Analyzing Finetuned Vision Models for Mixtec Codex Interpretation

Anonymous ACL submission

Abstract

Throughout history, pictorial record-keeping has been used to document events, stories, and concepts. Examples include the Foggini-Mestikawi Cave, the Bayeaux Tapestry, and the Tzolk'in Maya Calendar. The pre-Columbian Mixtec society also recorded many works through graphical media called codices that depict both stories and real events. Mixtec codices are unique because the depicted scenes are highly structured within and across documents. As a first effort toward translation, we created two binary classification tasks: gender and pose. The composition of figures within a codex is essential for understanding the codex's narrative. We labeled a dataset with around 1300 figures drawn from three codices of varying qualities. We finetuned the VGG-16 and ViT-16 models, measured their performance, and compared learned features with expert opinions found in literature. The results show that when finetuned, both VGG and ViT perform well, with the transformer-based architecture (ViT) outperforming the CNN-based architecture (VGG) at higher learning rates. We are releasing this work to allow collaboration with the Mixtec community and domain scientists.

1 Introduction

Vast amounts of historical and cultural documents are encoded in pictographic systems (Sampson, 2015). Representations such as Egyptian hieroglyphics use pictorial representations corresponding to words to express concepts. Other pictorial systems that display scenes that evoke a known narrative have also been used throughout the world. Rules govern the depiction of years, dates, names, class, ceremonies, and gender (Jansen, 1988). **The implicit grammatical rules can contribute to a deterministic interpretation of these ancient narratives.** Mixtec codices are highly structured and have fairly rigid conventions for the representation of people (Boone, 2000), such as loincloths on men and skirts on women. Consequently, the depiction of persons in these codices follows consistent patterns. Unfortunately, due to the ravages of time and conflict, only a few of these codices are available at present. However, computational analyses of the codices and their underlying structures may help researchers better understand the remaining works. In this paper, we explore how models such as VGG-16 (Simonyan and Zisserman, 2015) and ViT-16 (Dosovitskiy et al., 2021) perform when used to classify these low-resource patterns and understand the features they find important in this task.

2 Mixtec Codices

The researchers labeled data from three popular sources: The Codices Vindobonensis Mexicanus I (Lehmann and Smital, 1929; Unbekannt, 1449), Selden (Caso, 1964; Bakewell and Hamann, 2023), and Zouche-Nuttall (Nuttall, 1902; Forstmann, 2023). Codex Vindobonensis Mexicanus I describes both the mythological and historical founding of the first Mixtec kingdoms, Codex Selden follows the founding of the kingdom of *Jaltepec* and its ruler, *Lady 6 Monkey*, and Codex Zouche-Nuttall primarily illustrates the life and conquests of *Lord 8 Deer Jaguar Claw*, but also details the histories of his ancestors. While several other Mixtec codices are extant, their condition has been significantly degraded and is not amenable to our current machine-learning pipeline. Each codex is made of deerskin folios, and each folio comprises two pages. The Codex Vindobonensis Mexicanus I contains 65 pages, Selden 20 pages, and the Zouche-Nuttall facsimile edition 40 pages. We have chosen to use the Zouche-Nuttall facsimile edition over the complete 84-page edition because of its restored quality and high-quality scans available.

Codex	Total	Gender		Pose		Quality		
		Man	Woman	Standing	Not Standing	a	b	c
Nuttall	264	256	8	101	163	263	1	0
Selden	307	74	233	32	275	254	46	7
Vindobonensis Mexicanus I	714	573	141	253	461	569	123	22
Totals	*1285*	*903*	*382*	*386*	*899*	*1086*	*170*	*29*

Table 1: The counts of the **a**, **b**, and **c** labeled data items and the number of Man and Woman labels of each quality.

2.1 Data Processing

We used the Segment Anything Model (SAM) (Kirillov et al., 2023) from Facebook AI Research to extract individual figures from the three source codices[1]. Figures are representations of people or gods in Mixtec mythology and are composed of different outfits, tools, and positions. Their names are represented by icons placed near their position on a page. Each figure's file was then named according to the page it was found on, its quality as either `a`, `b`, or `c`, and its order within the page. An `a` quality rating indicated the entire figure was intact, regardless of minor blemishes or cracking, and could be classified by a human annotator as man or woman, standing or not. A `b` rating means that while the previous characteristics of the figure could be determined, significant portions of the figures were missing or damaged. The `c` rated figures were missing most of the definable characteristics humans could use to classify the sample.

