InceptionV3 and DeiT on Novel Street View Dataset for Pedestrian Risk Classification in Lagos

Chukwunwogor Faithful Chibuokem
Department of Computer Science
University of the People
595 E. Colorado Boulevard. Suite 623
Pasadena, CA 91101, USA
chukwunwogorchibuokem@my.uopeople.edu

Abstract

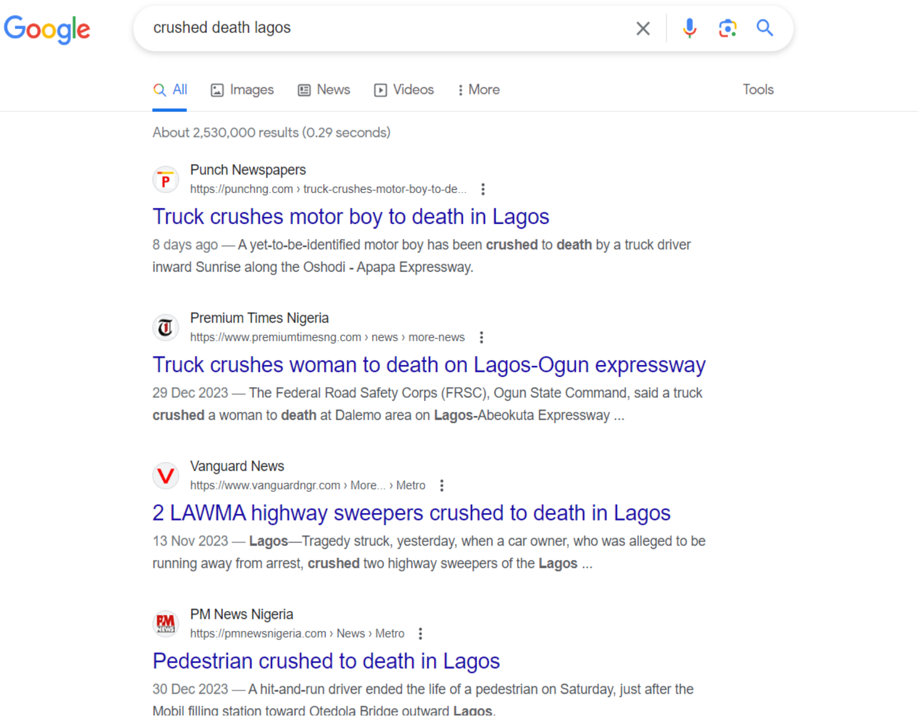
This research investigates the pressing issue of Road Traffic Injuries (RTIs) and their correlation with pedestrian risk in the densely populated urban environment of Lagos, Nigeria. With RTIs being a major contributor to trauma-related fatalities, particularly in regions undergoing rapid urbanization, this study pioneers the application of deep learning methodologies to address the challenges associated with pedestrian safety. The article introduces a novel dataset, GSVLagos1, consisting of 1384 Google Street View (GSV) images from Lagos, classified into four risk categories based on pedestrian counts. Leveraging state-of-the-art deep learning models, namely InceptionV3 and Data-efficient image Transformer (DeiT), the study explores transfer learning and fine-tuning techniques on this unique dataset. The research aims to bridge the existing gap in literature by proposing an innovative approach to pedestrian risk classification using GSV imagery.

Key contributions include the proposal of an automated pedestrian risk classification task, establishment of the GSVLagos1 dataset, demonstration of Transformer models' efficiency in comparison to convolutional approaches, and an investigation into the impact of normalization layer modifications on model performance. The study underscores the significance of utilizing GSV and street view imagery for comprehensive urban risk assessment and identifies potential areas for future research and improvement. Detailed insights into dataset creation, algorithmic considerations, model architectures, and training strategies are provided. Experimental results show comparable performance between InceptionV3 and DeiT, with validation accuracies reaching 69.51% and 68.64%, respectively. The study acknowledges challenges in dataset annotation, proposing recommendations for future research endeavours.

This research contributes to the advancement of knowledge in urban risk assessment by proposing a deep learning-based solution for pedestrian risk classification. The findings have implications for urban planning, infrastructure development, and traffic safety initiatives in developing urban areas.

1. Introduction and Related Work

Road Traffic Incidents (RTIs) are the prominent cause of trauma-related deaths and the third leading cause of overall deaths in Nigeria, with about 40,000 deaths per year (Ibrahim, Kumazhege, & L'Kama, 2023). This correlates positively with the rates of urbanization and population density in Nigeria, with infrastructure unable to keep up (Adebayo, Akinsanya, Coker, & Jolaawo, 2023). Unsurprisingly, Lagos, Nigeria's most densely populated city, shows high incidences of pedestrian morbidity and mortality with significant underreporting. (Babatunde, et al., 2015) found RTIs to occur most in highly populated areas. A Google search reveals several incidences in barely a month (Figure 1). Urbanization in Nigeria is influenced by road access (Nathaniel, 2020). Eventually, with inadequate road expansion and other infrastructure adjustments, congestion on the roads leads to disarray in traffic and elevates driver stress levels (Ryder, Gahr, Egolf, Dahlinger, & Wortmann, 2017).



crushed death lagos

About 2,530,000 results (0.29 seconds)

Punch Newspapers
https://punchng.com › truck-crushes-motor-boy-to-de...

Truck crushes motor boy to death in Lagos

8 days ago — A yet-to-be-identified motor boy has been **crushed** to **death** by a truck driver inward Sunrise along the Oshodi - Apapa Expressway.

