of the encoding and decoding processes (see appendix of Kingma and Welling [2013] for detailed expressions).

8 Conclusion

To conclude, in this project we explored weak supervision for semantic segmentation - eliminating the need for pixel-level supervision by using weak localization cues obtained from visual explanation modalities Selvaraju et al. [2016] associated with deep models which were subsequently refined via efficient inference over densely connected conditional random fields. Taking the task a step further we also studied scene understanding via probabilistic generative models, identifying certain pathological consistencies as well as inconsistencies with the adopted and proposed approaches. In terms of future work, we plan on integrating both of the avenues explored in this project to model a scene in a joint fashion and perform inference in a semi-supervised setting. In addition, it might also be interesting to explore the usage of natural language as data-modality allowing us construct scenes from natural language descriptions and performing inference in the opposite hop as well.

References

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

Joao Carreira, Rui Caseiro, Jorge Batista, and Cristian Sminchisescu. Semantic segmentation with second-order pooling. In *European Conference on Computer Vision*, pages 430–443. Springer, 2012.

Arjun Chandrasekaran, Deshraj Yadav, Prithvijit Chattopadhyay, Viraj Prabhu, and Devi Parikh. It takes two to tango: Towards theory of ai's mind. *CoRR*, abs/1704.00717, 2017. URL `http://arxiv.org/abs/1704.00717`.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016. URL `http://arxiv.org/abs/1606.00915`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Xuming He, Richard S Zemel, and Miguel Á Carreira-Perpiñán. Multiscale conditional random fields for image labeling. In *Computer vision and pattern recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE computer society conference on*, volume 2, pages II–II. IEEE, 2004.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589, 2014.

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European Conference on Computer Vision*, pages 695–711. Springer, 2016.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. *arXiv preprint arXiv:1802.10171*, 2018.

Guosheng Lin, Chunhua Shen, Anton Van Den Hengel, and Ian Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3194–3203, 2016.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

N. Payet and S. Todorovic. (rf)2 – random forest random field. In *NIPS*, 2010.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

Ramprasaath R. Selvaraju, Abhishek Das, Ramakrishna Vedantam, Michael Cogswell, Devi Parikh, and Dhruv Batra. Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *CoRR*, abs/1610.02391, 2016. URL `http://arxiv.org/abs/1610.02391`.

Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, pages 1–8. IEEE, 2008.

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304. Ieee, 2011.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, pages 3483–3491, 2015.

Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Ramakrishna Vedantam, Ian Fischer, Jonathan Huang, and Kevin Murphy. Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*, 2017.

Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *IEEE CVPR*, 2017.