Latent Representation of Topographic Classes in Remote Sensing Image Data using Autoencoders

Hannes Stärk

June 13, 2019

Contents

1 Vorwort

2 Kurzfassung

3 Abstract

4 Introduction

4.1 Topic Overview

4.2 Related Work

4.3 Contributions

4.4 Outline

5 Setup

5.1 Data

The available data are 1024x1024[1] images of the two United States cities Jacksonville in Florida and Omaha in Nebraska taken from the US3D Dataset that was partially published to provide research data for the problem of 3D reconstruction (Bosch et al. 2019). The images for each recorded area cover one square kilometer and can be divided into four categories with the first one being optical satellite images with three channels (RGB). Secondly visible and near infrared satellite images with eight channels (VNIR). Thirdly digital surface models (DSM).[2] And lastly semantic labeling with five different categories.

The optical images were taken by the WorldView-3 satellite of Digital Globe from 2014 to 2016 and contain[3] seasonal and daily differences in vegetation and sun positions.[4]. [5] Each pixel of an image is described by three bytes representing the intensity[6] of either red, green or blue.
Also collected by WorldView-3 were the VNIR images[7] which contain eight channels for eight different bands of the[8] spectrum with a ground sample distance of 1.3 meters[9]. These images were taken over the course of all twelve months

[1] Sehr spezifische Angabe fÃ¼r einen Einleitungssatz
[2] Der Satz ist zu kurz und kann mit dem vorherigen vielleicht verbunden werden.
[3] Ist dieser Begriff passend?
[4] Die Logik des Satzes ist m.E. nach nicht konsistent?
[5] Hierbei wurden die Bilder zu verschiedenen Jahreszeiten- zeiten aufgenommen, die dazu fuehren, dass die Erscheinungen der Szene sich in einem hohen Masse unterscheiden. Zum Beispiel fuehren unterschidedliche Sonnenstaende zu veraenderten Schatten und Reflexionen. plus wetter
[6] die reflektierte Intensitaet des zugehoerigen Punktes auf der Erdoberflaeche in den Wellenlaengen
[7] VNIR am Anfang? Und uber msi statt vnir reden
[8] the? Warum ein bestimmter Artikel?
[9] Verhaelt es sich anders zu den RGB-Bildern?

making them usable for training models that can handle seasonal appearance differences which are even more distinct than in the RGB data because certain wavelengths capture shadows and vegetation especially well. Overall this data offers more detail than the three channel RBG pictures. The eight channels of the imagery correspond to the following wavelengths:[10]

- Coastal: 400 - 450 nm
- Blue: 450 - 510 nm
- Green: 510 - 580 nm
- Yellow: 585 - 625 nm
1. Red: 630 - 690 nm
2. Red Edge: 705 - 745 nm
3. Near-IR1: 770 - 895 nm
4. Near-IR2: 860 - 1040 nm

The given DSMs were collected using light detection and ranging technology (Lidar). They have a single channel that describes the height of each pixel with a greater number representing a higher distance to the ground. [11]
Lastly there are semantic labeled pictures with one channel of a single byte encodes one of five different topographic classes. Those classes are vegetation, water, ground, building and clutter. The semantic labeling was done automatically from lidar data but manually checked and corrected afterwards.[12]
For all four categories of data the area covered in a single image is one square kilometer[13] and they contain a lot of oblique view of buildings, often with sunshine casting good shadows making the data ideal for training models that should detect them.[14]

5.2 Experiments

With knowledge about the provided Data we can start thinking about possible experiments to gain insight into the

6 Architecture

6.1 Loss Function

Entropy

A part of the loss function used in the variational autoencoder is based on Kullback-Leibler divergence. To understand Kullback-Leibler divergence it seems

[10] Ist es sinnvoller zuerst die Wellenlaengen einzufuehren und hiernach auf die verschiedenen Bilddaten. Darueber hinaus die Inhalte von der Aufnahme zu trennen?
[11] Gefallen dir die Saetze zu DSMs? Ich formuliere hier einen anderen Satz.
[12] Welchem Zweck dienen diese Daten? Bzw. warum wurden diese Daten gelabelt? Das ist eine nicht unwichtige Information. Du erwaehntest oben schon die 3D-Reconstruction.
[13] Wiederholung?
[14] Logik: Abweichungen von der zentralperspektive, die dazu fuehren dass Objekte mit starken Hoehenaenderungen sichtbare Schatten werfen... Ich komme nicht mit der Logik des Satzes klar.