Premium Times Nigeria
https://www.premiumtimesng.com › news › more-news

Truck crushes woman to death on Lagos-Ogun expressway

29 Dec 2023 — The Federal Road Safety Corps (FRSC), Ogun State Command, said a truck **crushed** a woman to **death** at Dalemo area on **Lagos**-Abeokuta Expressway ...

Vanguard News
https://www.vanguardngr.com › More... › Metro

2 LAWMA highway sweepers crushed to death in Lagos

13 Nov 2023 — **Lagos**—Tragedy struck, yesterday, when a car owner, who was alleged to be running away from arrest, **crushed** two highway sweepers of the **Lagos** ...

PM News Nigeria
https://pmnewsnigeria.com › News › Metro

Pedestrian crushed to death in Lagos

30 Dec 2023 — A hit-and-run driver ended the life of a pedestrian on Saturday, just after the Mobil filling station toward Otedola Bridge outward **Lagos**.

Figure 1: A Google Search on 7th January 2024 reveals an astonishing number of fatal pedestrian incidents in Lagos over few days.

Pedestrian safety is established to be negatively impacted by urbanization, when there is inadequate infrastructural development (Buhari, Aponjolosun, Oni, & Sam, 2020; Asaju, Olawepo, & Ojekunle, 2020). Locations like markets and bus stops experience heightened vehicular and human activity. Notably, safety measures such as delineators for pedestrian crossings, traffic control features, and speed indicators may be either absent or ignored. (D, Thomas, & J, 2015). Hence, it is necessary to quickly identify these high risk urban locations for infrastructure upgrade and strict enforcement of regulatory law.

There is a scarcity of research examining the application of deep learning to address Road Traffic Injuries (RTIs) using Google Street View (GSV) (Anguelov, et al., 2010) images. Current studies often depend on conventional statistical and geospatial methods (Mehta, et al., 2023; Adebayo, Akinsanya, Coker, & Jolaawo, 2023; Olusina & Ajanaku, 2017), along with remote sensing technologies (Adeofun & Oyedepo, 2011). This research gap underscores the need for a deep learning approach that leverages pedestrian scenes from GSV to identify high-risk areas for Pedestrian Injuries (PIs) and other RTIs.

Analysis based on remote sensing imagery lacks on-the-ground socio-economic semantic information necessary for a deeper understanding of human-space interactions (Zhang, Chen, Zheng, Chen, & Wang, 2021). Consequently, identifying urban functions using top-view imagery often falls short in quality (Voorde, Jacquet, & Canters, 2011). In contrast, proximate sensing perspectives like streetscape or street view imagery captured at high spatial resolution provide detailed urban information (Zhang, et al., 2020; Chen, Lu, Ye, Xiao, & Yang, 2022).Various studies highlight the efficiency of street view images in identifying urban functions (Xu, et al., 2022; Biljecki & Ito, 2021). Leveraging the advantages of GSV and other street view imagery, researchers have conducted studies to classify urban elements (Alhasoun & González, 2019). Computer vision methods are employed to learn features of urban structure for tasks such as street-level image classification (Santani, Ruiz-Correa, & Gatica-Perez, 2018), street-level mobility prediction (Zhang, Wu, Zhu, & Liu, 2019), street-level accessibility (Najafizadeh & Froehlich, 2018), building classification (Kang, Körner, Wang, Taubenböck, & Zhu, 2018), road-scene vehicle detection (Shyam & Pranjay, 2022), and land-use investigation (Srivastava, Vargas-Munoz, & Tuia, 2019).

Recently, researchers have utilized street view images to estimate pedestrian volume. Image-based pedestrian detection techniques can automatically count pedestrians with high accuracy (Chen, et al., 2020; Yin, Cheng, Wang, & Shao, 2015). For instance, the integration of street view images with machine learning algorithms has demonstrated high reliability in estimating pedestrian volume at the street level compared to field audits (Chen, et al., 2020). Therefore, the application of cutting-edge computer vision methods on street view images represents a promising approach for PI risk classification based on number of pedestrians in geolocated GSV images

This paper introduces a novel dataset comprising 1384 GSV images from the city of Lagos. These images are classified into four categories depending on the number of pedestrians observed in the image. I build on ImageNet pretrained InceptionV3 Convolutional Neural Network (CNN) (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016) and Data-efficient image Transformer (DeiT) (Hugo, et al., 2021) models to carry out transfer learning and fine-tune on the dataset. I investigate decay rate and regularization methods on InceptionV3 and DeiT to improve model generalization on the unique dataset. I also modify InceptionV3 normalization layer to test transfer learning and fine-tuning performance on a new normalization technique. My approach puts to test the ability of state-of-the-art (SOTA) deep networks to generalize to a unique dataset of developing urban areas, which are often underrepresented in image datasets (Cordts, et al., 2016; Neuhold, Ollmann, Bulo, & Kontschieder, 2017; Yu, et al., 2020), due to a lack of diversity in datasets used to pre-train SOTA models. My fine-tuned model conducts risk classification by leveraging on representations of roadside pedestrians from the GSV input image.

