IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Convolutional Neural Network modeling for Land Surface Temperature Super Resolution

Manuel Cabeza Gallucci
Badr Soulaimani
Nicolas Delbono
André Cheker Burihan

Encadrants:
Carlos GRANERO BELINCHON (IMT Atlantique)
Aurélie MICHEL (ONERA-DOTA)
Xavier BRIOTTET (ONERA-DOTA)
Thomas CORPETTI (CNRS)

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Index

Scientific context

Through satellite imaging, it is possible to obtain high temporal resolution of land surface temperature images. However, the spatial resolution which is obtained through this method is not enough for certain applications. Thus, image processing techniques are being implemented in order to get the desired spatial resolution.

We can distinguish two main methods to procure this spatial resolution. In the first place, we have methods which use the reflective domain of the electromagnetic field. These methods use both classical and artificial intelligence methods. On the other hand, there are methods which use the thermal infrared (TIR) domain. Unlike their counterparts on the other class of methods, these ones only use classical approaches.

Nevertheless, these approaches have certain limitations. Mainly, the need of generalizing that, the parameters at a higher resolution are still the same than those at a lower resolution. Also, we need to have images of the visible and near-infrared (VNIR) and short wave infrared (SWIR) domains at a time and space which are not too far apart. [1]

The methods which use the thermal infrared domain are therefore, the more interesting methods for innovation using AI methods. We will present some of the classical methods in this category.

These methods work on the following way. First, they generate a model between LST and VNIR-SWIR features, like the Normalized Difference Vegetation Index (NDVI), at a 1km scale. Then, they grab the VNIR-SWIR image at 250m, which we they have a priori, and apply the previously obtained model to them. Thus, LST images are obtained at this better resolution. Finally, a residual estimation is used to correct the result. [1]

TsHARP is widely used due to its simplicity, as it is the basic kernel-driven method. ATPRK obtains better results but it is a little bit more complex. ATRPK has two parts, regression and area-to-point kriging (ATPK), with kriging being a method for spatial interpolation based on Gaussian processes. ATRPK incorporates fine spatial resolution ancillary data into ATPK for image downscaling by regression

modeling. [2] Finally, there is also the AATRPK method, which improves ATPRK by allowing the regression coefficients to change throughout the image. [3]

This project intends to obtain high temporal resolution of land surface temperature images using a Convolutional Neural Network (CNN) with the aim of overcoming some of the disadvantages of the classical methods, namely the need to have the NDVI images (obtained from the VNIR and SWIR images) at the same time and position as the LST images. The model would still need these images to get trained, but afterwards it would be able to produce LST images at a higher resolution without them. It does, nevertheless, still assume that the hypothesis do not change with the scale.

Methodology

We aim to add texture detail to a CNN for superresolution. As such, we decided to build upon the Multi-residual U-Net architecture used in Modis Land Surface Temperature Super-Resolution [1], as it was shown that this architecture could outperform other neural network methods such as VDSR and DMCN as well as statistical methods such as ATPRK. Our implementation differs because of one modification. We preserved the original U-Net max-pooling for the encoder downsampling and used bilinear interpolation on the decoder instead of using strided convolutional layers to avoid striking problems. The MODIS LST images we used are of size 64x64 at 1 km spatial resolution and the output should be at 256x256 with a spatial resolution of 250 m.

In order to incorporate the texture information from the MODIS NDVI images (which are of size 256x256 at 250 m), we decided to modify the loss function. We introduced the Mixed Gradient Loss, which combines the traditional Mean Squared Error (MSE) with a new Mean Gradient Error (MGE) term. This MGE term is responsible for carrying texture information present in NDVI images. The Mixed Gradient Loss is calculated like so:

$$\mathrm{MixGL}(Y,\hat{Y}) = \alpha MGE + \beta MSE \begin{cases} MGE = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}(G(i,j)-\hat{G}(i,j))^2 \\ MSE = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}(Y(i,j)-\hat{Y}(i,j))^2 \\ \alpha+\beta=1 \end{cases}$$

The gradients are calculated using 3x3 Sobel filters and the hyperparameters alpha and beta are to be tuned. The training process involved utilizing a large database of over 12,000 (64x64 at 1 km) MODIS LST images captured during both day and night periods, along with their corresponding 6,017 (256x256 at 250 m) NDVI day images. These images were captured between the years 2015 and 2020, and were preprocessed to remove any pairs that contained either clouds or ocean. Once the pre-processing was complete, the images were split into training and validation sets using a 75/25 ratio.

