Monitoring Night Skies with Deep Learning*

Yuri Galindo[1], Marcelo De Cicco[2,3,4], Marcos G. Quiles[1], and Ana C. Lorena[5]

[1] Univ. Fed. São Paulo - UNIFESP, São José dos Campos - SP, Brazil
[2] INMETRO, Divisão de Metrologia Cientifica, Rio de Janeiro - RJ, Brazil
[3] Observatório Nacional, Rio de Janeiro - RJ, Brazil
[4] EXOSS, Rio de Janeiro - RJ, Brazil
[5] Inst. Tecn. Aeronáutica - ITA, São José dos Campos - SP, Brazil

Abstract. The surveillance of meteors a.k.a. "shooting stars" is important due to the possibility of predicting future threatening impacts, uncovering new meteor showers, and gaining insight on our solar system. Camera-based surveillance systems often capture non-meteor objects such as planes and clouds, producing a large quantity of data that must be filtered. The automation of this filtering task greatly reduces the workload of astronomers, that otherwise may need to classify over 100 images per night. This represents a challenge due to the variety of non-meteor objects and environment conditions e.g. rain, clouds, and camera glitches. We apply Convolutional Neural Networks for this classification task and propose the use of calibration in order to improve the identification of confident predictions. In a dataset composed of images from the EXOSS surveillance system, our method is able to automatically classify 60% of the images with high confidence while maintaining accuracy and precision of 98% and recall of 99%. Additionally, we use a revised methodology for partitioning the data that takes into account the region of the captures and show that the usual methodology of randomly partitioning the data leads to overestimation of performance. Code is made available at github.com/yurigalindo/DeepLearningMeteors.

Keywords: Deep Learning · Computer Vision · Uncertainty Estimation

1 Introduction

EXOSS[6] (*Exploring the Southern Sky*) is a non profit organization that studies and monitors the incidence of meteors in Brazil [4, 5]. The project is centered on the citizen science model, counting with 56 stations equipped with CCTV (*Closed-circuit television*) cameras that monitor the night sky by capturing moving objects, such as meteors. The surveillance of meteors can be used to aid defense strategies against possible impacts on populated areas, to detect new meteor showers, and to study the origin of meteors and ultimately of the Universe.

* Supported by FAPESP (2018/20508-2)
[6] http://press.exoss.org/

EXOSS gathers information about the location, trajectory and velocity of the captured objects, among other data. The cameras undesirably capture other events, such as the passage of aircrafts, storms, birds, insects and upper atmosphere electrical discharges. The captures must then be filtered to keep the database populated solely by meteor information. This task is currently done by astronomers involved in the project, that may need to classify over 100 images a day in order to maintain the database free of non meteor captures.

The dataset used in this work consists of 5,620 labeled images, with roughly the same number of examples per class (namely meteor versus non-meteor). Difficulties of detecting meteors from automatically captured images include the high variance of the non meteor class, that may contain any non meteor object; the object of interest being small, out of focus or partially occluded; the image containing glitches and noises; and the presence of various objects in the same picture. Objects that trigger non meteor captures can also be present in meteor captures, such as passing satellites and clouds. Examples of captures are present in Figure 1.

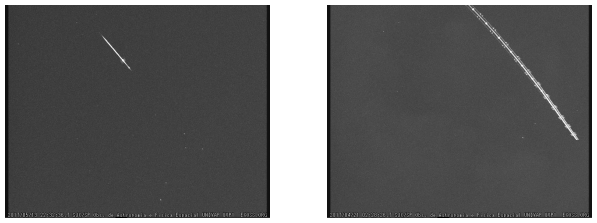

Fig. 1: Meteor (left) and non-meteor (right) examples.

We apply Convolutional Neural Networks (CNN) to the task of classifying the images captured by a meteor surveillance system as meteor or non-meteor. We use a Resnet50 network pre-trained on ImageNet [15] and fine-tune it to our data, while maintaining the first 33 layers frozen. We calibrate the likelihood given by our model on a separate dataset to improve the confidence estimate of the predictions, and use this estimate to select the most confident predictions. The goal is to leave only images for which the CNN is not confident to be examined by the human experts, reducing their overload. With this method we are able to achieve 98% of accuracy by predicting on 60% of the data, improving upon the baseline 90% of accuracy without any thresholding and 94% of accuracy predicting on 60% of the data with thresholding but no calibration. Our use of calibration increased the granularity of the choice of the threshold, providing more freedom for setting the desired accuracy level. The identification of uncertain examples is especially relevant for surveillance systems since novel events may be of interest for human inspection.

