Identifying Cancer Cells with CovNets

Chris Pearce - Stanford University - cpearce@stanford.edu

Overview

Detecting cancer cells in slide images is a laborious manual process. Recent advances in deep neural networks have presented the opportunity to automate this analysis, and are an active area of research. The winning entry in the 2016 TUPAC challenge was a CovNet model from Lunit Inc (Paeng et al. 2017)

Challenges

Key issues in developing effective models include

- **Small Datasets**: A typical dataset may comprise fewer than 1000 training images
- **Sparsely Populated Results**: An image of 2000x2000 pixels may contain one single cancer cell of 30x30 pixels
- **Image Variability**: Slides are manually prepared by pathologists, with significant variation in exposure levels
- **Model Saturation**: The binary nature of many of the classification tasks (presence or absence of a cancer cell) can easily cause models to saturate

Research Aims

The aim of this research is to train a classifier to identify the presence of cancer cells in slide images. The primary model seeks to build a lightweight classifier that can detect the presence of cancer cells in individual slide samples. As an extension, a generator model is being trialled that seeks to generate a probability map modelling the likelihood of the presence of cancer cells at all points across the slide

Model

The classifier model is illustrated below. 256x256 samples of the images are passed through a four layer CovNet to a binary classifier determining whether an image contains a cancer sell. To address problems with model saturation,

- The model is trained at a learning rate of 1e-5.
- Input data is subsampled after the first epoch, so the model is presented with all misclassified observations, and only 1/3 of correctly classified observations.
- SVM loss is used

A second generator model is under development. Inspired by the pix2pix model (Isola et al. 2016), the design uses a generator with skip layer connections to try and produce a probability map of mitosis presence

Dataset and Processing

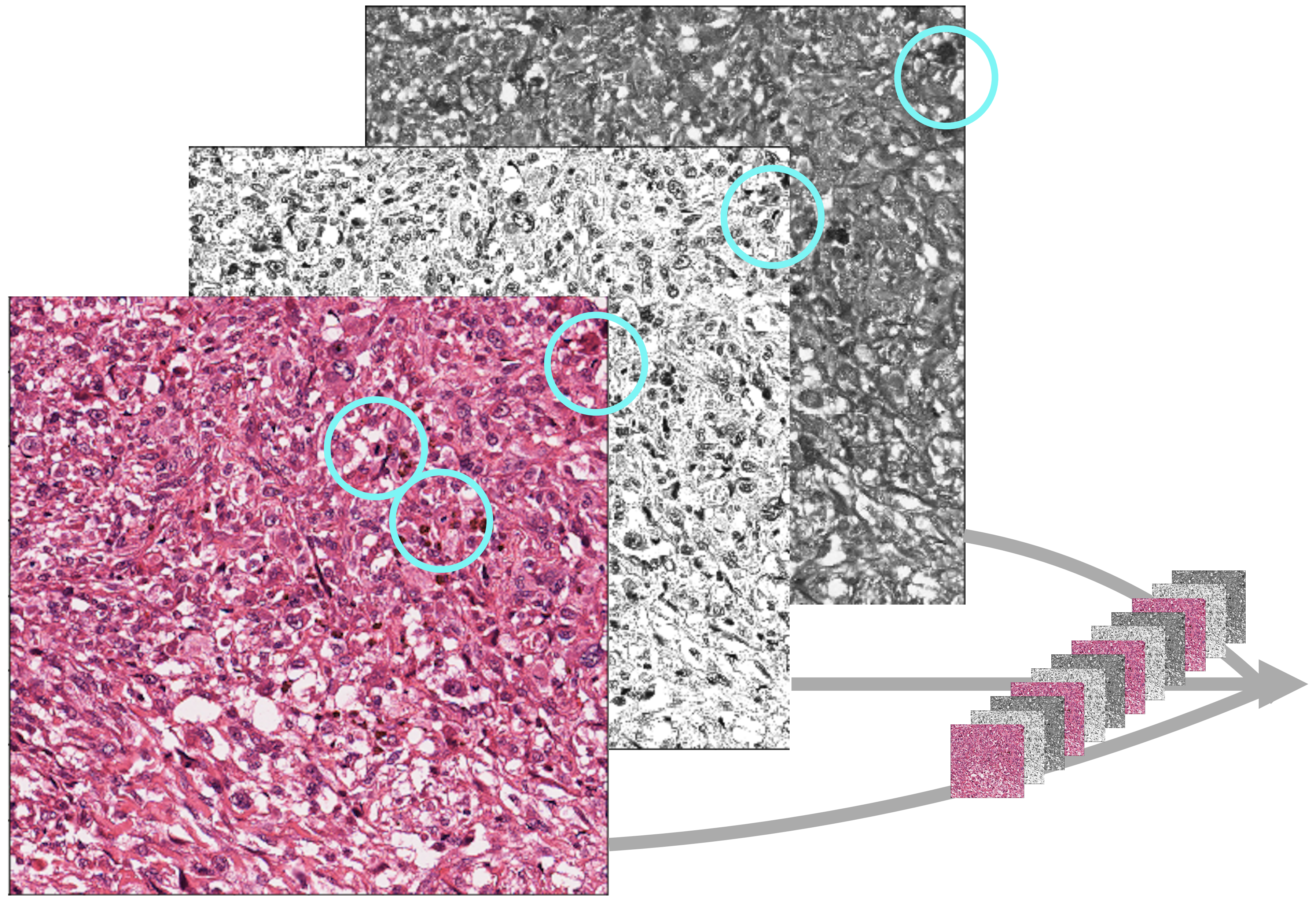
This research uses the Tumour Proliferation Assessment Challenge 2016 mitosis detection dataset, which comprises approximately 650 labelled 2000x2000 pixel slide images. http://tupac.tue-image.nl