2.2 Labeling Procedures

Once the figures had been segmented and graded, we added additional classification labels to each figure. Literature describes representations of gender and poses in Mixtec codices to guide our classifications (Boone, 2000; Smith, 1973; Jansen, 1988; Williams, 2013; Lopez, 2021). We propose two binary classification tasks: Gender (`man`/`woman`) and Pose (`standing`/`not standing`). These two categories represent meaningful distinctions in Mixtec codices and allow for the exploration of deeper, more complex investigations into the structure of these documents. Two team members tagged the images for both categories independently and then verified the results with each other using the process of inter-rater reliability (Hallgren, 2012).

[1]Each codex we used were high-quality and designated as free for non-commercial use or provided by national libraries

2.3 Dataset Statistics

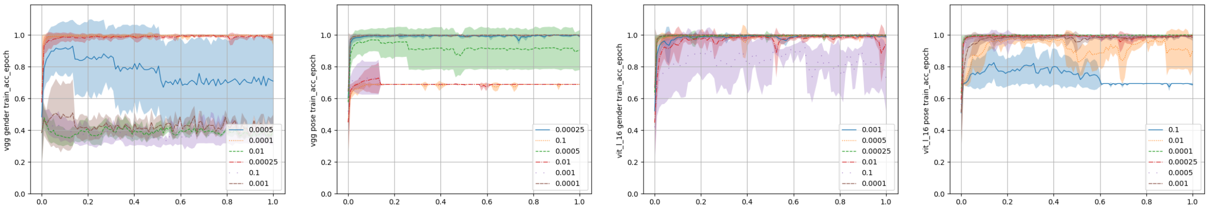
Codex Vindobonensis Mexicanus I represents the largest proportion of the 1285 figures with 714, Codex Selden has 307, and Codex Zouche Nuttall is the smallest with 264. Codex Vindobonensis Mexicanus I contains 573 men and 141 women, Selden 74 men and 233 women, and Zouche-Nuttall 256 men and 8 women. This imbalance in each dataset can be attributed to the fact that each is centered on a different figure. The Pose category follows a similar proportion split, however, a `not standing` position outweighs `standing`, for each codex. The reason for this is unclear, although given the number of ceremonies that each codex describes, which entails a seated or kneeling position, this balance intuitively makes sense. The quality of the figures is largely dominated by the `a` classification with 1086 figures, followed distantly by `b` at 170 figures, and `c` comprising only 29 figures. Of these totals, the Zouche-Nuttall accounts for 263 `a`, only one `b` designation, and zero `c` figures. The Selden contains 254 `a` classifications, 46 marked with `b`, and 7 `c`. Finally the Vindobonensis Mexicanus I has 568 `a` figures, 123 `b`, and 22 `c`. Given the small number of `c` samples across all three codices, we use all three categories in the model training and testing pipelines. These numbers can be viewed in Table 1.

3 Experiment

We describe the preprocessing, finetuning, and execution steps of this pipeline. We explore the hyperparameter space for each model first to find the optimal configuration to use during execution.

3.1 Preprocessing

For our model pipeline preprocessing, the figures are moved to tensors and then normalized to 224x224 pixels. We bias the loss function by weighting each class in the loss function by its in-



(a) VGG-16 - Gender (b) VGG-16 - Pose (c) ViT-16 - Gender (d) ViT-16 - Pose

Figure 1: Training accuracy vs. percentage to completion for a given run. Graphs execution across learning rates. Smaller learning rates converged faster across all runs, while some larger learning rates failed to converge.

verse. Finally, due to the overall limited number of figures, and to prevent overfitting, we augmented the entire dataset by using random flips and blocking to increase the number of samples for training. The dataset is then split into training, testing, and validation sets, 60%, 20%, and 20% respectively. We set aside eight reference images to monitor which features of gender and pose are prevalent in activation and attention maps throughout training.

3.2 Models

Both CNNs and transformers are used in image classification (Lu et al., 2021). We fine-tuned popular vision models VGG-16 and ViT-16 to perform classification tasks and improve computational efficiency. We first imported the models and their pre-trained weights from the PyTorch library. We then unfroze the last four layers and heads of each model for training, as they are responsible for learning complex features specific to our classification tasks (Olah et al., 2017). Finally, the fully connected layer of each model was replaced by one matching our binary classification task.

3.2.1 Hyperparameters

Next, we explored the number of epochs, batch size, and learning rate of each of our models. We experimented with different batch sizes, ranging from 32 to 128, and opted for an average value of 64 as no particular size significantly outperformed the others. Once we finalized the hyperparameter space, we selected the loss function and optimizer according to the best practices associated with our pretrained models, VGG and ViT.