We also apply a different methodology for partitioning the dataset into training and test data that takes into account the station (location) of the captures,

differing from previous work [3,6,9,14,19] that randomly selected images for composing the training and test partitions. We show that the previous methodology leads to overestimation of performance in our dataset: a model with estimated 97.7% of accuracy by predicting on randomly separated images actually obtains 90% of accuracy when predicting on unseen regions. Our proposed methodology provides a more reliable estimate of the performance of the model for new captures. This is specially relevant for contexts in which new capturing stations can be added at any time, such as EXOSS.

This paper is structured as follows: Section 2 describes other work done on image classification of meteor data. Section 3 presents the dataset and methods used in this work. Section 4 presents the experiments performed and the obtained results. Section 5 discusses the results and presents future work perspectives, and Section 6 concludes this work.

2 Related Work

Previous works have tackled the problem of classifying captures from meteor surveillance systems with shallower neural network models [6], CNNs [3,9,14,19] and Recurrent Neural Networks [6, 14]. The datasets range from small, with 1,660 captures, to very large datasets with over 200,000 captures. These works reported accuracies varying between 84.35% to 99%, but separated training and test data randomly, which may lead to overestimation of performance and is further discussed in Section 4.

Galindo and Lorena [9] made use of a dataset of 1000 meteor images and 660 non meteor images, employing deep CNNs that ranged from 18 to 101 layers. The effectiveness of transfer learning and data augmentation was evaluated and found to be advantageous on this dataset. An 18 layer CNN pre trained in Fashion-MNIST [28] achieved 96% of accuracy and 0.94 of F1 score. Data augmentation based on image flipping and rotation were found to worsen accuracy, which may be related to the fact that the meteor trajectories produced are unlikely.

Marsola and Lorena [19] worked with a reduced subset of the dataset used in [9], and employed data augmentation, dropout and other common strategies adopted in deep learning studies. The achieved average accuracy was of 84.35%.

De Cicco et al. [6] employed a random forest classifier based on engineered and selected features related to the trajectories and light curves of each object, achieving 90% of precision and 81% of recall in a dataset comprised of 200,000 CAMS (Cameras for Allsky Meteor Surveillance) captures of which approximately 3% contained meteors. A CNN with five convolutional and two fully connected layers was trained from scratch and used to classify images that contained meteors. The reported precision and recall rates for the test set were 88% and 90%, respectively. A Long Short Term Memory Network (LSTM) was also trained to classify light curve tracklets (pieces of the detected trajectory), achieving 90% of precision and 89% of recall.

Gural [14] used Deep Learning techniques in a balanced dataset of 200,000 CAMS captures and reported up to 0.99 of F1 score, with 99.94% of precision

and 99.6% of recall. As in De Cicco et al. [6], they employed CNNs for classifying images and LSTMs for classifying light curve tracklets. The LSTM architectures achieved better results than on [6], but were surpassed by the CNN approach. They experimented adding additional channels to the CNN input images by incorporating multiple frames of the video capture, but this additional information did not lead to an improvement of predictive performance. The final architecture was a CNN comprised of four layers with input of a single frame, trained from scratch.

Cecil and Campbell-Brown [3] achieved 99.8% of accuracy by identifying regions of interest and applying CNNs to the identified regions. Their training set consisted of 50,475 images, and the test set of 5,485 images.

We contribute upon these works by showing that the methodology for training and testing the predictive models should take into account the region and date metadata, and by applying calibration and confidence thresholding to identify uncertain examples that may lead to mistakes in automatic classification. The identification of uncertain examples makes it possible to identify novel events that may be of interest to astronomers, and was not explored on these previous works.

3 Materials and Methods

This section describes the materials and methods employed in this work.

3.1 The data

Meteors, popularly known as "shooting stars", are objects from outer space that enter the Earth's atmosphere at high speed, suffering ablation (loss of mass) that produces light and ionisation. By analyzing this phenomenon, astronomers can infer physical and chemical properties of the object. These objects may be fragments of asteroids or comets, and represent material remaining from the formation of the solar system. The passage of a comet may leave behind debris that enters the Earth as meteors, in events knows as "meteor showers" that attract the interest of human observers and can greatly increase the risk of hazard to spacecrafts [21].

The EXOSS program counts with volunteers for setting up and maintaining surveillance stations across Brazil with cameras that monitor the night sky. These cameras record throughout the night and employ the UFOCapture software [20] for detecting events of moving objects. The recorded events, or captures, are stored on the EXOSS servers in different formats, e.g. videos of the captured event, light information, and images that summarize the event. These images are currently filtered manually by volunteer astronomers afterwards.