My main contributions can be summarized as follows:

1. I propose the task of automated pedestrian risk classification by considering the number of pedestrians in a GSV input and estimating a certain risk class for the corresponding location
2. I establish a novel dataset from an underrepresented area and delineate a reasonable classification approach to train deep learning networks to identify risk.
3. I demonstrate that Transformer models achieve comparable performance against convolutional approaches on the unique risk classification dataset due to its ability to capture the global context, long-range dependencies between pixels.
4. I demonstrate that changing the normalization layers of a convolutional network decreases transfer accuracy.
5. I demonstrate that by training computer vision models to classify pedestrian population on labelled GSV dataset, it is possible to automate risk assessment and PI hotspots in developing cities.

2. Dataset

I developed a model dataset; GSVLagos1, consisting of 1384 GSV images. To extract images from GSV API, I first designed a custom map route using Google My Maps. Next I extracted its KML data in text format containing positional data for every ~25 meters on the route, in longitude and latitude. Iteratively, each line is fed to the GSV API to retrieve the street view image for each location. Alongside positional data, GSV API requires image resolution, API key, heading and pitch. If heading is not supplied GSV API uses its algorithm to derive a heading it deems appropriate. First I developed an algorithm to calculate heading of x_1 taking reference from x_2

```
Algorithm 1: Calculate Heading from $(x_2)$ to $(x_1)$
    Inputs:
    $(x_1)$: $(\mathrm{lat}_1, \mathrm{lon}_1)$
    $(x_2)$: $(\mathrm{lat}_2, \mathrm{lon}_2)$
    Outputs:
    Heading: The initial bearing from $(x_2)$ to $(x_1)$ in degrees.
    Algorithm Steps:
    $\mathrm{lat}_1, \mathrm{lon}_1 \leftarrow \mathrm{deg2rad}(\mathrm{lat}_1, \mathrm{lon}_1)$
    $\mathrm{lat}_2, \mathrm{lon}_2 \leftarrow \mathrm{deg2rad}(\mathrm{lat}_2, \mathrm{lon}_2)$
    $\Delta\mathrm{lat} = \mathrm{lat}_1 - \mathrm{lat}_2$
    $\Delta\mathrm{lon} = \mathrm{lon}_1 - \mathrm{lon}_2$
    $a = \sin^2\left(\frac{\Delta\mathrm{lat}}{2}\right) + \cos(\mathrm{lat}_1) \cdot \cos(\mathrm{lat}_2) \cdot \sin^2\left(\frac{\Delta\mathrm{lon}}{2}\right)$
    $c = 2 \cdot \mathrm{atan2}(\sqrt{a}, \sqrt{1-a})$
    $\mathrm{bearing} = \mathrm{atan2}(\sin(\Delta\mathrm{lon}) \cdot \cos(\mathrm{lat}_2), \cos(\mathrm{lat}_1) \cdot \sin(\mathrm{lat}_2) - \sin(\mathrm{lat}_1) \cdot \cos(\mathrm{lat}_2) \cdot \cos(\Delta\mathrm{lon}))$
    $\mathrm{bearing} = \mathrm{rad2deg}(\mathrm{bearing})$
    $\mathrm{heading} = (\mathrm{bearing} + 360)\backslash\%360$
```

GSV API results returned using headings generated by algorithm 1 are as expected. However, upon investigation, I found too many instances of invalid pictures (invalid pictures are addressed below). Therefore I tried again without passing in heading data using algorithm 1. This time allowing GSV use a suitable heading. Investigation of these second set of results reveal a high incidence of rearview images. Although a less desirable outcome, it is not necessarily a disadvantage, and invalid images are much less this time.

To classify the images, I defined four categories of risk. The number of visible pedestrians in a single image determines the class.

Classes	*High_risk*	*Medium_risk*	*Low_risk*	*Invalid*
No of Peds	<10	3-10	>3	X

Table 1: Classes for the GSVLagos1 Dataset

Classified as invalid are images returned which do not show a proper view of the highway, road or street. This is done to prevent unnecessary diversity in the dataset and also to prevent misclassification resulting from biased views.

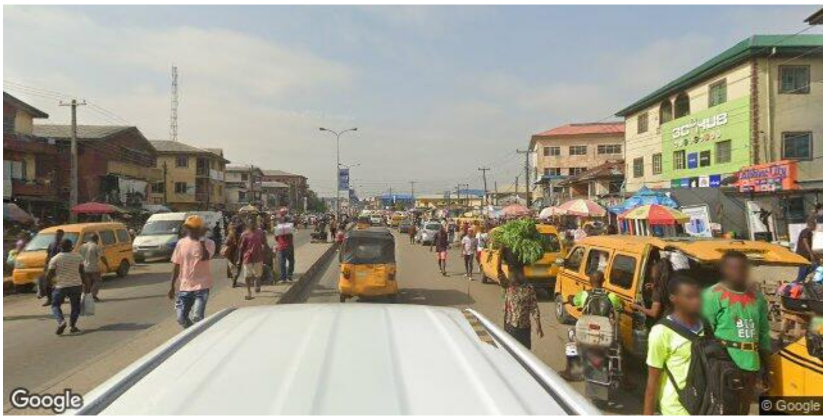


Figure 2: Sample of image in the dataset classified as 'High_risk'

