Egocentric Vision - transfer and multi-modal learning

Apostol Laurentiu
Politecnico di Torino
S317518@studenti.polito.it

Zerbini Ilaria
Politecnico di Torino
S300790@studenti.polito.it

Shahin Saeidpour
Politecnico di Torino
S290004@studenti.polito.it

Abstract

This report presents a possible technique, based on the application of transfer learning and the combination of different data modalities, in order to classify videos in the field of egocentric action recognition.
First, the transfer learning capabilities of the I3D model are shown by applying it on two datasets, Epic Kitchens and ActionNet, with different classification architectures.
Then, focusing on ActionNet, the RGB features extracted with I3D are combined with EMG in order to study the capabilities of multi-modal learning in this field, comparing the results with those achieved by using only the RGB or EMG data. The chosen architectures for this purpose are based on four different cores: multi-layer perceptron, LSTM, CNN and Transformers.

1. Introduction

In recent years, due to the growing amount of accessible data, there has been a strong interest in the field of egocentric vision, which involves analyzing and understanding visual data from a first-person perspective. Egocentric vision has numerous applications, such as human-computer interaction, healthcare monitoring, and surveillance systems. Therefore, the ability to recognize and understand actions performed by individuals in egocentric videos is becoming more and more required.
This report focuses on the problem of egocentric action recognition using both separate modalities, specifically RGB (visual) and EMG (electromyography), and then trying a multimodal approach.
The first aim of the project consists in showing the potential of transfer learning in this field.
Transfer learning has proven to be a powerful technique in various computer vision tasks, allowing researchers to exploit models pre-trained on large-scale datasets and to apply them to similar contexts, requiring only a fraction of the computational power and data, compared to training from scratch, while still obtaining a proficient model of similar performance.
In this study, the transfer learning capabilities of the I3D (Inflated 3D ConvNet) model are investigated, by applying it to subsets of two benchmarks datasets, namely Epic Kitchens and ActionNet. By exploiting different classification architectures, such as multi-layer perceptron, LSTM (Long Short-Term Memory) and Transformer, the effect of this technique is tested using various sampling techniques, in order to fully understand its capabilities.
Then, focusing on the ActionNet dataset, two different data modalities are explored (EMG and RGB), through the prospective of multimodal learning; afterwards, the results obtained by combining the RGB features extracted through I3D with EMG data, which captures muscle activity signals, are compared with the ones obtained using solely RGB or EMG data, in order to asses the effectiveness of this technique.
In order to perform this last task, the three architectures above are once again employed, with the addition of a Convolutional Neural Networks (CNN), both for training and for feature extraction.
Each architecture is assessed in terms of its ability to exploit the multimodal information and capture the complex temporal patterns distinctive of video data.

2. Related Work

The field of egocentric vision and its applications have gained significant attention in recent years. Several studies have explored different techniques and approaches to address this task, and thanks to the availability of large-scale datasets and advancements in deep learning techniques, remarkable achievements have been attained.
One of these progresses is the introduction of **Transfer learning**, which has emerged as a powerful technique in computer vision tasks, enabling models to leverage pre-trained knowledge from large-scale datasets and apply them to similar domains.
In the context of egocentric action recognition, transfer learning has shown promising results in improving classification performance. One notable work is the application of the I3D (Inflated 3D ConvNet) model by Carreira and Zis-

serman (2017) [1]. They demonstrated the effectiveness of transfer learning both by passing from a massive 2D CNN model, trained on images, to a 3D CNN, better suited for video data but much harder to train from scratch, and by applying their I3D model on various benchmark datasets, achieving state-of-the-art results.
This study motivated the first half of the present work, which aims to dig deeper into the capabilities of I3D, mixing transfer learning with fine tuning and freezing of the I3D layers.
Damen et al. [2] in 2018 gave a huge contribution to the egocentric classification field, by introducing Epic Kitchen, one of the first large-scale dataset of this type of videos, to which they have added more data and a robust data collection pipeline in 2021 [3].
Exploiting this new source of data, a possible classification approach is described by Kazakos at al. [6] in 2021, where the *"Audio-Visual Temporal Binding"* technique is introduced. In this new procedure three different modalities (RGB, Flow and Audio) are combined and the technique used can be summarized as *"the combination of modalities within a range of temporal offsets"*.
The present work, for its second half, explores this approach, combining visual (RGB) and electromyography (EMG) data, on the basis of their complementary nature (visual and muscle activity signals), aiming to enhance the performance of action recognition in egocentric videos.
This is made possible thanks to the introduction, by Delpreto, Chaoliu et al. in 2022, of a new multimodal dataset for action recognition, in the form of ActionNet [4].
In summary, the field of egocentric action recognition has witnessed significant progress through the exploration of transfer learning techniques and the integration of multimodal data. The present work builds upon these advancements and aims to contribute to the understanding of egocentric vision by investigating the transfer learning capabilities of the I3D model applied to Epic Kitchens and evaluating the effectiveness of multimodal learning using RGB and EMG data on the ActionNet dataset.