The dataset used in this paper consists of 2,971 meteors and 2,649 non meteors images from the EXOSS initiative, taken by cameras from 52 different stations and recorded between January of 2016 and May of 2019, with a median of 40 captures per station.

3.2 Convolutional Neural Networks

Convolutional neural networks (CNNs) are feed forward neural networks inspired by the visual cortex [18] that aggregate the weights into kernels. The kernels perform a 2D convolution with the inputs, resulting in a 2D output that preserves spatial information and goes through an activation function, resulting in an "activation map". The activation map resulting from each kernel will serve as an input channel for the subsequent layers. The kernels from lower layers have been shown to capture low-level features such as curves and lines [30], and the last layers can capture high-level features such as the presence of an object.

The ability of CNNs of learning abstract and generic features has led to the adoption of the activation of high-level layers as descriptive features for various applications, such as style transfer [10], image captioning [27], and transfer learning [26]. Transfer learning for CNNs works on the assumption that the high-level features learned for classification in one domain are useful for classification on a new domain. This approach has been successful for applying deep learning models in problems with small amounts of training data [7]. A network trained on a large amount of data can be used as a feature extractor for a classification algorithm, or as an initialization for the weights that will be further trained in the new data to extract more specific features, in a process called fine tuning.

Our model is based on the Resnet50 architecture presented in [15], pre-trained in the ImageNet dataset. We fine-tune the network to our data while maintaining the weights of the first 33 layers fixed, also known as "frozen" layers. We selected this architecture through cross-validation on the task of predicting on new regions. Our exploration of architectures was focused on the sizes of 18 to 101 layers of the Resnet family presented in [15], and in determining how many layers to freeze.

3.3 Estimating Confidence

Deep Learning models for classification often adopt a softmax function in the last layer [2], that guarantees that the outputs are positive and sum up to 1, resembling probabilities. These models are generally trained using the Cross-Entropy loss function [11], that is minimized when the predicted probabilities correspond to the true likelihood of the predictions given the data.

This probabilistic interpretation of the output of Neural Networks for classification has led to the usage of confidence thresholds for estimating uncertainty in practical applications, as in [12] and [23]. Classifications for which the largest predicted probability is under a threshold are deemed uncertain and are not output by the model. This approach trades coverage (the ratio of predicted examples) for higher accuracy in the predictions, and can increase accuracy in the reduced set containing the examples with higher predicted confidence.

The use of the softmax output as an estimate of uncertainty has been criticized in works such as [22] and [8]. Nguyen et al. [22] shows that Deep Neural Networks trained in ImageNet may assign high probabilities of a certain class for examples of unrecognizable images such as abstract shapes and white noise.

Gal [8] shows that a model can output a high probability even when predicting on out of distribution examples, and proposes a framework for using Deep Learning models for Bayesian modelling.

Guo et al. [13] showed that the probabilities produced by modern neural networks are miscalibrated: the confidence produced by the softmax function does not match the average accuracy of new predictions for that confidence level. Miscalibration is closely related to increased model capacity, that makes it possible for the neural network to minimize the loss on the training set by being overconfident on the correct predictions. The proposed solution is to use post-processing calibration methods that are trained on top of the model predictions on a validation or calibration set. Among these methods are the use of a logistic regression on top of the uncalibrated predictions, known as Platt scaling [25], temperature scaling [13], a variant of Platt scaling that is comprised of the multiplication by a constant before the softmax layer, and isotonic regression.

Isotonic regression refers to the problem of fitting a non-decreasing function that minimizes mean squared error [1]. A common implementation is non-parametric and uses the pool-adjacent violators (PAV) algorithm, fitting a piece-wise constant function [29]. The PAV algorithm works by finding neighborhoods of points that violate the non-decreasing property and replacing then with their mean. In the context of calibration, the ordering of the points is given by the uncalibrated probabilities, and the function is the true label of each data point. The violation of the non-decreasing property happens for wrong predictions, in which the algorithm assigns a higher value to a point of the negative class or vice-versa. Since the PAV algorithm replaces these regions with the observed mean, the isotonic fit effectively bins uncalibrated probabilities together, replacing them with the true observed accuracy.