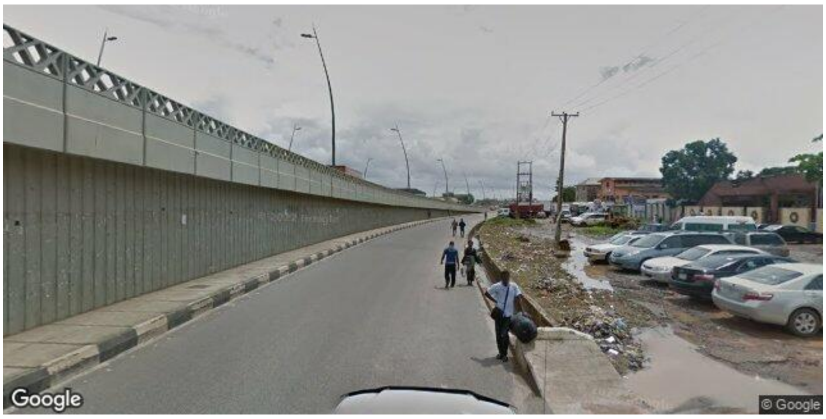


Figure 3: Sample of image in the dataset classified as 'Medium_risk'

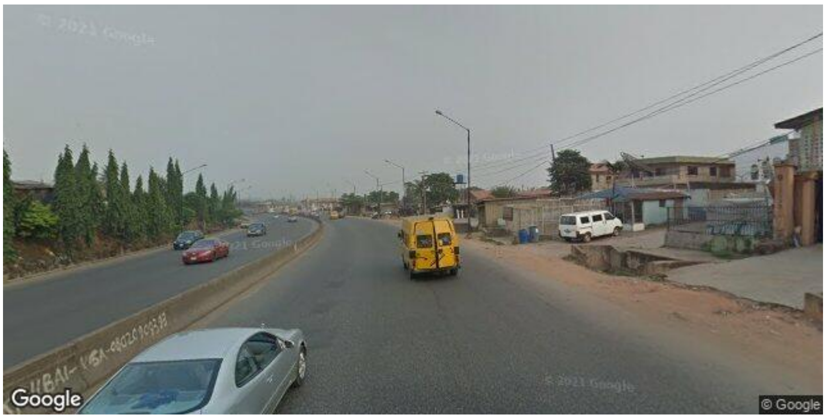


Figure 4: Sample of image in the dataset classified as 'Low_risk'

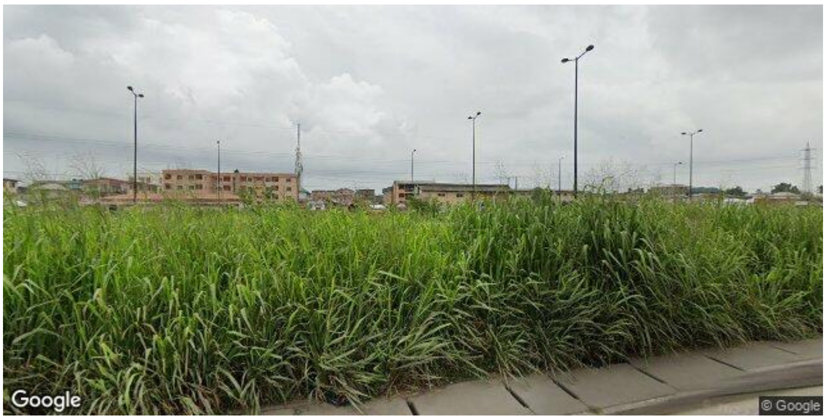


Figure 5: Sample of image in the dataset classified as 'invalid'

Classifying the dataset was far from straightforward. Due to high levels of variations in images obtained from the API, it was often difficult to classify images. Angles in the image vary, and pedestrians range from fully visible to barely identifiable. This is clearly a limitation to accuracy, and thus a better set of classes is recommended in future study. One suggestion is to classify images by level of urban activity rather than simply doing a headcount of pedestrians, which results in some ambiguity.

Horizontal flip is carried out as a weak data augmentation procedure (Debenedetti, Sehwag, & Mittal, 2023). Thus the dataset is expanded to 2768 images. Data augmentation procedures such as vertical flip crop and color manipulation, etc., are not considered relevant to this particular dataset. The dataset is then divided into the test and validation dataset in an 80/20 ratio.

3. Architecture

I will introduce the architectures used to compare performance on the GSVLagos 1 dataset. I have chosen two prominent SOTA models; InceptionV3 and DeiT.

3.1. InceptionV3

InceptionV3 is a convolutional neural network (CNN) designed by Google for image classification tasks (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Introduced in 2015 by Szegedy et al., it builds upon previous Inception architectures, featuring a deep and interconnected structure of inception modules (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). These modules leverage various convolutional and pooling layers to extract diverse features from input images.

Key features of InceptionV3 include the use of factorized convolutions, reducing network parameters while maintaining high accuracy (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Factorized convolutions enable learning both local and global features, enhancing model accuracy. The network employs 1x1, 3x3, and 5x5 convolutional filters to extract different features, with 1x1 filters reducing dimensionality.

Batch normalization stabilizes training and reduces internal covariate shift (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016). Pre-trained on the ImageNet dataset, InceptionV3 allows for transfer learning, where the model is fine-tuned for specific tasks using smaller datasets. This versatility makes it effective in various applications like object detection, image segmentation, and video classification.

InceptionV3's architecture, with approximately 4 million parameters, is more efficient than some counterparts like VGG (Wirayasa, 2021). The absence of a fully-connected layer and its replacement with a pooling layer contributes to a smaller model size, enabling faster computations. The model comprises initial convolution layers, inception convolution, repeated sections, and a stop-learning layer to prevent overfitting, with the second last layer being fully connected (Bhatia, Bajpayee, Raghuvanshi, & Mittal, 2019). InceptionV3 has demonstrated high performance across various benchmarks, establishing itself as a powerful and versatile model for image classification tasks.