To validate our approach, we used a single high-resolution ASTER LST image at 250m with the corresponding NDVI image. We then applied our model, as well as classical methods such as TSHARP, ATPRK, and AATPRK, and evaluated the results using the Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR) metrics.

Model description

The Multi-residual U-Net architecture used is based on the U-Net architecture, which was first introduced by Ronneberger et al. in 2015 [4]. It consists of a contracting path and a symmetric expanding path with long skip connections. The contracting path uses convolutional and max pooling layers to decrease input size and increase feature channels, while the expanding path upsamples the feature maps and merges them with corresponding maps from the contracting path for precise localization. The final layer generates the output.

The Multi-residual U-Net introduced by Nguyen et al. (2022) [1] differs in three main ways:

Firstly, the network takes a bicubic-interpolated low resolution image as input and maps it to a residual image which is the difference between the desired high resolution output and the input. The residual connection helps with the learning process and is similar to the residual correction step in statistical LST sharpening techniques.

Secondly, the architecture downsamples the feature map using a convolutional layer instead of a max-pooling one. The convolution operation considers all neuron values in its perceptive field, preserving large-scale spatial dynamics and local information of the LST image's feature maps. Moreover, the U-Net's fully convolutional encoder has been replaced with a residual-style encoder. This is achieved by replacing two consecutive convolutional layers with one residual unit followed by a convolution block. This approach uses residual learning to solve the problem of deep network degradation caused by gradient vanishing/exploding.

Finally, the Bridge also undergoes a modification by placing a transformation structure composed of one residual block (same as ResNet) between the encoder and the decoder. This transformation structure plays the role of a connection, transforming the features at the last encoding block to the first block of the decoding path, increasing the performance of the generator network in super-resolution objectives.

As for our model, we made a slight modification to the encoder and the decoder part. The downsampling in the encoder is done using the Original U-Net MaxPooling layer and not a strided convolution. As for the decoder part, the upsampling is done using a bilinear interpolation instead of a strided transposed convolution. This helped us get rid of the striking problems that we got on the output images when using strided convolutions.

Database

The MODIS instrument works on both Terra and Aqua spacecraft. It has an observation bandwidth of 2,330 km and observes the entire surface of the Earth every one to two days. Its detectors measure 36 spectral bands between 0.405 and 14.385 µm, and it acquires data at three spatial resolutions: 250 m, 500 m and 1000 m.

In our project, we focus on two MODIS products to build the database. The first is MODIS MOD11A1. This product provides daily Land surface temperature (LST) and emissivity per pixel with a spatial resolution of 1 kilometer (km) in a 1200 by 1200 km grid. We will use these images to train the model for super-resolution.

The second product used is MODIS MOD09GQ. This product provides an estimate of the surface spectral reflectance of the red and near-infrared bands of the Terra Moderate Resolution Imaging Spectroradiometer (MODIS) 250 meters (m), In addition to the 250 m surface reflectance bands, there are the quality assurance (QA) layer and five observation layers. These reflectance bands will then be used to calculate the normalized difference vegetation index (NDVI) which will be used to calculate the loss function during model training.

We use images of the central part of Europe which corresponds to the h18v04 tile in the MODIS geographical distribution. We thus cut our temperature images and our vegetation indexes from 1200x1200 and 4800x4800 to small patches of 64x64 and 256x256 respectively. By using quality masks, we eliminate images with sea or cloud pixel coverage. We also eliminate the vegetation index patches presenting NaN values due to the reflectance values of the two bands presenting values close to zero.

To ensure that the NDVI patches and the LST patches correspond to the same geographical location, we run a series of validation tests on the database images. We first check the visual correspondence between LST and NDVI images during different steps of the processing (The whole images before cutting them - The small patches after cutting - The patches after saving them in tiff format). We also check that the values of the NDVI images are between 1 and -1 and the values of the LST images are logical (Generally between 260K and 320K). In addition, we check that the geographical coordinates of the top left pixel of both the NDVI image and the LST image are the same.

