CS 2241 Project Proposal – Image Compression for Classification Tasks

Sezim Yertanatov, Jared Ni, Dennis Du

Question to answer

Image classification models, particularly deep neural networks, often require large datasets of images for effective training. This poses significant challenges in terms of storage space for the training data and bandwidth for transmitting images during both training and inference. During model inference, especially in client-server architectures, users may need to send images to a remote server for classification. Reducing the size of these images through compression can alleviate these challenges.

Standard lossy compression techniques focus on reducing the size of the image while ensuring that the reconstructed image is of sufficiently high quality that it appears acceptably similar to the original image to a human user. However, in the context of image classification, the main result we care about is whether the model outputs the correct class, regardless of how lossy the compression algorithm is or how poor the quality of the compressed image is. In other words, the goal is to compress the image as much as possible, no matter how bad the reconstructed version looks, as long as the classifier can output the correct class with an accuracy that is comparable to a classifier that operates in a standard manner. The idea then is to create a compression algorithm that preserves the features of the image that are most important to an accurate classification and removes all extraneous information.

Therefore, our central research question is: Can we develop a lossy image compression algorithm, specifically tailored for image classification tasks, that achieves significantly higher compression ratios than standard methods while maintaining classification accuracy comparable to a model operating on uncompressed images? In essence, the goal is to maximize image compression without sacrificing classification performance. We hypothesize that by focusing on preserving classification-relevant features rather than visual fidelity, we can achieve significantly higher compression ratios than are possible with traditional methods. We aim to move beyond perceptual quality as the primary metric and instead optimize for classification quality.

To elaborate, we aim to investigate the following sub-questions along the way:

- Trade-off Analysis: What is the precise relationship between compression ratio and classification accuracy for classification-aware compression? How does this trade-off

compare to that of standard compression methods like JPEG? Can we quantify this relationship?

- Feature Preservation: What are the key features or image regions that are most crucial for accurate classification by a given model or class of models? Can we design a compression scheme that explicitly prioritizes the preservation of these features while aggressively compressing less relevant information?
- Algorithmic Efficiency: Can we develop compression and decompression algorithms that are computationally efficient, thus ensuring that the benefits of reduced storage and transmission are not offset by excessive computational overhead? This is particularly important for inference, where latency is often a critical factor.
- Robustness: How robust is our proposed compression method to variations in image content (does it overfit)? Will it perform consistently across different datasets?

Approach

There are several directions we can pursue, ranging from very basic transformations of the image to much more complex compression algorithms. These directions should hopefully prove to be reasonably orthogonal to each other so that multiple methods can be applied to one image to achieve the best tradeoff between compression rate and prediction quality.

Basic ideas include blurring the image somewhat, which can easily reduce the size of the image without obscuring what the image represents. For example, we could take 3×3 blocks of the image and average them into a single pixel. In order to add additional expressiveness, we might indicate how similar or different the pixels in each original 3×3 block were. That way, we can distinguish between a 3×3 block whose average is a green pixel simply because everything was originally a green pixel and a block that originally contained significantly more information. This could be useful in helping us preserve information, such as whether that block was originally quite uninteresting or whether it represented a highly textured part of the image or the boundary of an object.

A more complex approach would be to identify regions of interest (ROIs) that are likely to be important in determining the image class and encode those regions with a higher quality compression. Very lossy compression schemes, which can achieve high compression rates, can then be used for the rest of the image. It may even make sense to completely remove certain unimportant parts of the image. These ROIs could be identified by training an image segmentation model to compute a saliency map indicating how important each pixel of the image is, or we could design a heuristic for identify important regions (such as regions that

contain more variation in the pixel values are appear to be surrounded by some sort of boundary).

If possible, we can consider training the model directly on the compressed image files instead of first uncompressing them, which would reduce the computational cost of running the model. For this, we would need to identify a compression scheme that inherently preserves the structure of the image in a way that is learnable for a classification model.

The plan then is to write code to implement these various ideas and then run experiments that involve training and evaluating an image classifier on our compressed images to determine whether we can get a compression rate while achieving competitive levels of accuracy from the model. We expect this will be an iterative process in which we progressively improve our algorithm based on the results of our experiments.

If possible, we wish to further explore a balanced compression scheme that aggressively compresses non-essential features for classification while preserving some visual quality for features deemed important in the classification. It will help us to better interpret the decisions made by the model. To accomplish this, we will need to identify an algorithm for quickly identifying features of the image that are important. This algorithm should ideally strike a balance between computational efficiency and accuracy in pinpointing classification-relevant information. We will investigate methods that leverage both image-intrinsic properties, such as edges and textures, and model-specific information, like activation patterns within a pre-trained classifier. The goal is to develop a reliable and fast method for distinguishing between image regions that can be heavily compressed and those that require more careful preservation.

Resources

We will likely need access to some GPUs in order to train an image classification model on our proposed compressed images, so that we can evaluate whether our compression scheme is able to preserve prediction accuracy. We anticipate that Google Colab should be sufficient.

Backup plans

Should our primary approach prove unviable, we will consider an alternative method still involving image compression based on regions of interest, without employing any classifiers. Instead, our goal would be to maximize image compression by applying varying compression rates to different image segments (ex: ), ensuring that the perceived quality remains indistinguishable from the original to the human eye. We would likely conduct a user study to objectively evaluate the effectiveness and visual fidelity of our compression technique.

Related work

- Compact Representation for Image Classification: To Choose or to Compress? https://www.cv-foundation.org/openaccess/content_cvpr_2014/html/Zhang_Compact_Representation_for_2014_CVPR_paper.html
- High-dimensional signature compression for large-scale image classification https://ieeexplore.ieee.org/abstract/document/5995504
- Effects of JPEG compression on image classification https://www.tandfonline.com/doi/abs/10.1080/01431160210142842
- Compressed Learning for Image Classification: A Deep Neural Network Approach https://www.sciencedirect.com/science/article/abs/pii/S1570865918300024
- Compression Helps Deep Learning in Image Classification https://www.mdpi.com/1099-4300/23/7/881
- ROI-based Deep Image Compression with Swin Transformers https://arxiv.org/abs/2305.07783
- Variable Rate ROI Image Compression Optimized for Visual Quality https://openaccess.thecvf.com/content/CVPR2021W/CLIC/papers/Ma_Variable_Rate_ROI_Image_Compression_Optimized_for_Visual_Quality_CVPRW_2021_paper.pdf
- The Image Calculator: 10x Faster Image-AI Inference by Replacing JPEG with Self-designing Storage Format https://www.dropbox.com/scl/fi/dxpt42wip2esjp0i9vjoc/image_calculator_sigmod2024_crc_final.pdf?rlkey=mhxdg4dka1mhn3x5wne2it29x&e=1&dl=0