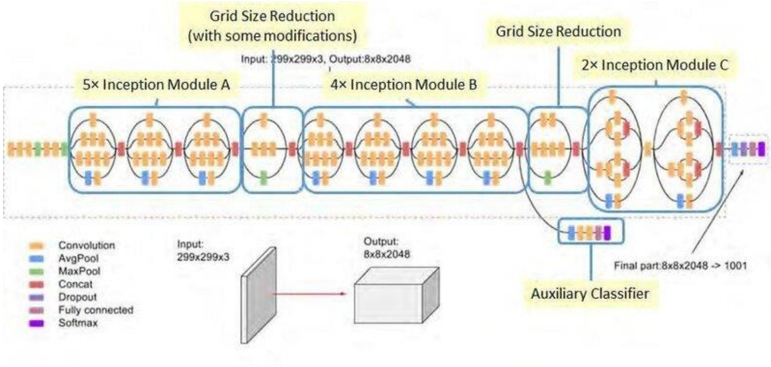


Figure 6: Architecture of InceptionV3 Network (Tsang, 2018)

3.2. Data-efficient image Transformer

The DeiT (Data-efficient image Transformer) model is built upon the transformer architecture, initially introduced for natural language processing tasks. The transformer architecture is adapted for computer vision, specifically image classification. The model processes images as sequences of tokens, inspired by the Vision Transformer (ViT) (Dosovitskiy, et al., 2020).

Multi-Head Self-Attention Layers (MSA)

The attention mechanism in DeiT is based on trainable associative memory with (key, value) pairs. A query vector $(q \in R^d)$ is matched against key vectors $(K \in R^{k\times d})$ using inner products. The attention output is obtained through weighted sums of value vectors. The attention function is defined as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \quad (1)$$

where the SoftMax function is applied row-wise, and $(\sqrt{d})$ provides normalization (Ashish, et al., 2017).

Transformer Block for Images

The DeiT model employs a full transformer block that includes Multi-Head Self-Attention (MSA) and a Feed-Forward Network (FFN) with skip-connections. For image processing, the architecture is inspired by ViT, where the input RGB image is decomposed into fixed-size patches. Each patch is linearly projected to maintain its overall dimension. Positional information is incorporated using fixed or trainable positional embeddings. The class token, appended to the patch tokens, is used for classification, and the transformer processes batches of tokens, including the class vector.

Distillation Through Attention

DeiT introduces distillation to leverage knowledge from a teacher model, which can be a convnet or a mixture of classifiers. Two axes of distillation are explored: soft distillation, minimizing the Kullback-Leibler divergence (Hinton, Vinyals, & Dean, 2015), and hard distillation, using the teacher's hard decisions. The distillation loss function is defined as:

$$\begin{aligned} L_{\mathrm{global}} = (1-\lambda) \cdot L_{\mathrm{CE}}(\psi(Z_s), y) + \lambda\tau^2 \\ \cdot \mathrm{KL}\big(\psi(Z_s/\tau), \psi(Z_t/\tau)\big) \quad (2) \end{aligned}$$

where (Z_t) and (Z_s) are logits of the teacher and student models, respectively. Hard-label distillation introduces a variant where the hard decision of the teacher is taken as a true label:

$$L_{\mathrm{hardDistill}} = \frac{1}{2} \cdot L_{\mathrm{CE}}(\psi(Z_s), y) + \frac{1}{2} \cdot L_{\mathrm{CE}}(\psi(Z_s), y_t) \quad (3)$$

where $(y_t = \mathrm{argmax}_c Z_t(c))$. The distillation token is introduced as a new token interacting with class and patch tokens, facilitating learning from the teacher's output.

4. Experiment and Results

I performed experiments with the models described in section 3 and dataset GSVLagos1 containing images classified according to levels of pedestrian risk.

4.1. Input Image Transformation

For each model, I devised two distinct image transformation pipelines tailored for training and testing the dataset. In the training pipeline, I implemented resizing, reducing the input image to 255 to match InceptionV3's input, and 384 for DeiT. For each training set transformation pipeline, I added a tensor transformation for converting pixel values to a normalized range of [0, 1], and a Normalize transformation that adjusted values based on predetermined mean and standard deviation. This normalization process ensures that the model is trained on consistent and standardized input data. For the testing pipeline, I performed the same resize transformation as testing, maintaining image proportions, and applied the same tensor and normalize transformations as done on the training set. The tensor transformation converts pixel values from the conventional range of [0, 255] to [0, 1], while Normalize standardizes the pixel values using the predefined mean and standard deviation.

4.2. Learning Rate

I took cue from (He K. , Zhang, Ren, & Sun, 2016) to choose learning rate as 0.1 × b/256, where b is the batch size of the network. During fine-tuning, I divided the learning rate by 10, to prevent destructive effects on the existing weights of the pretrained network.

4.3. Cosine Learning Rate Decay Strategy

I made use of cosine annealing strategy by (Loshchilov & Hutter, 2016), the learning rate (ηt) is determined using the cosine function, gradually decreasing from the initial value (η) to 0 over T batches. The equation for ηt is given by:

$$\eta_t = \frac{1}{2}\left(1 + \cos\left(\frac{t\pi}{T}\right)\right)\eta \quad (4)$$

This scheduling, termed "cosine" decay, is compared against well-known exponential decay strategy (He T. , et al., 2018). The comparison reveals that cosine decay slows at the start, transitions linearly, and slows again, potentially enhancing training progress compared to step decay.