3.3 Execution

Model training and inference were performed on an Nvidia A100 on the HiPerGator cluster using PyTorch 2.1 and CUDA 11. For both VGG and ViT, each run took up to 25 minutes to complete. Before the first and after the last epoch of training, an activation map for VGG and an attention map for ViT is output for each reference image. We then ran the testing phase of the model pipeline using the optimal hyperparameters found during training and validation. Testing is run 30 times for each model and classification task and the performance scores are averaged to measure the reliability of the model.

4 Results

For each training and validation run, we collected metrics such as accuracy, F1, recall, loss, and precision. The accuracy results from training for varying levels of learning rates are presented in Figure 1 for both VGG and ViT and both classification conditions. ViT performs consistently higher than VGG for these different learning rates, however, both returned strong results for each metric. The testing results for both ViT and VGG were high with a small standard deviation, around 98% and 1% standard deviation for both (see Table 2). Hyperparameter investigations revealed that the accuracy for training and validation converged around 100 epochs and the ideal learning rate was 0.00025.

Model	Task	Test Accuracy $\pm$ (stddev)
VGG-16	Gender	0.978 $\pm$ (0.009)
VGG-16	Pose	0.978 $\pm$ (0.01)
ViT-16	Gender	0.977 $\pm$ (0.009)
ViT-16	Pose	0.974 $\pm$ (0.009)

Table 2: Testing accuracy and their standard deviations for VGG-16 and ViT-16.

5 Discussion

The purpose of the experiments has been to explore the aforementioned research questions, namely: *Can CNN and transformer-based models be fine-tuned to classify figures from a Mixtec Codices dataset?* and *Does the model identify the same*

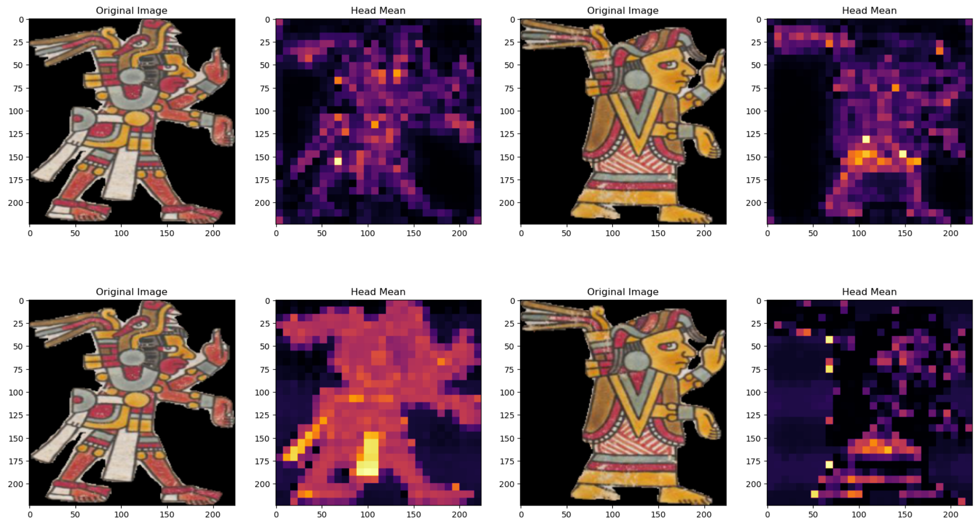


Figure 2: ViT-16 Mean Attention Maps for Man and Woman. The top row shows original images and attention maps extracted before the first epoch of training for `man` (left), and `woman` (right). The bottom row shows original images and attention maps after the last epoch of training. The model shows increased attention in the loincloth area for `man`, and the skirt area for `woman`, which follows expert opinion.

features experts do? To answer the first question, we analyze and compare the performances of both the pretrained ViT and VGG models. Based on accuracy alone, there is not much to distinguish the two. Both models achieve great results across training, validation, and testing phases when using an appropriate learning rate. Smaller learning rates require more epochs to converge, as the steps are smaller, but are less likely to miss a minimum loss. On the other hand, Larger learning rates require fewer epochs, but may not converge. As we can see in Figure 1, ViT converges for almost all learning rates, and so could be used in environments where compute resources are lacking.

5.1 Class Activation Maps

We assigned reference images for each class (`man`/`woman`, and `standing`/`not standing`) to understand which features each model learned, as well as to compare these learned features to those highlighted by experts. During training, we generated visualizations of activation and attention per pixel to view how the models learned important features over time. In the left image in Figure 2, the ViT model assigned higher attention to areas corresponding to loincloths on men. On the right, ViT shows increased attention to the poncho area on a woman. These are both features noted by domain experts (Boone, 2000).