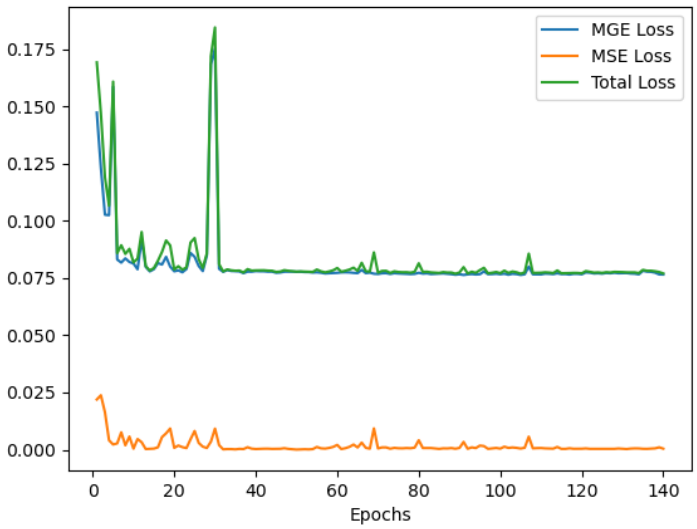
After downloading and processing the images from 2015 to 2020, we obtain 12034 temperature images of 64x64 at a spatial resolution of 1km. These images are coupled to 6017 vegetation index images of 256x256 at a 250m spatial resolution (Day and night temperature images are coupled to the same vegetation index calculated on the basis of daytime reflectance images). These images will constitute the training database of our model.

To validate the performance of our model, we use an image taken on January 26, 2017 in the Strasbourg region from the ASTER database. This database presents temperature images with a spatial resolution of 90m. We use linear sub-sampling to have the image at a spatial resolution of 250m to be able to use it to validate the performance of our model with metrics.

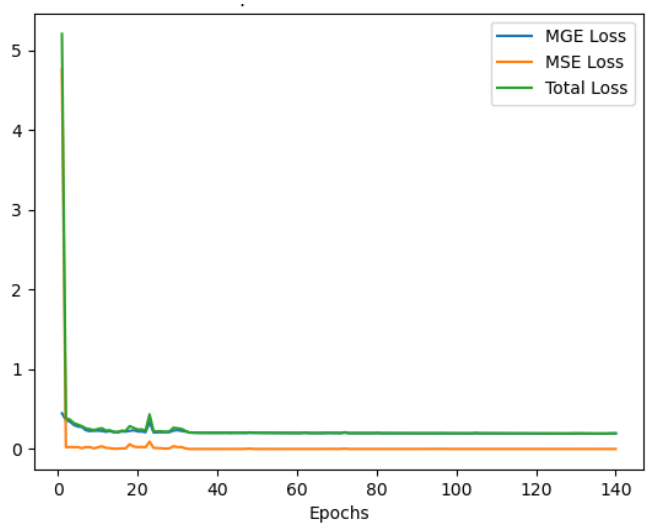
Results

Many different values of alpha and beta were tried and they yielded similar results, the training shown in this case was using alpha=0.001 and beta=0.999. We observe that the MSE loss goes down quite quickly while the MGE does not change over time, this might mean that the model cannot learn from the gradients and is thus not able to retain the information correctly.

Apart from the different alpha and beta combinations that were tried (from alpha 0.001 to 0.999 and beta=1-alpha), the idea of changing alpha and beta during training was entertained but again yielded no positive results.