3. Datasets

In order to test and ascertain the capabilities of the approach presented, two benchmark datasets have been chosen: Epic Kitchens and ActionNet.
As mentioned in the section above, Epic Kitchen is a large-scale egocentric video dataset, in which 32 participants have recorded themselves performing various kitchen tasks, using first-person perspective.
It is important to highlight that no subject followed a script, hence not only the number of different recorded actions is conspicuous but there is also a significant variety in how these actions have been carried out.
In particular this feature could make Epic Kitchen a valid choice for training models with good generalization potential (and hence good transfer learning potential), since it allows for the extrapolation of the generic conception of an action, which can be better adapted to other contexts, decreasing the risk of overfitting.

DATASET	Rec	#Gen_Act	#Spec_Act
EpicKitchens	P01,P08,P22	8	25
ActionNet	S04	12	

Table 1. In the table some summary statistics about the dataset used are reported; in particular *Rec* indicates the specific subsets of the dataset taken into account for the work, *#Gen_act* is the number of generic class action while *#Spec_Act* is number of specific action in the dataset. Moreover, in the present, split D1 represents the recording P08, D2 the P08 and D3 P22.

Instead, for what concerns ActionSense, called ActionNet in the present, it is a multimodal dataset which contains the data registered by wearable sensors during kitchen-related activities. Its aim is to serve as a benchmark dataset for studies about human interactions with objects and the environment, during activities of daily living, with the intention of creating more capable and collaborative robot assistants.
The sensors capture motion, force, and attention information; they include eye tracking with a first-person camera, forearm muscle activity sensors, a body-tracking system using 17 inertial sensors, finger-tracking gloves, and custom tactile sensors on the hands.

3.1. Preprocessing

The preprocessing of the data divides itself into different streams, depending on the dataset and on the modalities involved.
Due to the lack of computational power, the work has been done on 3 different subsets of Epic Kitchen, called in the present D1, D2 and D3, and on one subset of ActionNet, called S04.
The first preprocessing step for both Epic Kitchens and ActionNet consisted in translating the action descriptions into action categories, which would be the target of the training process. The Epic Kitchen subset used was already divided into 8 action categories, while for ActionNet 12 different categories were chosen.
Each one of the Epic Kitchen's subsets were comprised of the videos split into frames, so no preprocessing was needed in this regard.
Each sample, made of a certain number of frames, was split into clips and then different sampling methods and sampling parameters were tested.

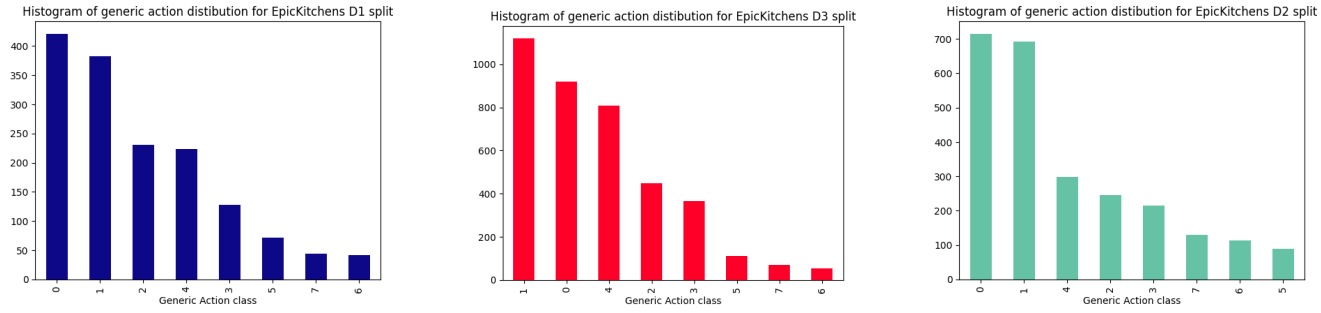

Figure 1. Distribution of generic action classes in the various splits of Epic Kitchens used

Each sampled frame was re-scaled, random cropped, gray-scale transformed and normalized.
Finally, every clip was then passed through the I3D pre-trained model and the features from its last layer, before the fully connected layer, were extracted and saved, in order to be used as input to the classification algorithms.

Regarding the ActionNet subset used, it was comprised of a fully-fledged video, so the first step was dividing it into frames. In order to save space and speed up the process, the original frame-rate was maintained while the down-scaling was applied already at this phase. After this initial step, the same Epic Kitchen RGB pipeline was applied to ActionNet.