4.4. Optimization Algorithm

I maintained RMSprop for InceptionV3 and Adam for DeiT as used in each original work to achieve best results.

4.5. Feed Forward Classifier Head

Initially I experimented with a single layer, but this is soon seen to hurt transfer performance. Therefore for InceptionV3 and DeiT, I designed a robust classifier head with a single hidden layer

Layer	Shape
Fully connected	(input_features, 512)
ReLU	(512,)
Dropout (30%)	(512,)
Fully connected	(512, 4)

Table 2: Feed Forward Neural network for the classifier head

4.6. Batch and Group Normalization

Group Normalization (GN) exhibits advantages over Batch Normalization (BN) by dividing channels into groups, thereby enhancing robustness and efficiency, particularly in scenarios with limited batch sizes. This approach, introduced by (Wu & He, 2018) in their study, offers improved performance and stability, addressing some limitations associated with Batch Normalization. Group Normalization's adaptability makes it a promising alternative for optimizing convolutional neural networks in various computer vision applications (Wu & He, 2018).

Pretrained InceptionV3 uses BN for its logits. However, I modify the architecture, replacing BN with GN. My intention is to see if pretrained parameters generalize well to GN and potentially improve model performance. However, this is not the case. There is no significant improvement up to epoch 23. Validation accuracy remains nearly the same with epoch 0. Using GN is therefore not possible with pretrained architectures already using BN. To take advantage of the benefits of GN, one must train from scratch. Thus GN is dropped in this case and BN is returned.

4.7. Label Smoothing

In InceptionV3 and DeiT, the final layer is a fully-connected layer outputting predicted confidence scores, denoted as (z_i) for class (i) out of (K) labels. Normalizing these scores using the softmax operator yields predicted probabilities (q_i). During training, the negative cross-entropy loss $(\mathcal{L}(p, q) = -\sum_{i=1}^{K} q_i \log p_i)$ is minimized, where (p_i) is the truth probability distribution constructed based on the true label (y). Label smoothing, introduced for Inception-v2 (Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016), modifies the true probability distribution to $(q_i = (1 - \epsilon))$ if $(i = y)$, and $(\epsilon/(K - 1))$ otherwise. The optimal solution adjusts (z_i) to $(\log\big((K - 1)(1 - \epsilon)/\epsilon\big) + \alpha) if\, (i = y)$, and (α) otherwise, encouraging a finite output from the fully-connected layer for better generalization. Label smoothing mitigates overfitting by preventing excessively distinctive output scores, addressing the limitation of the traditional training objective. I set ϵ to 0.1 for training.

4.8. Training Details and Results

I selected a batch size of 11. Using the equation defined in 4.2., learning rate is 0.0045 and 0.00045 for transfer learning and fine-tuning respectively. Training is carried out on Nvidia Tesla P100 GPU.

For InceptionV3, I carried out transfer learning for 50 epochs and then fine-tuned for another 50 epochs. Total time is 14+31 minutes.

Maximum validation accuracy is 0.6951 achieved at epoch 44 before accuracy begins to drop over the next few epochs.

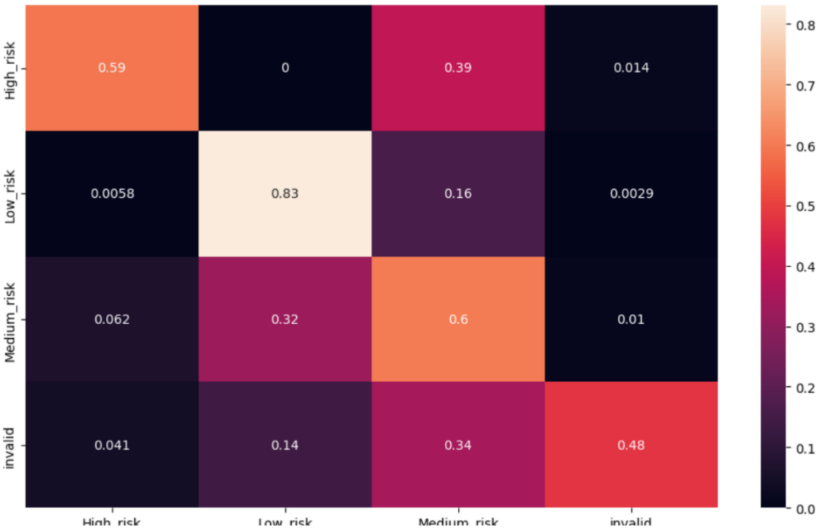


Figure 7: Confusion Matrix for InceptionV3 validation accuracy

For DeiT. I carried out transfer learning, I carried out transfer learning for 9 epochs and then fine-tuned for another 9 epochs. I kept the batch normalization layer frozen during fine-tune. Total time for training and fine-tune is 11+27 minutes. Maximum validation accuracy is 0.6864 occurring at epoch 4 before accuracy begins to drop over the next few epochs

