Automatic Assessment of Open Street Maps Database Quality using Aerial Imagery

Boris Repasky
STELarLab
Lockheed Martin Australia
Australian Institute for Machine Learning
University of Adelaide

Dr Timothy Payne
STELarLab
Lockheed Martin Australia

Dr Anthony Dick
Australian Institute for Machine Learning
University of Adelaide

***Abstract*—Open data initiatives such as OpenStreetMap (OSM) are a powerful crowd sourced approach to data collection. However due to their crowd-sourced nature the quality of the database heavily depends on the enthusiasm and determination of the public. We propose a novel method based on variational autoencoder generative adversarial networks (VAE-GAN) together with an information theoretic measure of database quality based on the expected discrimination information between the original image and labels generated from OSM data. Experiments on overhead aerial imagery and segmentation masks generated from OSM data show that our proposed discrimination information measure is a promising measure to regional database quality in OSM.**

I. Introduction

OpenStreetMaps (OSM) is an open data initiative of geographic data which is entirely comprised of community contributors [1]. While a fantastic project, such sources of information depend on the enthusiasm and determination of the public. In this paper we propose a novel method for automatically assessing database quality, and apply it to the OSM database. Our method is well motivated by the probabilistic graphical model formalism of variational autoencoders (VAEs). However, to avoid their reliance on element-wise losses, we go beyond VAEs, and incorporate the generative adversarial network's (GAN) discriminator into the reconstruction objective.

II. Related work

To our knowledge, no work has previously been done to assess the quality of segmentation masks in an unsupervised fashion, particularly in the context of aerial imagery or remote sensing. However, a number of related works are described in this section.

A. Weakly supervised and Unsupervised Feature Extraction.

A lack of labeled data has been an ongoing problem in the overhead and remote sensing domain [2]–[4]. Due to this many authors have attempted to develop weakly supervised or unsupervised techniques to extract useful features [2]. Mou *et al.* [5] proposed a fully Conv-Deconv network with residual learning for unsupervised spectral-spatial feature learning of hyperspectral images. Their network used a encoder-decoder paradigm to learn features from reconstructing the input 3D hyperspectral patch [2]. Romero *et al.* [3] introduced the use of greedy layer-wise unsupervised pre-training coupled with their Enforcing Lifetime and Population Sparsity algorithm to efficiently learn sparse features.

In work most comparable to ours, Singh *et al.* [4] used techniques from image semantic in-painting to extract features useful for semantic segmentation in overhead imagery. The authors use an adversarial training scheme to adversarially corrupt the network input with a binary in-painting mask, forcing the decoder to fill in the missing information [4].

B. Image-to-Image Translation

Image-to-image translation is the task of transferring one representation of an image to another [6]. Examples of image-to-image translation include colouring in monochrome images, denoising images and generating real images from drawings [6]–[8]. A number of authors have used GANs, VAE-GANs or conditional GANs specifically for the task of generating realistic images from segmentation masks [6], [9], [10]. Of particular note, Isola *et al.* used a U-Net based conditional GAN to generate overhead images from rasters of google maps.

III. Dataset Collection

We used the Arcgis API Powered by Esri to fetch aerial imagery over major metropolitan areas of both Australia and New Zealand. Then for each geographical region we used the OpenStreetMap API to fetch the raw geodata from the OSM database in XML format. We wrote a custom renderer using the Python Imaging Library (PIL) to convert the OSM XML into binary segmentation masks using the map features available in OSM.

For these experiments we restricted ourselves to the primary features *'building'* and *'highway'*. The *'building'* tag includes all structures from small dwellings to large stadiums, while the *'highway'* tag includes all roads, highways, lanes and access ways. Figure 1 shows an example of the aerial imagery together with the rendered binary segmentation masks.

IV. Method

To automatically assess the quality of the OSM database we take advantage of a key property of VAEs, namely that they are formalised in terms of probabilistic graphical models. Thus, using similar reasoning to VAEs we construct the problem as

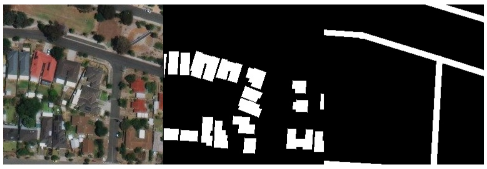

Fig. 1. An example of the rendered segmentation masks together with the aerial imagery served by Esri through ArcGIS

follows, let $X = \{x^{(i)}\}_{i=1}^{N}$ be a collection of N i.i.d image samples and $L = \{l^{(i)}\}_{i=1}^{N}$ be a collection of N i.i.d label samples. We assume both images and labels are generated by some random process involving a shared latent random variable z. Furthermore, we assume that the prior $p_{\theta^*}(z)$ and the likelihoods $p_{\theta^*}(x \mid z)$, $p_{\theta^*}(l \mid z)$ come from differentiable parametric families. Then similar to VAEs, despite the fact that the true parameters θ^* and latent vector z are hidden from us, we can use variational Bayesian methods to approximate the intractable posteriors $p_{\theta^*}(z \mid x)$ and $p_{\theta^*}(z \mid l)$.

Let us assume for the moment that we have successfully obtained the posteriors $p_{\theta^*}(z \mid x)$ and $p_{\theta^*}(z \mid l)$, let us further assume that $p_{\theta^*}(z \mid x)$ in some sense represents the "true" distribution of the latent vectors z. Then we can measure the information lost when $p_{\theta^*}(z \mid l)$ is used to approximate $p_{\theta^*}(z \mid x)$ by calculating the Kullback–Leibler (KL) divergence

$$D_{KL}\left(p_{\theta^*}(z \mid x) \mid p_{\theta^*}(z \mid l)\right).$$

In fact this is precisely the expected discrimination information for x over l. Another way of interpreting this is that if the label l contained the same information content as the image x the KL-divergence would vanish.