For the EMG modality, comprised of 8-channel data from each arm (so 16 channels in total), there were two distinct preprocessing pipelines, both based on the wave nature of EMG data, used in order to test out two different types of architectures:

- Each channel was rectified by taking the absolute value of each EMG reading, and then a low-pass filter with cutoff frequency 5 Hz was applied. All 8 channels from each armband were then jointly normalized and shifted to the range [0, 1] using a min-max scaler.
  Then, the processed data from across all 16 forearm channels were summed together and used as an indicator of overall activation. After this feature-engineering step, the new processed wave segments, each corresponding to an individual action, was normalized with respect to its length, either by 0-padding or by truncating. The reference length value was chosen after inspecting the length distribution of each action's EMG data.
- The same type of length normalization described above was first applied to each individual channel of every action's EMG readings; afterwards, each data flow from each channel was individually transformed into a spectogram (2D matrix). Finally all the spectograms were grouped together, forming a distinct 3D tensor of shape [16 channels x frequencies_number x time_steps] for each action.
  After these preprocessing steps, the spectograms were still split into clips (dividing them into equal parts along the time axis of the 3D matrices). This was done in order to simplify the multimodal integration.

Finally, for what concerns the multimodal approach, the spectogram EMG preprocessing was followed. The raw spectograms were, however, not used. Instead, a CNN classifier was trained on the EMG spectogram data and the features from the last layer, before applying the fully connected classification layer, were extracted. These features were then added to the features extracted with I3D, clip by clip, in order to create a "fused" set of features.

3.2. Data exploration and visualization

Clustering techniques were applied both for data exploration and for testing their capabilities in relation to egocentric vision tasks. These methods are not only characterized by an easy and straightforward pipeline but they are also unsupervised and allow for the usage of 2D visualizations of the results.
After the RGB preprocessing steps and features extraction, the initial data was transformed into a tensor of dimensions [Number_of_actions x Number_of_clips x Number_of_I3D_features].
For each action, only the central clip was considered as input for the clustering pipeline. Since the number of I3D features is quite large (1024 features per clip), two different procedures for features reduction have been tested, and their results have been compared: PCA (Principal Component Analysis) and TSNE (T-distributed Stochastic Neighbor Embedding).
The target embedding space dimension was set to 2 for both techniques and the 2D vectors per action, resulting from both, were used, separately, as inputs for the K-Means clustering algorithm.
For each of the two sets of 2D vectors the K target number of clusters was then assessed by means of a classic knee-graph analysis, using the silhouette score as the fit quality

metric. The chosen K number of clusters were, hence, 5 for the PCA vectors and 3 for the TSNE vectors.
In order to visually inspect the results, aside from plotting the 2D graphs and the clusters, the nearest sample to each centroid was extracted, which originally corresponded to a clip and then the central frame of each of these clips was selected and visualized. Fig.2 shows the resulting frames obtained applying TSNE. Recalling that the preferred number of clusters was 3, less than the actual 8 classes of actions in the dataset, it appears that the sets represents different "domains" or "environments", in this case stove-fridge-sink. This is consistent to the output obtained with the PCA.
Therefore it can be reasonably assumed that this method isn't suited for capturing the dynamic features needed for action recognition tasks, since it seems to be better suited for extracting/classifying environmental features.

4. Chosen Architectures

In this section a brief introduction of the deep learning architectures used for the classification tasks is given.
Due to limited computational power, all the architectures implemented were basic. Moreover, the scope of this work isn't to find the new state-of-the-art procedure, but to explore different techniques and how they behave without any fine-tuning on their hyperparameters.

4.1. Multi-layer perceptron

The multilayer perceptron (MLP) is a feedforward neural network model in which multiple linear layers are stacked one on top of the other, with an activation function in-between layers. It is particularly well-suited for tasks involving pattern recognition, classification and regression, due to its ability to learn complex relationships within datasets. Usually the linear layers are used as the last layers of complex architectures like CNNs and Transformers.
The MLP implemented in this work consists of a sequential stack of 3 linear layers, each followed by a Rectified Linear Unit (ReLU) activation function, in order to introduce non linearity.

4.2. LSTM

Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture capable of capturing long-term dependencies in sequential data, while combating the vanishing/exploding gradient problem [5]. A fundamental difference with classic RNNs is the introduction of the LSTM cells, which enable the model to selectively retain or forget information over time. Each LSTM cell consists of three main components: an input gate, a forget gate, and an output gate. These gates regulate the flow of information: the input gate determines how to update the so called "long-term memory" of the system, the forget gate controls which information to remove from said long-term memory, and the output gate governs the release of relevant information from the cell, in the form of either "short term memory" or final output of the unraveled network.
Through this mechanism, LSTM can learn and store relevant long-term dependencies while selectively discarding irrelevant information, making it particularly well suited for tasks involving sequential data such as natural language processing, time series analysis, and video classification.
In this work the structure chosen is comprised 3 stacked LSTM layers, on top of which there is a fully connected layer, made by a simple linear layer, responsible for mapping the LSTM's output to the target of the classification (action labels).