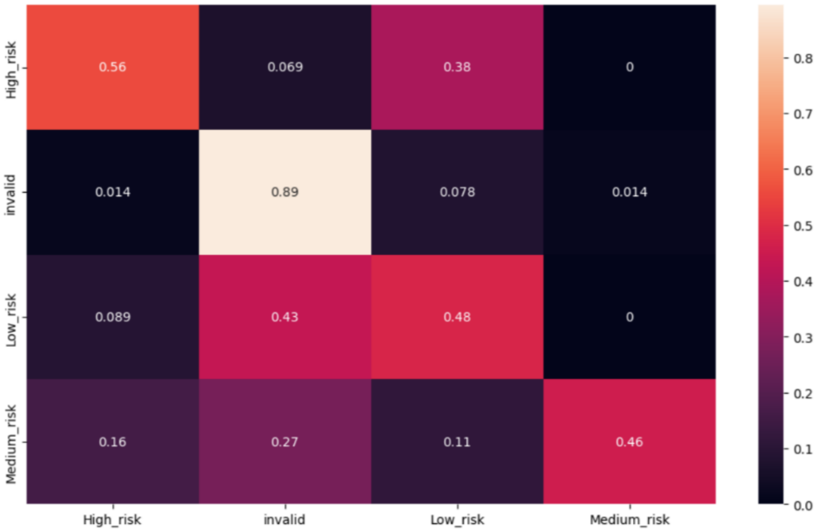


Figure 8: Confusion Matrix for DeiT validation accuracy

Validation accuracies for each class are similar between the InceptionV3 and DeiT. Note that the 'invalid' class has the highest accuracy. This is likely because it is the most contrasting of all four classes. As noted earlier, wide variation between examples were observed in the annotation of the dataset. To achieve greater accuracy, it is necessary to scale the dataset, so the model can achieve better generalization.

5. Conclusion

I have performed experiments to evaluate the accuracy of two prominent computer vision architectures on a novel dataset I created for pedestrian risk classification. InceptionV3 and DeiT trained on GSVLagos1 show similar performance across classes. Taking that into consideration, it would appear that transformer architectures in general perform reasonably well compared with SOTA convolutional architectures for computer vision, image and scene recognition in a Lagos context. I also observed that modification of the normalization layers of a pretrained convolutional network is counter productive. In future work, I will expand the dataset, consider alternative labels that give more clarity, and perform more experiments with other emerging architectures.

References

Adebayo, H., Akinsanya, F., Coker, A., & Jolaawo, O. (2023). Thematic Mapping of the Trend of Road Traffic Crashes in Nigeria: A Tool for Advancing Sustainable Safety of Lives. *KIU Journal Of Humanities*, 163-181. Retrieved from https://kampalajournals.ac.ug/ojs/index.php/kiuhums/article/view/1582

Adeofun, C., & Oyedepo, J. (2011). INTEGRATION OF GIS, GPS, GSM AND REMOTE SENSING,(3GR) FOR ROAD ACCIDENT REPORTING AND MANAGEMENT. *Journal of Agricultural Science and Environment, 11*(2), 111-121.

Alhasoun, F., & González, M. (2019). Streetify: using street view imagery and deep learning for urban streets development. *IEEE International Conference on Big Data (Big Data)*, 2001-2006.

Anguelov, D., Dulong, C., Filip, D., Frueh, C., Lafon, S., Lyon, R., . . . Weaver, J. (2010). Google street view: Capturing the world at street level. *Computer, 43*(6), 32-38.

Asaju, J. A., Olawepo, R. A., & Ojekunle, J. A. (2020). A Study of the Factors That Influence the Rate of Pedestrian Accidents in Lagos Ikorodu Expressway, Nigeria.

Ashish, V., Noam, S., Niki, P., Jakob, U., Llion, J., Aidan, N. G., . . . Illia, P. (2017). Attention is all you need. *NIPS*.

Babatunde, A., Solagberu, R. A., Balogun, I. A., Mustafa, N. A., Ibrahim, M. A., Oludara, A. O., . . . Osuoji, R. I. (2015). Pedestrian Injuries in the Most Densely Populated City in Nigeria—An Epidemic Calling for Control. *Traffic Injury Prevention, 16*(2), 184-189. doi:https://doi.org/10.1080/15389588.2014.921817

Bhatia, Y., Bajpayee, A., Raghuvanshi, D., & Mittal, H. (2019). v2 and Recurrent Neural Network. *Twelfth International Conference on Contemporary Computing (IC3)*, 1-6.

Biljecki, F., & Ito, K. (2021). Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning, 215*. doi:https://doi.org/10.1016/j.landurbplan.2021.104217

Buhari, S. O., Aponjolosun, M. O., Oni, B. G., & Sam, M. W. (2020). Sustainable urban mobility: An approach to urbanization and motorization challenges in Nigeria, a case of Lagos state. *Journal of Sustainable Development of Transport and Logistics, 5*(2).

Chen, L., Lu, Y., Sheng, Q., Ye, Y., Wang, R., & Liu, Y. (2020). Estimating pedestrian volume using Street View images: A large-scale validation test. *Computers, Environment and Urban Systems, 81*. doi:https://doi.org/10.1016/j.compenvurbsys.2020.101481

Chen, L., Lu, Y., Ye, Y., Xiao, Y., & Yang, L. (2022). Examining the association between the built environment and pedestrian volume using street view images. *Cities, 127*. doi:https://doi.org/10.1016/j.cities.2022.103734

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., . . . Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. *Proc. of the IEEE Conference on Computer Visionand Pattern Recognition (CVPR)*.

D, A. Q., Thomas, D. K., & J, J. M. (2015). The walking environment in lima, peru and pedestrian–motor vehicle collisions: an exploratory analysis. *Traffic injury prevention, 16*(3), 314-321.