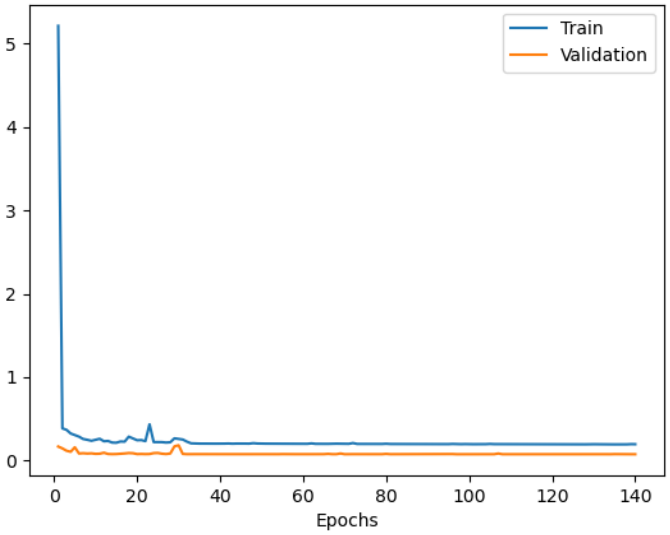


Validation losses

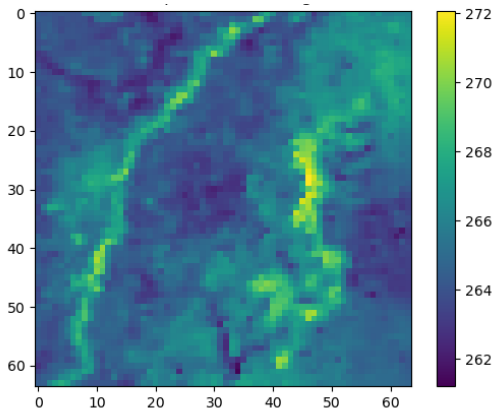


Train losses

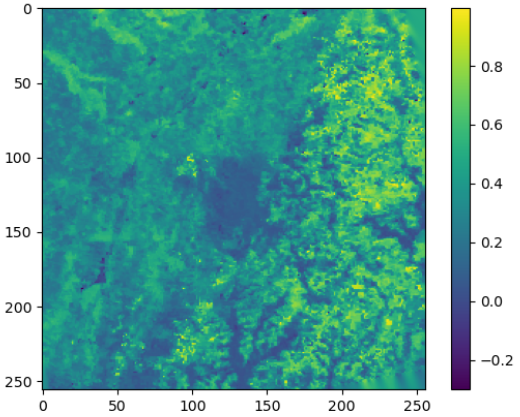
Model losses



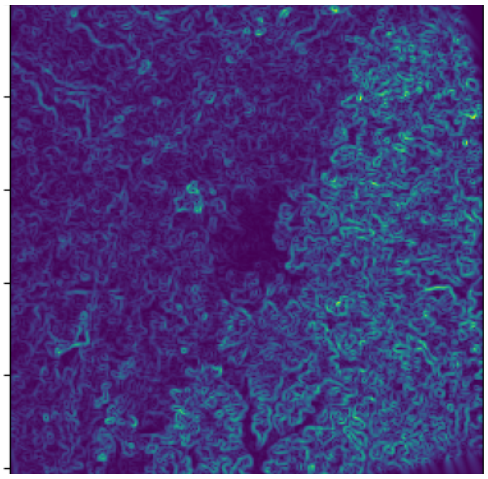
IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom



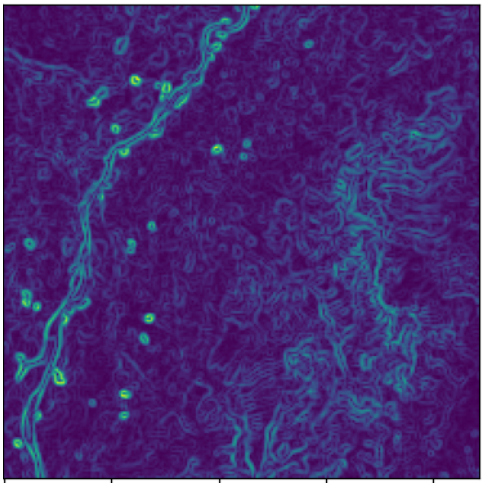
Original LST image (1km, 64x64)



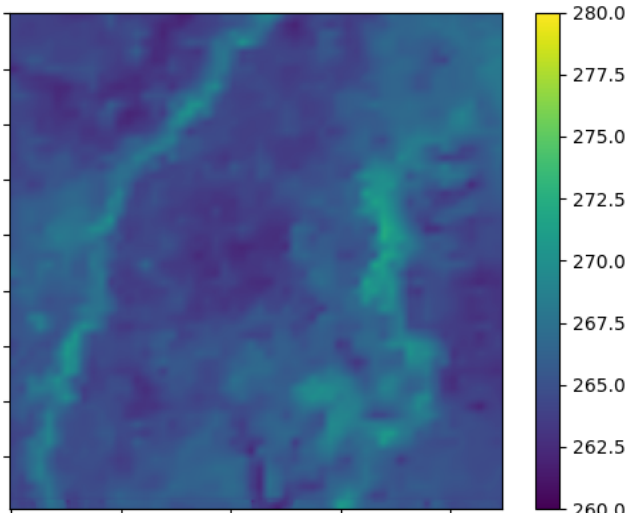
Input NDVI image (250m, 256x256)