Our context requires the identification of uncertain predictions in order to automatically classify examples while maintaining high accuracy, leaving the most difficult classifications and novel events for human inspection. As in [12] and [23], we use confidence thresholds to identify the predictions that the model can make while maintaining high accuracy. We improve this technique by applying the calibration strategy discussed on [13], achieving better confidence estimates that results in better accuracy and coverage.

Our approach uses a calibration set, in which we calibrate the probabilities and choose the desired threshold based on the accuracy and coverage needs for the application. This practical use of calibration results in more reliable confidence estimates on the test set, leading to a more effective trade off of accuracy versus coverage.

4 Experiments

The CNN models were trained with cross-entropy loss for 30 epochs with the Adam algorithm [16]. A set of 500 random images separated from the training set was used to detect the epoch with best validation accuracy for early stopping. The following algorithm parameters were considered: $\beta_1 = 0.9, \beta_2 = 0.999$, a

learning rate of 3e-4, and a learning rate decay of 0.99. The experiments were run using the PyTorch [24] library on the Google Colaboratory platform, on a Tesla K80 GPU with 2,496 CUDA cores and 12 GB memory. The developed code is available at GitHub[7].

4.1 Effect of data splitting

In possession of metadata regarding the region and timestamp of the captures, a high correlation between some captures can be observed. Captures from similar periods of time may register the same object multiple times, for example. A common occurrence of this phenomena is related to storms, when the same group of clouds may be the focus of hundreds of pictures, as displayed in Figure 2. Some types of objects are also restricted to certain regions, and regions have differing frequencies of types of objects. Cameras close to airports may have frequent captures of aircrafts, rarely seen in other regions, while some regions have insects that do not appear on others.

In order to take account of the correlation present in the data, we propose splitting the training and test data by different dates of capture or different regions. The date split prevents the presence of images from the same event in the training and test sets, and the region split also prevents this effect while measuring the generalization for unseen events and new camera and ambient conditions.

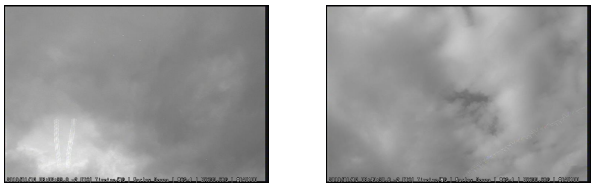

Fig. 2: Captures from the start and end of a storm, a series spanning 93 pictures.

Table 1: Observed performance with different splitting criteria

Criteria for splitting the data	Accuracy	Precision	Recall
Random	97.7 ± 0.5 %	98 ± 1 %	97.9 ± 0.9 %
Distinct dates	95 ± 1 %	94 ± 1%	97.7 ± 0.7 %
Distinct regions	90 ± 7 %	90 ± 9 %	95 ± 1 %

The experiment in Table 1 is a comparison between the estimated average accuracy performance (and standard deviation) using the methodologies of randomly splitting the data, separating images by date of capture, and separating

[7] https://github.com/yurigalindo/DeepLearningMeteors

by region. The separation between regions was performed by randomly selecting 10 regions (average of 1,620 images) for the test set and using the images of the remaining 42 regions for training. For the methods of randomly separating the data and separating by regions, we evaluated performance on four different training and test splits with five different network initializations for each split, totaling twenty experiments for each methodology. The separation between dates was performed by selecting six months across the three years of captures (1,417 images) for the test set. Due to the concentration of captures of some regions on specific periods, such as regions that had heightened activities on some periods or that were added or removed more recently, the months were selected as to have all regions present on both training and test sets, which prevented evaluation with different choices of months for testing. We evaluated this methodology with fifteen different network initializations.

The accuracy estimated by testing the model on new date periods is of 95%, lower than the 97.5% of accuracy obtained when predicting on randomly selected new images. This is an evidence of the overestimation of performance due to the presence of images with similar timestamps on training and test data that may come from the same event. The lower estimated accuracy of 90% when predicting on new regions befits this hypothesis and suggests trouble with generalization to unseen objects and new station setups. The greater standard deviation of the region split when compared to the random split indicates that some separations of regions are harder than others due to greater dissimilarity between the training and test sets, whereas when randomly splitting the data the dissimilarity between the training and test set is more constant. This analysis cannot be extended to the date separation since only one split was evaluated, in face of the temporal constraints for splitting the data.

Separating training and test sets by region is the appropriate setup for our application, since new stations are constantly added to the EXOSS program and measuring the performance of the model when applied to new regions is more representative of how the model would be applied in practice. This approach might not be feasible in some cases, such as surveillance programs that have only a single station, in which case the separation by date should be used to prevent the presence of images of the same event on training and test sets. This methodology will be adopted on the following experiments and should be taken into account for future works regarding meteor surveillance. This finding is also applicable to other classification problems focused on surveillance data that rely on automatic criteria to capture events, such as the detection of animal or human activity.

