Model Card - Flower classification

1) Model Details

- **Person or organization developing the model:** Developed by Google Research and first introduced by Dosovitskiy et al. The weights were converted from the timm repository by Ross Wightman.
- **Model date:** Released in 2020.
- **Model version:** Patch 16
- **Model type:** Vision Transformer (ViT), a Transformer encoder model designed for image recognition tasks.
- **Training algorithms and parameters:** The ViT model is trained by using supervised learning on ImageNet-21k. It uses the Transformer architecture which was originally used for NLP tasks and processes images as a sequence of patches.
- **Paper:** "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Dosovitskiy et al.
- **Citation:** See BibTeX citation at the end of the card.
- **License:** Apache 2.0 License.
- **Where to send questions or comments about the model:** Contact the Hugging Face community, where the model is hosted.

2) Intended Use

- **Primary intended uses**

This model is intended for image classification tasks. It has been pre-trained on a large dataset (ImageNet-21k), which makes it a flexible base model for various downstream tasks.

- **Primary intended users**

Researchers, Developers, and any type of machine learning practitioners who want to use a Transformer-based model for image recognition tasks.

- **Out-of-scope use cases**

This model is not fine-tuned for tasks outside image classification. The pre-trained weights are not designed for image generation or object detection, and applying the model to other domains without appropriate fine-tuning may lead to suboptimal results.

3) Factors

- **Relevant factors**

Hardware factors of camera and lens type as well as environmental factors of lighting and humidity to prevent blurred or distorted images can affect the effectiveness of the model in extracting important features.

- **Evaluation factors**

The performance varies with the dataset size, diversity, and the distribution of the existing training images across the categories.

4) Metrics

- **Model performance measures**

 Standard image classification metrics such as accuracy, top-1, and top-5 accuracy were used for evaluation in the paper.

- **Decision thresholds**

 Not applicable as this model gives scores for each class, which can be turned into probabilities using a softmax function. These probabilities show how likely it is that an image belongs to each class.

- **Variation approaches**

 Larger model variants (such as ViT-Large) and higher input resolutions (384x384) can improve performance. Fine-tuning also leads to better results for domain-specific tasks.

5) Evaluation Data

- **Datasets**

 The ViT model was pre-trained on ImageNet-21k, a large-scale dataset with 14 million images and 21,843 categories (test data split).

- **Motivation**

 ImageNet-21k is one of the largest and most diverse image datasets available, making it suitable for pre-training models that can generalize well to various downstream tasks.

- **Preprocessing**

 Images are resized/rescaled to the same resolution (224x224) and normalized across the RGB channels with mean (0.5, 0.5, 0.5) and standard deviation (0.5, 0.5, 0.5). The exact details about the preprocessing can be found here.

6) Training Data

Training data was sourced from the same dataset as the evaluation data and was preprocessed in the same way (training data split).

7) Quantitative Analyses

- **Unitary results:** The ViT model achieved strong performance on several image classification benchmarks. For specific accuracy values, refer to tables 2 and 5 in the original paper.

8) Ethical Considerations

- **Bias and Fairness**

 Since the model is trained on ImageNet, which contains images from various online sources, there may be inherent biases in the data, such as underrepresentation of certain demographics or overrepresentation of Western cultural elements. These biases may affect the model's performance in real-world applications.

- **Potential Misuse**

 This model is intended for image classification tasks and should not be used in critical applications such as medical diagnosis, surveillance, or applications that can negatively impact individuals' privacy or security without proper oversight and testing.

9) Caveats and Recommendations

- This model performs best when fine-tuned on a task-specific dataset.
- Larger input resolutions (e.g., 384x384) can improve classification accuracy but come at the cost of increased computational resources.

10) BibTeX Entry and Citation Information

```
@misc{wu2020visual,
      title={Visual Transformers: Token-based Image Representation and Processing for Computer Vision},
      author={Bichen Wu and Chenfeng Xu and Xiaoliang Dai and Alvin Wan and Peizhao Zhang and Zhicheng Yan and Masayoshi Tomizuka and Joseph Gonzalez and Kurt Keutzer and Peter Vajda},
      year={2020},
      eprint={2006.03677},
      archivePrefix={arXiv},
      primaryClass={cs.CV}
}
@inproceedings{deng2009imagenet,
  title={Imagenet: A large-scale hierarchical image database},
  author={Deng, Jia and Dong, Wei and Socher, Richard and Li, Li-Jia and Li, Kai and Fei-Fei, Li},
  booktitle={2009 IEEE conference on computer vision and pattern recognition},
  pages={248--255},
  year={2009},
  organization={Ieee}
}
```