Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Jiří Mayer
Název práce	Semi-supervised Learning in Optical Music Recognition
Rok odevzdání	2022
Studijní program	Informatika **Studijní obor** Softwarové a datové inženýrství

Autor posudku	Milan Straka **Role** Oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Text posudku:

The goal of the thesis was to improve optical music recognition using semi-supervised learning – by using unannotated data of sheet music in addition to a supervised dataset. To that end, the author propose to jointly train a model on two tasks: (1) the supervised task at hand, (2) proposed reconstruction (denoising) task, which processes sheet music masked with randomly generated squares and tries to predict the original image. The proposed architecture is evaluated in three settings: (a) using the MUSCIMA++ dataset (training set of 10 supervised pages and 50 unsupervised pages), (b) using the MUSCIMA++ and the CVC-MUSCIMA datasets (training set of 99 supervised pages and 551 unsupervised pages), (c) knowledge transfer from printed to handwritten sheet music. In all experiments, the semi-supervised approach does not surpass the supervised one, but shows better training stability. The used hyperparameters are clearly described and explored.

The thesis text is written in English and is of a high quality. In addition to the optical music recognition itself, the existing datasets, the proposed methodology, and the performed experiments, the author also describes semi-supervised learning itself in detail (on 9 pages). Even if the presented information is not directly used in the experiments, it can serve as a good introduction to the topic.

I consider the thesis to be of high quality, given that the author demonstrated independent, high-quality scientific work in a rather unexplored area.

I recommend the thesis to be defended.

Remarks and Suggestions

- I consider the choice of the pixelwise F1 metric inappropriate. The explanation given by the author is

  > Even though these object detection metrics tell more about the actual usefulness of the model (we care about detected objects, not pixels), we chose to evaluate our experiments directly at the pixel level. Adding these evaluation metrics adds unnecessary complexity to our experiments and it is not needed for our goal – measuring the impact of unsupervised data.
  >
  > …
  >
  > It is important to note, that the two articles regarding music symbol detection (Hajič jr. et al. [2018], Dorfer et al. [2017]) use object detection F1 score, which is not directly comparable to our pixelwise F1 score.

  Using a custom metric prevents any comparison to previous work, which I find severe. Furthermore, the author themselves consider the existing metric better, so it seems the only reason for not using it was the work required to implement it.
- I find the lack of batch normalization to be a huge disadvantage of the proposed architecture. Batch normalization has been a de-facto standard regularization technique

for convolutions, and has been used almost routinely with convolutions. The author even mentions on page 36

> It may be the case, that using batch normalization instead or dropout (like Hajič jr. et al. [2018]) has the same effect of regularizing the network. However our goal is not to find the optimal architecture, but to measure the impact of unsupervised data. For that reason, we did not explore this option further.

I find this reasoning faulty because suboptimal architectures can easily behave differently to the high-performing ones.

I also believe the inclusion of batch normalization could alleviate the "dying ReLU" problem described on page 40 – when batch normalization is used, "half" of the convolution outputs would be positive all the time, avoiding the situation when no gradients flows back through a ReLU because the former steps of the training pushed the activations too much below zero.

- I find the reason for using gated skip connections unconvincing. Quoting from page 37:

  > The gated mode seems to perform worse than the solid one. Despite this, we chose to train all the experiments in this gated mode as we believe it helps the model to learn representations during unsupervised training.

  The claim about better representations is not supported by any evidence, on the contrary, the supervised results are better with the solid skip connections. Given that the semi-supervised representations are trained using denoising, I do not think removing the skip connections should necessarily result in better representations (for the masked regions, the skip connections do not contain any interesting data anyway).
- Training both the supervised and semi-supervised task jointly is a potential reason why the performance of the semi-supervised approach does not surpass the supervised one (because the model must perform well in two quite different tasks simultaneously).
  I would also consider the (overwhelmingly dominant) approach of first pre-training the model on the semi-supervised task, followed by finetuning on the supervised task (the approach is mentioned in Section 6.1, but no reason for choosing the joint training is given).
- It might be interesting to evaluate the semi-supervised pre-training on some higher-level task like a complete end-to-end OMR.
- I would consider exploring the model architecture more.
  - I would try more ResNet-style convolutional blocks on every stage.
  - I would consider not reducing the channels in the upsampling convolution; instead, I would concatenate the channels from the residual connection and the upsampling block and let the following convolution to reduce the number of channels.
  - Using Adam with convolutions can yield suboptimal results. I would therefore consider using AdamW or RMSProp (as in EfficientNet).

Remarks to the Text

- The citations that are not part of a sentence are systematically formatted incorrectly.
  Instead of for example (page 3):

  > The largest dataset of handwritten music sheets is CVC-MUSCIMA, containing 1000 pages (Fornés et al. [2011]).

  the author should use

  > The largest dataset of handwritten music sheets is CVC-MUSCIMA, containing 1000 pages [Fornés et al., 2011].

  achieved via the `\citep` command.
- I noticed two systematic errors in commas – first, there should be a comma before "and" when it connects two independent clauses; second, there should not be a comma before "that" when it begins a restrictive relative clause (and it cannot begin a non-restrictive relative clause).

- Page 17, the end of the smoothness assumption:
  > The opposite statement also holds; if the two input points are separated by a low-density region, the outputs must be distant from each other.

  I believe that is incorrect – there is no reason why one class could not form several clusters. Indeed, the *Semi-supervised Learning* book by Chapelle et al. states that
  > If, on the other hand, they are separated by a low-density region, then their outputs need not be close.

  Later, in the cluster assumption, Chapelle et al. state that
  > Note that the cluster assumption does not imply that each class forms a single, compact cluster: it only means that, usually, we do not observe objects of two distinct classes in the same cluster.
- On page 18, the author state that the consistency regularization follows from the cluster assumption – I believe it follows from the smoothness assumption instead (cluster assumption considers several datapoints, while here we consider a neighborhood of a given datapoint, which does not necessarily contain other datapoints).
- I would move the Section 4.3.2 Adversarial Autoencoders after 4.3.3 Generative Adversarial Networks, given that AAE is a combination of both VAE and GAN.
- On page 34/35, I could not understand the term "oversampling" from the definition
  > The article also employs a technique called oversampling, where a tile is sampled up to five times if it contains no pixel of the target segmentation class.

  I had to read the referenced paper to understand it (I thought a tile without the target class pixels will be used [sampled] 5 times during training). Furthermore, given that the author considers only notehead/staffline segmentation, does oversampling has any effect at all?
- The idea of dropping larger parts of the input image to obtain better representations has been used by various previous works. To mention a few:
  - Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and L´eon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. JMLR, 2010.
  - Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In CVPR, 2016.
  - Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR, 2021.
  - Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, Ross Girshick: Masked Autoencoders Are Scalable Vision Learners. In CoRR, 2021.

  The first work proposes to perform image denoising by masking (yet of randomly selected pixels), and the second one masks larger regions of the input image. The third and fourth paper employ masking of random input patches as a pre-training task, with the latter paper achieving extremely good performance when trained solely on the ImageNet-1k dataset.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 31. květen 2022 **Podpis**