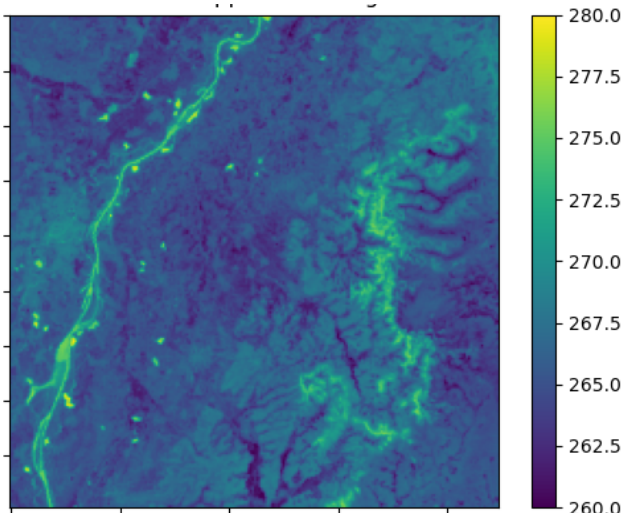
NDVI gradient
(250m, 240x240)
(normalized 0.0-1.0)



ASTER ground truth gradient
(250m, 240x240)
(normalized 0.0-1.0)



Model output LST (250m, 240x240)



ASTER ground truth LST (250m, 240x240)

For the ASTER image used as the test set, we got the following results when removing the 16 pixel edges from the outputs.

Metric	Our Model	TSHARP	ATPRK	AATPRK	Bilinar
PSNR	22.36	22.57	22.77	22.62	22.73
SSIM	0.43	0.45	0.47	0.47	0.45

Even if the original image is subdivided into 4 32x32 images (thus increasing the test set by x4) or even 16 16x16 images the results are very similar, we find that there is not a very big difference between the model, the classic methods and a simple bilinear upscaling.

Conclusions

Since our test size is only one image, the results are not that conclusive. Nevertheless, since the metrics of the models output and the bilinear transform are very similar we can conclude that the super resolution doesn't seem to perform as well as the statistical methods for this ASTER image. This might be for one of two reasons. First, as we saw in the previous images, the NDVI gradient and the ASTER gradient are not very similar in the leftmost half. This goes against one of the main hypotheses proposed during training, and since it is not valid, the result might differ from what is expected. Second, the test image itself might be hard to super resolve given the terrain and situational characteristics. Since all methods give similar results we cannot conclude with a high certainty that the model is much worse. In general, the idea of not having the NDVI as an input of the model (as was proposed in the beginning of the project) yields results that are not very satisfactory using the presented training and validation scheme. However, to fully conclude the model effectiveness at doing super resolution when compared to classic methods a bigger validation test set should be used.

However, we should note that the results shown suffer from the training problem highlighted in the above paragraph and solving this problem could possibly yield better results**.** In addition, we can also observe some fluctuations in the evolution of the loss function. This is probably due to the batch size being too low (The batch size is set to 4 due to GPU memory limitations). Thus, another path to explore in order to improve training is to use smaller images (32x32 or 16x16) and increase the batch size. Nevertheless, the impact of using smaller images on the performance should be studied.

Notwithstanding, a better result might be achieved if one were to use the findings from the following paper, *[5] Super resolution reconstruction method for infrared images based on pseudo transferred features* which used the infrared images to perform upscaling (instead of the NDVI) and used a GAN architecture.

Bibliography

[1] *Binh Minh Nguyen, Ganglin Tian, Minh-Triet Vo, Aurélie Michel, Thomas Corpetti, et al..* Convolutional Neural Network Modelling for MODIS Land Surface Temperature Super-Resolution. 2022 30th European Signal Processing Conference (EUSIPCO), Aug 2022, Belgrade, Serbia

[2] *Qunming Wang, Wenzhong Shi, Peter M. Atkinson, Yuanling Zhao*, Downscaling MODIS images with area-to-point regression kriging, Remote Sensing of Environment, Volume 166, 2015, Pages 191-204,

[3]Q*unming Wang, Wenzhong Shi, Peter M. Atkinson,* Area-to-point regression kriging for pan-sharpening, ISPRS Journal of Photogrammetry and Remote Sensing, Volume 114, 2016, Pages 151-165.

[4] *O. Ronneberger, P. Fischer, T. Brox*, U-net: Convolutional networks for biomedical image segmentation, in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234-241.

[5] *Shengyan Zhu, Caiqiu Zhou b, Yongjian Wang* Super resolution reconstruction method for infrared images based on pseudo transferred features.