Debenedetti, E., Sehwag, V., & Mittal, P. (2023). A Light Recipe to Train Robust Vision Transformers. *ArXiv (Cornell University)*. doi:https://doi.org/10.1109/satml54575.2023.00024

Dosovitskiy, A., Beyer, L., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., . . . Sylvain, G. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint. *arXiv:2010.11929*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., & Li, M. (2018). Bag of Tricks for Image Classification with Convolutional Neural Networks. *arXiv:1812.01187v2*.

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hugo, T., Cord, M., Matthijs, D., Francisco, M., Alexandre, S., & Herve, J. e. (2021). Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2010.11929*.

Ibrahim, U., Kumazhege, S., & L'Kama, J. (2023). Prevalence of Automobiles Road Traffic Accidents in Nigeria: The Imperative of Integrating Road Safety Education (RSE) into Automobile Technology Education. *InternationalJournal of Education and National Development*, 10-24.

Kang, J., Körner, M., Wang, Y., Taubenböck, H., & Zhu, X. X. (2018). Building instance classification using street view images,. *ISPRS journal of photogrammetry and remote sensing, 145*, 44-59.

Loshchilov, I., & Hutter, F. (2016). SGDR: stochastic gradient descent with restarts. *CoRR, abs/1608.03983*.

Mehta, A., Kim, D., Allo, N., Odusola, A. O., Malolan, C., & Nwariaku, F. E. (2023). Using parallel geocoding to analyse the spatial characteristics of road traffic injury occurrences across Lagos, Nigeria. *BMJ Global Health*. doi:https://doi.org/10.1136/bmjgh-2023-012315

Najafizadeh, L., & Froehlich, J. E. (2018). A feasibility study of using Google street view and computer vision to track the evolution of urban accessibility. *Proceedings of the 20th international ACM SIGACCESS conference on computers and accessibility*, 340-342.

Nathaniel, S. P. (2020). Modelling urbanization, trade flow, economic growth and energy consumption with regards to the environment in Nigeria. *GeoJournal, 85*(6), 1499-1513.

Neuhold, G., Ollmann, T., Bulo, S. R., & Kontschieder, P. (2017). The mapillary vistas dataset for semantic understanding of street scenes. *IEEE International Conference on Computer Vision (ICCV)*, 5000-5009.

Olusina, J. O., & Ajanaku, W. A. (2017). Spatial analysis of accident spots using weighted severity index (WSI) and density-based clustering algorithm. *Journal of applied sciences and environmental management, 21*(2), 397-403.

Ryder, B., Gahr, B., Egolf, P., Dahlinger, A., & Wortmann, F. (2017). Preventing traffic accidents with in-vehicle decision support systems-The impact of accident hotspot warnings on driver behaviour. *Decision support systems, 99*, 64-74.

Santani, D., Ruiz-Correa, S., & Gatica-Perez, D. (2018). Looking south: Learning urban perception in developing cities. *ACM Transactions on Social Computing, 1*(3), 1-23.

Shyam, & Pranjay, e. a. (2022). Infra Sim-to-Real: An efficient baseline and dataset for Infrastructure based Online Object Detection and Tracking using Domain Adaptation. *IEEE Intelligent Vehicles Symposium (IV)*.

Srivastava, S., Vargas-Munoz, J. E., & Tuia, D. (2019). Understanding urban landuse from the above and ground perspectives: A deep learning, multimodal solution. *Remote sensing of environment, 228*, 129-143.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *arXiv preprint arXiv:1512.00567.*

Tsang, S. (2018). *Medium*. Retrieved from Review: Inception-v3 — 1st Runner Up (Image Classification) in ILSVRC 2015: https://sh-tsang.medium.com/review-inception-v3-

Voorde, T. V., Jacquet, W., & Canters, F. (2011). Mapping form and function in urban areas: An approach based on urban metrics and continuous impervious surface data. *Landscape and Urban Planning, 102*(3), 143-155. doi:https://doi.org/10.1016/j.landurbplan.2011.03.017

Wirayasa, I. K. (2021). Comparison of Convolutional Neural Networks Model Using Different Optimizers for Image Classification. *International Journal of Sciences*.

Wu, Y., & He, K. (2018). Group normalization. *European Conference on Computer Vision (ECCV)*.

Xu, Y., Zhou, B., Jin, S., Xie, X., Chen, Z., Hu, S., & He, N. (2022). A framework for urban land use classification by integrating the spatial context of points of interest and graph convolutional neural network method. *Computers, Environment and Urban Systems, 22*. doi:https://doi.org/10.1016/j.compenvurbsys.2022.101807

Yin, L., Cheng, Q., Wang, Z., & Shao, Z. (2015). 'Big data' for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. *Applied Geography, 63*, 337-345. doi:https://doi.org/10.1016/j.apgeog.2015.07.010.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., . . . Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2633-2642.

Zhang, F., Wu, L., Zhu, D., & Liu, Y. (2019). Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS Journal of Photogrammetry and Remote Sensing, 153*, 48-58.

Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., . . . Huang, Z. (2020). Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems, 81*. doi:https://doi.org/10.1016/j.compenvurbsys.2020.101478

Zhang, Y., Chen, Z., Zheng, X., Chen, N., & Wang, Y. (2021). Extracting the location of flooding events in urban systems and analyzing the semantic risk using social sensing data. *Journal of Hydrology, 603*(C). doi:https://doi.org/10.1016/j.jhydrol.2021.127053