necessary to explain entropy from the field of information theory. In short, entropy is a measure for the minimum average size an encoding for a piece of information can possibly have.
Suppose there is a system $S1$ that can have four different states a, b, c, d and every one of those states is equally likely to occur, that means the probability $P(x)$ of each state x is $1/4$. Now the goal is to losslessly transmit all information about that system with the minimum average amount of bits. That can be done with onyl two bits for example like this

$$a : 00 \quad b : 01 \quad c : 10 \quad d : 11$$

However, if $P(a) = 1$ and $P(b) = P(c) = P(d) = 0$, zero bits will suffice to encode the information since it is always certain that the system is in state a. So the entropy of the system clearly depends on the probablities of each state. To see in which way, one can consider the system $S2$ with $P(a) = 1/2,\ P(b) = 1/4,\ P(c) = 1/8$ and $P(d) = 1/8$. In that case it would be best to encode the state with the highest probability with as few bits as possible since it has to be transmitted the most often. That means a is encoded with one bit as 0. When decoding the information there must be no ambiguities so while the encoding for b has to start with a 1 it cannot be 1 since we need to encode two more states so $b : 10$. Additionally if $c : 11$ there would be no space left for d: say $d : 111$ then if the transmitted information is $111111\ldots$ it could either be decoded to $ccc\ldots$ or $dd\ldots$. So c should rather be encoded as 110 which way $d : 111$ works. In the end a valid encoding that can transmit all information with the minimum average amount of bits is

$$a : 0 \quad b : 10 \quad c : 110 \quad d : 111$$

Here the states c, d are encoded with three bits instead of the two bits in the first example. But c and d are transmitted far less often than a which now only needs one bit. To be more precise half of all transmissions have one bit. Additionally a quarter of all transmissions have two bits. The sum of those probabilities multiplied with the respective amount of bits is the average amount of bits needed to transfer the information in a given encoding. So in the example, with $f(x)$ as the number of bits that encode a state x, that turns out to be $P(a)f(a) + P(b)f(b) + P(c)f(c) + P(d)f(d) = 1.75$. That means for $S2$ on average you only need 1.75 bits to encode a state and since that is also the minimum 1.75 is the entropy of $S2$.
In general in an optimal encoding $f(x)$ is the same as $\log_2 \frac{1}{P(x)}$. For $S1$ $P(x)$ is $1/4$ so the number of bits for x is $\log_2(4) = 2$ what matches the two bits the first encoding uses for the states of $S1$.
The entropy H of a system with a set of discrete events X and the probability distribution $P(x)$ for each $x \in X$ is

$$H(P) = \sum_{x \in X} P(x) \log_2 \frac{1}{P(x)} = -\sum_{x \in X} P(x) \log_2 P(x) \tag{1}$$

This is often written as the expectation for a given state x under the distribution P.

$$H(P) = E_{x\sim P}[-log_2 P(x)]$$

Intuitively if a system has high entropy, the size of the encodings are high on average and many states have small probabilities. This means it is hard to predict what state the system will be in at a given time since there is no state that can be guessed with high confidence. If entropy is low, zero for example, one can be confident that the system is in a certain state like in the previous example with $P(a) = 1$.

Cross Entropy

If the real distribution P of a system is unknown an estimate distribution Q could be guessed and encoding sizes $-log_2 Q(x)$ can be produced which will not be optimal for the true distribution P. Now with some data gathered and P known the used encoding sizes can be cross-checked with the expectation under the actual distribution resulting in the cross entropy $H(P, Q)$

$$H(P, Q) = E_{x\sim P}[-log_2 Q(x)] = -\sum_{x\in X} P(x) \log_2(Q(x)) \tag{2}$$

In machine learning tasks regarding classification this is often used as a loss function since the label of a piece of data gives us a distribution P with absolute cetrainty and $H(P) = 0$. With the inaccurate distribution Q that the model estimates $H(P, Q)$ will be greater than zero unless $P = Q$ where $H(P, Q) = H(P, P) = H(P) = 0$. So the lerning algorithm can try to minimize $H(P, Q)$.

Kullback-Leibler Divergence

Having computed the entropy $H(P)$ and cross-entropy $H(P, Q)$ of two distributions P, Q it is possible to compare those distributions by comparing $H(P)$ and $H(P, Q)$ through subtraction

$$\begin{aligned} D_{KL}(P \parallel Q) &= H(P, Q) - H(P) \\ &= E_{x\sim P}[-log_2 Q(x)] - E_{x\sim P}[-log_2 P(x)] \\ &= E_{x\sim P}[-log_2 Q(x) + log_2 P(x)] \\ &= E_{x\sim P}[log_2 \frac{P(x)}{Q(x)}] \\ &= \sum_{x\in X} P(x) log_2 \frac{P(x)}{Q(x)} \end{aligned} \tag{3}$$

where $D_{KL}(P \parallel Q)$ is called the Kullback-Leibler divergence of P and Q. This works because $D_{KL}(P \parallel Q)$ is zero if Q and P are the same since that means $H(P, Q) = H(P)$. On the opposite if Q is different from P then $H(P, Q)$ is

greater than $H(P)$ and therefore the KL divergence is greater than zero proportional to how differten Q and P are. In summary Kullback-Leibler divergence is a measure of how different two probability distributions are that is zero if they are the same and greater than zero if not.

The Loss Function

Intuitively the loss function defines a goal that the model should reach when training by minimizing the loss function. In the variational autoencoder one of those goals is that the distribution of the latent space is similar to a normal distribution since that makes it possible to sample latent variables from a normal distribution. From those sampled variables the decoder can generate an output that resembles the real distribution.
To define that goal in a loss function Kullback-Leibler divergence is used to force the distribution of the latent space to resemble a normal distribution.

7 Future Work

8 Conclusion

References

Bosch, Marc et al. (2019). "Semantic Stereo for Incidental Satellite Images". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, pp. 1524–1532.