4.2 Estimating Uncertainty

Our experimental setup consisted of separating three regions (average of 548 images) from the training set in order to create a calibration set. The calibration models are trained on the predictions made on this set, and evaluated at the region separated test set as described in Section 4.1. This experiment was per-

formed twenty times, with four different test set splits and five different network initializations for each split.

Table 2 shows the performance of the different calibration methods, operating at various confidence thresholds. We considered the confidence as the estimated probability of the predicted class. For instance, if the model predicts that an image has 80% of probability of being a meteor, the estimated confidence is of 80%. With the use of confidence thresholding, we only make predictions with confidence equal or greater than the threshold, and discard the remaining examples to be classified by humans afterwards. Coverage refers to the proportion of examples that are above the threshold and not discarded. The evaluated confidence thresholds for each model were 99%, 95%, 90%, 85%, 80%, 75% and 70%. Thresholds with performance similar to those displayed were omitted for brevity.

Table 2: Performance of thresholding strategies

Calibration method	Threshold	Accuracy	Coverage	Precision	Recall
None	99%	94 ± 8 %	70 ± 20 %	92 ± 10 %	98 ± 1 %
None	95%	92 ± 8 %	80 ± 10 %	90 ± 10 %	98 ± 1 %
None	None (50%)	90 ± 7 %	100%	90 ± 9 %	95 ± 1 %
Temperature Scaling	99%	99.3 ± 0.6 %	20 ± 20 %	99 ± 1 %	99.6 ± 0.9 %
Temperature Scaling	90%	98 ± 2 %	60 ± 30 %	98 ± 3 %	99 ± 1 %
Temperature Scaling	85%	97 ± 3 %	70 ± 30 %	97 ± 4 %	98 ± 2 %
Temperature Scaling	75%	95 ± 6 %	80 ± 20 %	94 ± 7 %	97 ± 2 %
Isotonic Regression	99%	97 ± 4 %	50 ± 20 %	99 ± 1 %	92 ± 20 %
Isotonic Regression	85%	94 ± 5 %	80 ± 10 %	96 ± 5 %	93 ± 5 %

The use of thresholding was effective at improving performance at the cost of predicting on less examples. The uncalibrated model obtained 94% of accuracy at the highest confidence level, reducing the error rate without any threshold from 10% to 6% by predicting on 70% of the examples. Calibrating the model pushes this effect further, achieving 99.3% of accuracy and 99.6% of recall with 20% of coverage for the temperature scaled model and 97% of accuracy and 99% of precision with 50% of coverage for the model with isotonic regression.

The use of calibration has led to a better control of the accuracy and coverage trade off. The 94% accuracy of the uncalibrated model at 99% confidence is similar to the temperature scaled model with 75% confidence and the isotonic regression model at 85% confidence. These models obtain 99.3% and 97% accuracy at the confidence level of 99% respectively, displaying performances that better reflect the high confidence. This indicates that the uncalibrated model was overconfident, a behavior consistent with the findings of [13]. Calibration also led to lower standard deviation of accuracy and precision, making the desired accuracy given a confidence level more reliable. The temperature scaled model provides more advantageous choices and a more fine grained relationship between coverage and accuracy. It obtains better accuracy and precision when operating at the same confidence level of other models, displaying a more real-

istic estimate of confidence and making it possible to choose confidence levels that can maintain high accuracy, precision and recall.

The choice of confidence threshold is dependent on the needs of accuracy, precision and recall of the application. For our application, the temperature scaled model operating at 90% confidence provides a balanced and appropriate choice, maintaining accuracy and precision at 98% and recall at 99% while classifying 60% of the images.

We also evaluated the use of Platt scaling, but the method performed poorly. For the threshold of 99% the model obtained null coverage, not predicting on any image. For the thresholds of 95% and 90% of confidence, the model would predominantly not make any predictions for the meteor class, obtaining undefined precision and recall. The achieved accuracy and coverage levels were also not better than the uncalibrated model. This poor performance is explained by the fact that the uncalibrated model was already calibrated to some extent, which violates some assumptions of the logistic model and is analysed in depth in [17]. A violated assumption that may have resulted on the model attributing low confidence to meteor predictions is the assumption that the scores of negative and positive predictions have the same variance, which is not applicable to our case in which the non-meteor examples have higher variance.