6 Summary

In this paper, we presented a low-resource dataset of figures from three Mixtec codices: Zouche-Nuttall, Selden, and Vindobonensis Mexicanus I. We extracted the figures using the Segment Anything Model from Facebook AI Research and labeled them according to gender and pose, two critical features used to understand Mixtec codices. Using this novel dataset, we finetuned the last few layers of CNN and transformer-based foundational models, VGG-16 and ViT-16 respectively, to classify figures as either man or woman and standing or not standing. We found that both models perform exceptionally well with this dataset, but that ViT-16 may be more reliable for varying learning rates. We confirmed that the models are learning the features said to be relevant by experts using class activation maps and targeted blocking of said features.

Given that these models can reliably classify figures from a low-resource dataset, this research opens the door for further processing and analysis of Mixtec Codices. The codices themselves are highly structured and carry a narrative woven through each scene. Finetuned state-of-the-art models could be combined to classify segmented figures within a scene, as well as classify the relationship between figures. These relationships would then be used to extract the narrative from a codex, as defined by subject matter experts.

7 Limitations

The Mixtec civilization produced many of the available codices, however, conquest and the passage of time have left us with only a few remaining high-quality samples (Boone, 2000). Fortunately, many of the surviving codices still contain examples of scenes and can be used to build a digitized corpus for machine processing. We chose popular models to demonstrate our method. We believe other architectures would have similar results. The quality results in both models show a specialized architecture is not required for accuracy. We have not yet explored more environmentally efficient models. Both models we adopt use pretrained classifiers, each trained on data not specific to our domain. The models inherit all biases previously encoded in the model. We have not investigated how these biases may affect downstream tasks. The finetuned models generated few errors in our investigation, however, we are unaware of how these biases may result in unintended effects.

We selected classification tasks that are well understood within the Mixtec research community, namely: `man` and `woman`, and `standing` and `not standing`. Many experts disagree on the interpretation of scenes across codices. For instance, some early 20th-century scholars have stated cannibalism and human sacrifice are depicted within the codices (Pohl, 1994), while others contend that these scenes should be understood as metaphorical interpretations (Lopez, 2021; Lopez and Collver, 2022). This work is an initial investigation into Mixtec and low-resource, semasiographic languages. We are prohibited from deeper explorations until we align our research direction with present communal, cultural, and anthropological needs. Support from Mixtec domain experts and native Mixtec speakers is essential for continued development.

References

Liza Bakewell and Byron Hamann. 2023. Codex selden.

Elizabeth Hill Boone. 2000. *Stories in red and black: Pictorial histories of the Aztecs and Mixtecs*. University of Texas Press.

Alfonso Caso. 1964. Códice selden. *Sociedad Mexicana de Antropología, Mexico City*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Sam Forstmann. 2023. Codex zouche-nuttall.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Maarten Jansen. 1988. The art of writing in ancient mexico: an ethno-iconological perspective. *Visible Religion*, 6:86–113.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything.

Walter Lehmann and Ottokar Smital. 1929. Codex vindobonensis mexic. 1. *Faksimileausgabe der mexikanischen Bilderhandschrift der Nationalbibliothek in Wien. Verlag von Anton Schroll & Co, Vienna*.

Felicia Rhapsody Lopez. 2021. *Women, Childbirth, and the Sticky Tamales: Nahua Rhetoric and Worldview in the Glyphic Codex Borgia*, chapter 4. University of New Mexico Press.

Felicia Rhapsody Lopez and Jordan Collver. 2022. How to read an aztec comic: Indigenous knowledge, mothers' bodies, and tamales in the pot.

Kangrui Lu, Yuanrun Xu, and Yige Yang. 2021. Comparison of the potential between transformer and cnn in image classification. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, pages 1–6. VDE.

Zelia Maria Magdalena Nuttall. 1902. Facsimile.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill*, 2(11):e7.

John Pohl. 1994. *The politics of symbolism in the Mixtec codices*. Vanderbilt University.

Geoffrey Sampson. 2015. Writing systems: methods for recording language. In *The Routledge handbook of linguistics*, pages 47–61. Routledge.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Mary Elizabeth Smith. 1973. *Picture writing from ancient southern Mexico; Mixtec place signs and maps.*, [1st ed.]. edition. The Civilization of the American Indian series, 124. University of Oklahoma Press, Norman.

Unbekannt. 1449. Bilderhandschrift: Sog. codex mexicanus bzw. codex yuta tnoho.

Robert Lloyd Williams. 2013. *The Complete Codex Zouche-Nuttall*. University of Texas Press.