Thus if we have access to the true posteriors $p_{\theta^*}(z \mid x)$ and $p_{\theta^*}(z \mid l)$ the discrimination information would serve as an ideal measure of the quality of our labels, as it measures how well the labels explain the original image in the latent space. Of course in reality we do not have the true posteriors, rather by using VAEs cleverly we can obtain the approximate distributions $q_{\psi}(z \mid x)$ and $q_{\phi}(z \mid l)$.

However we need two conditions to hold, that the parameters ψ are as close as possible to the parameters ϕ and that the latent variables z capture the latent information in the images well. To ensure these properties approximately hold, we used a combination of network architecture design, together with a novel training scheme.

A. Architecture

Our network architectures consist of two almost identically structured VAE-GANs, one which takes an image as an input, and one that takes a stacked collection of segmentation masks, one class for each channel. We use VGG-19 as the probabilistic encoder backbone, encoding to an output vector of $512 \times 2 \times 2$. We assume the prior over the latent variables is a centered isotropic multivariate Gaussian, and that the distributions $q_{\psi}(z \mid x)$ and $q_{\phi}(z \mid l)$ are multivariate Gaussians, therefore each sampled z is of dimension $256 \times 2 \times 2$ with half the VGG-19 output encoding the means and the other the variance. The decoder was a 9-layer network with 8 deconvolution layers, with leaky-relu as the activation functions. Finally, the discriminator was another VGG-19. The only difference between the two VAE-GANs is the number of channels they take in the initial input, with the image VAE-GAN having a three channel input, and the segmentation mask VAE-GAN having the same number of input channels as the number of classes (in this case two).

B. Training scheme

We first train the VAE-GAN on image inputs, initialising the weights of the VGG-19 with pretrained weights on ImageNet. Once training of the VAE-GAN on image inputs is complete we transfer all weights to the new VAE-GAN for training on segmentation masks as inputs. We then replace and reinitialise the first layer of the network in order to accommodate the change in the number of channels. Finally, we freeze the weights of both the decoder and the discriminator and retrain the encoder on the segmentation mask inputs.

Crucially, although the inputs have changed to the segmentation masks, the loss calculation is unchanged from that used when the images were inputs. That is to say the network is being trained to reconstruct the original images, and not the segmentation masks. Freezing the weights ensures that the retrained encoder must produce samples in a similar region of the latent space to the VAE-GAN trained on images in order to successfully reconstruct that image.

V. Experiments and Results

To test our method, we created a benchmark test set consisting of several hand chosen regions of Australia and New Zealand. Each region in the benchmark test set was carefully selected to ensure that all highways and roads were present. Then we removed some rendered elements and used the element-wise difference between the unaltered segmentation mask as a proxy for the database quality. Preliminary results show a correlation between our discrimination information measure and our proxy for database quality. However, further work is needed to investigate the robustness and nature of this correlation.

VI. Conclusion

We have introduced a novel method to automatically assess the quality of the OSM database. Our method involves the training of VAE-GANs in a novel training scheme. We develop a theoretical framework that allows us to compare the information content of an original image with labels of the content of that image. In particular, we approximately measure the expected discrimination information for the image x over the labels l, per sample of the latent vector z.

Acknowledgment

Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org

References

[1] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org , 2017.

[2] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, Dec. 2017, conference Name: IEEE Geoscience and Remote Sensing Magazine.

[3] A. Romero, C. Gatta, and G. Camps-Valls, “Unsupervised Deep Feature Extraction for Remote Sensing Image Classification,” *IEEE Trans. Geosci. Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, Mar. 2016, arXiv: 1511.08131. [Online]. Available: http://arxiv.org/abs/1511.08131

[4] S. Singh, A. Batra, G. Pang, L. Torresani, S. Basu, M. Paluri, and C. V. Jawahar, “Self-Supervised Feature Learning for Semantic Segmentation of Overhead Imagery,” p. 13.

[5] L. Mou, P. Ghamisi, and X. X. Zhu, “Fully conv-deconv network for unsupervised spectral-spatial feature extraction of hyperspectral imagery via residual learning,” in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Fort Worth, TX: IEEE, Jul. 2017, pp. 5181–5184. [Online]. Available: http://ieeexplore.ieee.org/document/8128169/

[6] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” *arXiv:1611.07004 [cs]*, Nov. 2018, arXiv: 1611.07004. [Online]. Available: http://arxiv.org/abs/1611.07004

[7] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, “Sketch2Photo: internet image montage,” *ACM Trans. Graph.*, vol. 28, no. 5, pp. 1–10, Dec. 2009. [Online]. Available: https://dl.acm.org/doi/10.1145/1618452.1618470

[8] B. Coll and J.-M. Morel, “A non-local algorithm for image denoising,” vol. 2, 07 2005, pp. 60– 65 vol. 2.

[9] L. Karacan, Z. Akata, A. Erdem, and E. Erdem, “Learning to Generate Images of Outdoor Scenes from Attributes and Semantic Layouts,” *arXiv:1612.00215 [cs]*, Dec. 2016, arXiv: 1612.00215. [Online]. Available: http://arxiv.org/abs/1612.00215

[10] S. E. Reed, Z. Akata, S. Mohan, S. Tenka, B. Schiele, and H. Lee, “Learning What and Where to Draw,” p. 9, 2016.