5 Discussion

The difference of accuracy between the data splitting criteria shows that the correlation between images from the same region and date of capture must be taken into account when estimating the performance of a model. Preventing images of the same event from being present when training and testing the model is essential, and both methods address this problem. The region split should be adopted when possible i.e. when there are enough stations and enough captures per station. This observation is also important for other applications that rely on automatic capture of events to generate images, in which there may be multiple images of a single event or different frequency of events depending on the region.

Our proposed use of calibration to improve thresholding was successful, managing to maintain high accuracy while still classifying over half of the images. This method is computationally efficient and straightforward to implement, and presents a simple way to estimate uncertainty that performed well in our application. The improved granularity provided by the calibration permits more freedom for choosing the confidence level appropriate for the application. This practical application of calibration indicates that more realistic confidence levels lead to better control of accuracy and better choices of thresholds.

6 Conclusion

In this paper, we addressed the problem of classification of meteor surveillance captures by using pre-trained Convolutional Neural Networks. We showed that

training and test data must be properly separated, ideally by regions and alternatively by date. This finding must be taken into account for future studies surrounding meteor classification, and also applies to other setups where images are automatically captured by surveillance systems. In order to make accurate predictions on unseen objects, we proposed the calibration of the confidence produced by our CNN and discriminated between confident and uncertain predictions, managing to improve accuracy by predicting on a reduced set of images.

7 Acknowledgments

We thank FAPESP for funding this research (grant 2018/20508-2), CNPq (grant 434886/2018-1), the EXOSS organization for providing the data and astronomy expertise, and Pete Gural for the discussions on meteor surveillance.

References

1. Barlow, R.E., Brunk, H.D.: The isotonic regression problem and its dual. Journal of the American Statistical Association **67**(337), 140–147 (1972)
2. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
3. Cecil, D., Campbell-Brown, M.: The application of convolutional neural networks to the automation of a meteor detection pipeline. Planetary and Space Science p. 104920 (2020)
4. De Cicco, M.: Bright meteor at the sky of espirito santo brazil. Copyright notices p. 76 (2017)
5. De Cicco, M.: Fireball captured on video by exoss stations at espirito santo's brazilian state. n Two slow meteors with spectra p. 12 (2017)
6. De Cicco, M., Zoghbi, S., Stapper, A.P., Ordoñez, A.J., Jack Collisno Peter S. Gural Siddha Ganju, J.L.G., Jenniskens, P.: Artificial intelligence techniques for automating the cams processing pipeline to direct the search for long-period comets. In: Proceedings of the IMC, Petnica (2017)
7. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning. pp. 647–655 (2014)
8. Gal, Y.: Uncertainty in deep learning. Ph.D. thesis, PhD thesis, University of Cambridge (2016)
9. Galindo, Y.O., Lorena, A.C.: Deep transfer learning for meteor detection. In: Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC) (2018)
10. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
11. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
12. Goodfellow, I.J., Bulatov, Y., Ibarz, J., Arnoud, S., Shet, V.: Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv preprint arXiv:1312.6082 (2013)
13. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 1321–1330. JMLR. org (2017)

14. Gural, P.S.: Deep learning algorithms applied to the classification of video meteor detections. Monthly Notices of the Royal Astronomical Society **489**(4), 5109–5118 (09 2019)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014)
17. Kull, M., Silva Filho, T.M., Flach, P., et al.: Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics **11**(2), 5052–5080 (2017)
18. LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D.: Handwritten digit recognition with a back-propagation network. In: Advances in neural information processing systems. pp. 396–404 (1990)
19. Marsola, T.C., Lorena, A.C.: Meteor detection using deep convolutional neural networks. In: Anais do Simpósito Brasileiro de Automação Inteligente. vol. 1, p. 104260 (2019)
20. Molau, S., Gural, P.: A review of video meteor detection and analysis software. WGN, Journal of the International Meteor Organization **33**, 15–20 (2005)
21. Moorhead, A.V., Cooke, W.J., Campbell-Brown, M.D.: Meteor shower forecasting for spacecraft operations (2017)
22. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
23. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences **115**(25), E5716–E5725 (2018)
24. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
25. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
26. Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M.: Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging **35**(5), 1285–1298 (2016)
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. IEEE transactions on pattern analysis and machine intelligence **39**(4), 652–663 (2016)
28. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
29. Zadrozny, B., Elkan, C.: Transforming classifier scores into accurate multiclass probability estimates. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 694–699 (2002)
30. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)