Due to the large image sizes being worked with, Macenko Principle Components Analysis was used to separate hematoxylin and eosin views of the slide

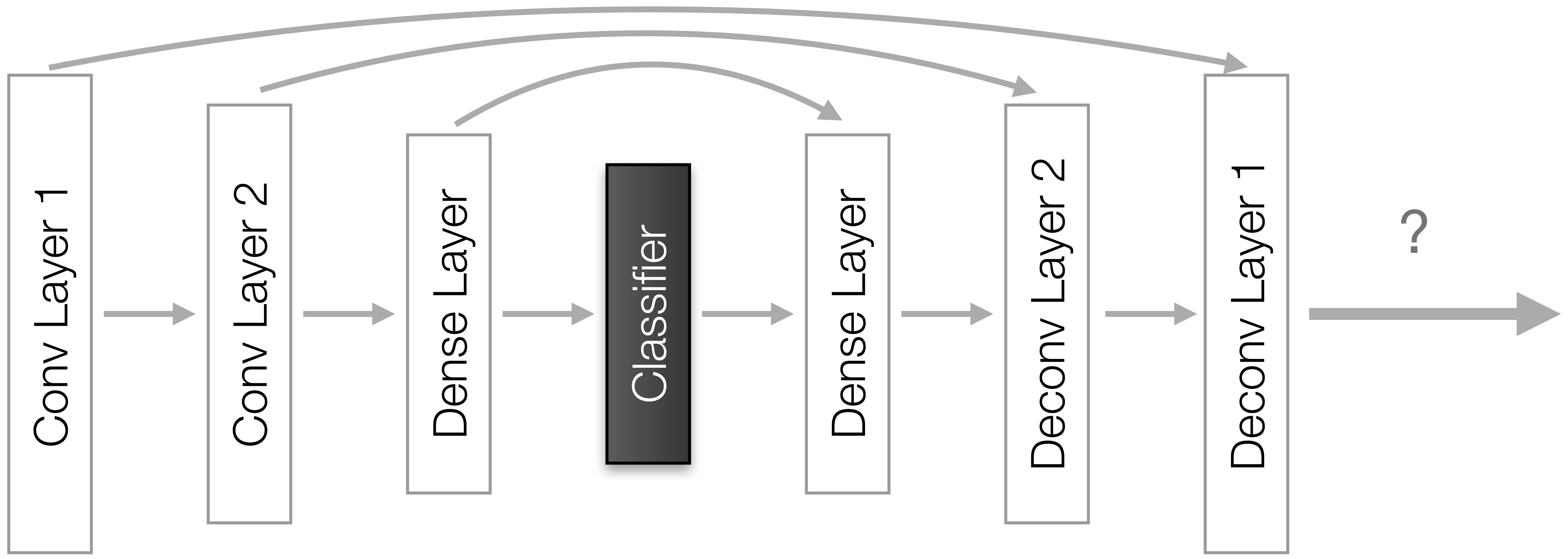
Results

The model achieves 87% binary classification accuracy after 30 epochs of training. Validation accuracy remains high even as test validation starts fitting at >98% accuracy. This model will be extended to calculate an estimate of the FC1 accuracy score

