Review for "Error-Free Tile-Based Segmentation Inference of Images Larger than GPU Memory"

Summary

The submission presents an exact approach to perform inference in fully-convolutional networks (FCNs) for inputs that are too large for the available GPU memory. The idea is to partition the input into rectangular regions of a certain size (Zones of Responsibility, or ZoR) and to include a border (halo) around each resulting input tile that is large enough to provide the context necessary to infer the output tile exactly. This is achieved by setting the halo size to half the receptive field size of the FCN.

A general approach to computing the halo size is introduced and is verified experimentally with the U-Net and Fully Convolutional DenseNet architectures, showing that an insufficiently large halo produces an inexact output (when compared against exact computation on the entire input) and that the proposed halo computation approach is sufficient to produce an exact output (up to floating point error).

Recommendation

Accept after minor revision.

Justification

The manuscript is clear and well-written, and generally provides adequate visual support for the concepts discussed. The proposed approach is described in sufficient detail to be replicated, and source code is provided via a GitHub repository. While no mathematical proof of the proposed approach's exactness is provided, the approach itself makes intuitive sense and exactness is verified empirically on a public benchmark.

I am very familiar with general CNN architecture concepts, but I am not familiar with the dense prediction literature specifically, so it's more difficult for me to assess the claim that *no published method fully explores a methodology for error-free tile-based (out-of-core) inference of arbitrarily large images*. I did a cursory literature search and could not find any work that contradicts this claim. In any case, the discussion of related work appears thorough, and the problem of exact computation of dense predictions on large inputs is of clear practical interest.

The issues that need to be addressed before publication are detailed below.

Detailed comments

Major concerns

My one concern with the submission is related to the following sentence: *The required halo can be calculated according to the Equation 1 for a general FCN architecture.*

I am concerned that this is too strong a claim. In particular, the 2^L term in Equation 1 appears to assume that pooling operations operate on 2x2 windows with stride 2x2. Would Equation 1 be correct for a FCN where pooling layers operate on 3x3 windows with stride 3x3?

In fact, the referenced Distill article (Araujo et al., 2019) goes into great length to derive a general framework for computing receptive fields in CNNs. I feel that the submission's contribution in that respect duplicates a subset of that effort. Since Araujo et al. (2019) provide software to perform these computations, wouldn't it be easier in general to run that code for a given network architecture?

To be clear, I think the paper's contributions are significant enough without that, but in my opinion Equation 1 would be better framed as a special case of Araujo et al. (2019)'s Equation 2.

Clarity

- *A special type of CNN which only uses convolutional layers is called a "fully convolutional neural network" (FCN) and it allows the alteration of the input image size.*
  - The phrasing *it allows the alteration of the input image size* was confusing to me at first. I think the authors mean that FCNs can be applied to inputs of arbitrary size?
- Sliding window and patch-based approaches could benefit from a brief introduction.
- The submission states that the proposed approach does not handle layers with dynamic inference or variable receptive fields. Would it be more accurate to say that the proposed approach has no memory benefits when the network architecture uses layers that increase the receptive field to the entire input, since exact computation is only possible using the entire input in that case?
- Readability would be improved if the term *halo border* was briefly explained in Section 1's last paragraph. In fact, I feel that page 4's last paragraph (*In other words, the image is broken down into…*) does a better job of explaining the proposed idea at a high level than Section 1's last paragraph.
- *Nonetheless, the presented methodology applies to any FCN…*
  - This seems in contradiction with the statement that the proposed approach does not handle layers with dynamic inference or variable receptive fields.

- *The halo must be half the receptive field.*
    - What happens if the receptive field size is odd-valued? Suggestion: *The halo must be half the receptive field (rounded down).*
- In page 6's last paragraph, batch normalization is presented as potentially problematic for exactness. While this is true when using batch statistics, aren't estimated population statistics generally used during inference?
- *Convolutional type was changed to SAME from VALID as used in the original paper*.
    - This sentence is confusing. Can you reformulate?
- I think displaying Ronneberger et al. (2015)'s Figure 1 with attribution would help clarify the discussion of the U-Net architecture.
- Why does the submission measure exactness with respect to three metrics? Isn't it sufficient to show that RMSE goes to zero to demonstrate exact computation?
- It would be helpful to explain when and why Luo et al. (2016)'s gradient-based approximation of the receptive field is not exact.

Grammar, style, and typos

- The formatting of textual citations adds a duplicate parenthetical citation (e.g. *Ronneberger et al. (Ronnerberger et al., 2015)* in Section 2's second paragraph).
- *Inference* (used as a verb) and *inferencing* are grammatically incorrect. *Perform inference* and *performing inference* should be used instead. Similarly, *operating on images which cannot be inferred in a single forward pass* (page 3) would read better as *operating on images for which inference cannot be performed in a single forward pass*, and *This tile-based inferencing* (page 5) would read better as *This tile-based inference*.
- *… the whole image being is broken down into non-overlapping regions.*
    - Missing word?
- Section 3's first paragraph mixes verb tenses.
- Mnih's name is misspelled as *Mihn* on page 3.
- *… is a combination of that kernels stride…*
    - *kernel's*
- *… when using halo values less 96.*
    - *less than 96*
- *… is shown relative by the runtime…*
    - *relative to*
- [Optional] I feel that the term "exact" would be a better choice than "error-free" to convey the fact that the proposed approach is *not* an approximate computation approach.