Development on the generator model is still ongoing. The sparse nature of the dataset is a major challenge, as it again causes the model to saturate easily. Current testing is focussing on training the classifier first, the locking the model before training the generator

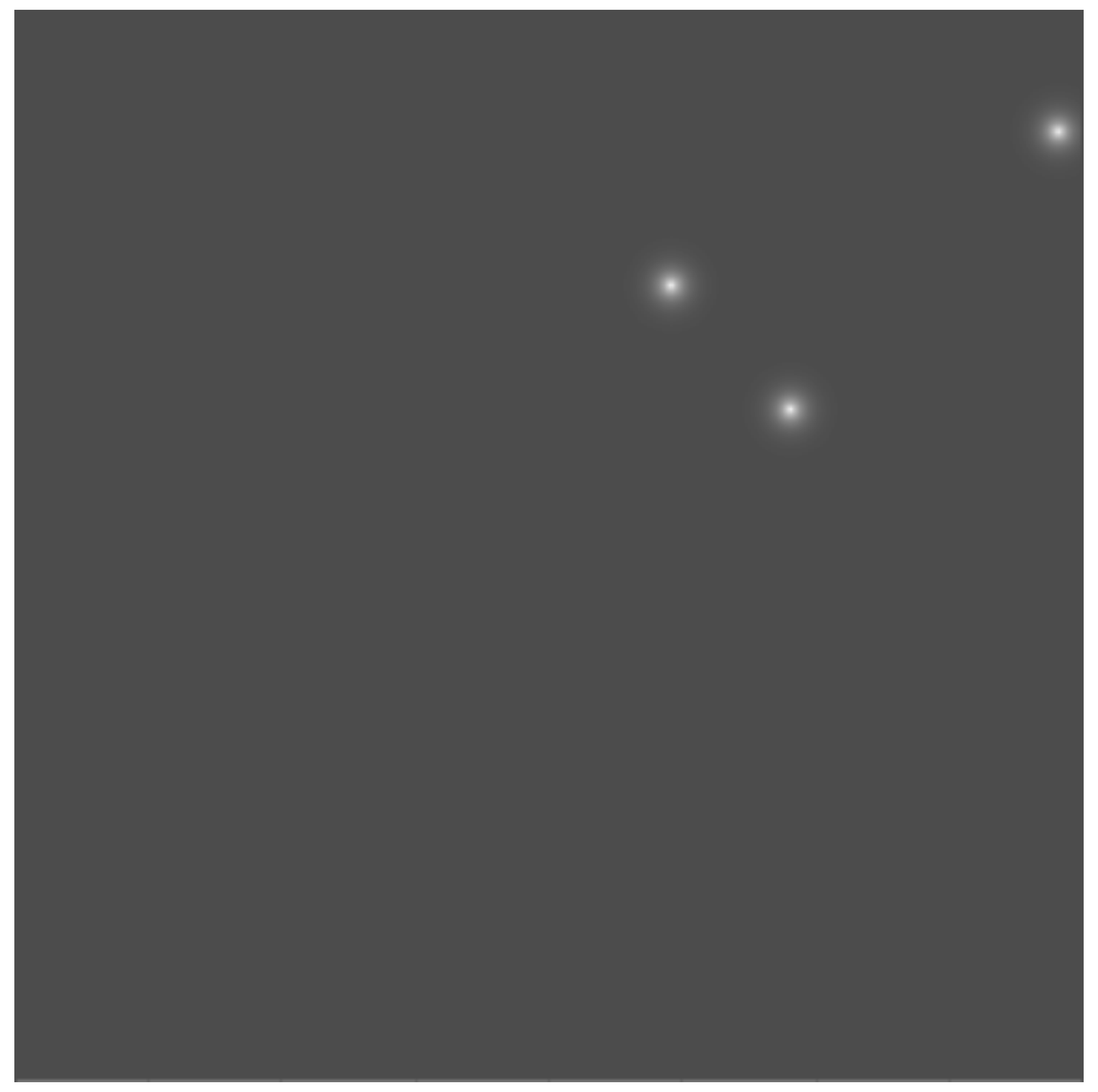


Automatic deconvolution of cancer slide images using Macenko Principle Components Analysis (Macenko et al. 2009)



Classifier Model

Proposed Skip Connection Generator Model



Probability Map