4.3. Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a type of neural network architecture whose aim is to learn a set of convolutional filters, stacked in layers, going from lower to higher levels of abstraction. It is based on an assumption of locality and if said assumption is met, the architecture tends to perform very well, which is why it is often used in the context of image classification [7]. In this work, the same assumption is made by applying the CNN architecture to a very different type of input data, spectograms.
The CNN architecture used was inspired by tinyVGG [10], which is designed to be both lightweight and instructive.
TinyVGG is based around a basic building block formed by a CNN layer, a Relu activation function, another CNN layer, another Relu activation function and finally a max pool layer. This basic block is repeated two times and then finally connected to a fully connected layer.

4.4. Transformers

Transformers are a revolutionary neural network architecture born in the context of natural language processing, which can be also easily adapted to the field of video action recognition.
Unlike traditional RNNs or CNNs, Transformers do not rely on sequential processing or convolutional operations; instead, they are characterized by a self-attention mechanism that allows them to capture contextual relationships between tokens in an input sequence [9].
Transformers consist of an encoder-decoder structure with self-attention layers. The self-attention mechanism allows the architecture, for each token in a given sequence, to pay attention to all other tokens simultaneously, that is to say that it attributes an importance value to each token in a sequence, with respect to the current token being analysed.
Focusing on relevant context, Transformers can effectively capture long-range dependencies and contextual information, enabling them to generate high-quality representations of the input.
Therefore, the combination of this structure with fully con-

Figure 2. Central frames of central clips nearest to the centroids obtained by applying K-means to TSNE data

nected feed-forward networks, provides Transformers with the ability to model complex relationships in a data-driven and adaptive manner.
In the present work the first layer of the constructed architecture consists of a positional encoding layer, which adds to each token (in our case each feature vector, corresponding to a clip) the information about its position in relation to the other tokens (the other clips). Thus the model can learn and capture important sequential relationships among clips. This is a required "pre-processing" step since the Transformer model itself does not inherently have a sequential processing mechanism.
For what concerns the core of the Transformer implementation, 3 encoder layers, made up by self-attention and feed-forward networks, are stacked on top of each other, and finally a fully connected linear layer head is applied to map the output to the target of the classification (action labels).

5. Experiments

5.1. Transfer learning

Before analysing the results [Table 3] some general statements:

- I3D itself is trained with a specific sampling scheme (16 frames per clip, uniformly sampled), hence an improvement of the score is expected for the sampling schemes that more closely resemble I3D's training preprocessing
- the D1 Epic Kitchens sample set had a significantly smaller number of samples compared with D2 and D3, so a higher testing error is expected for this set. This problem is even more pronounced with the ActionNet sample set.

The results seem to corroborate the expectations regarding the sampling and the data subsets. A notable exception seems to be the trend present with dense sampling, which, although worse then the uniform counterpart, seems to have an opposite trend, performing better with 5 samples per clip compared to 15. This might happen due to the fact that with dense sampling the selected central frames are adjacent, so the information they carry might be redundant after a certain number of frames.

The results also indicate the Transformer as a top performer in this task. Given the sequential nature of the input, this architecture is indeed well suited for taking advantage of the full range of information present in the data. The also bounded nature of the length of the sequences (the fixed number of clips of an action) provides an additional advantage when it comes to the usage of transformers.
Finally, returning to the datasets, the accuracy scores of ActionNet are, as expected, the worst ones, but it is worth noting that the small sample size proves the transfer learning potential of I3D, especially, again, when paired with a Transformer and a sampling method similar to the one used in training.

5.2. Exploring a new modality: EMG

The reason for this part is two-fold:

- exploring this modality's standalone capabilities in terms of egocentric action classification
- training a CNN classifier on this modality in order to use it as a means for feature extraction

The results [Table 2] regarding the first endeavour are inconclusive. Trying to use an LSTM classifier on the preprocessed EMG wave resulted in very poor performance. This can be in part attributed to less relevant information present in this modality and in part to the fact that a substantial amount of said information has been lost during the preprocessing phase. The CNN classifier, trained on the spectogram data, also appears to have achieved a similar results, but it has done so with a fraction of the training iterations (300 instead of 5000), when compared with all the other experiments, and the accuracy seemed to still be improving. This choice was made due to the computational power constraints. It would also be reasonable to assume a better performance from the spectogram based CNN classifier since most the information present in the original input data is kept during the preprocessing phase.
Regarding the feature extraction, although the classification performance was not great, the features themselves could

still contain useful residual information that could be used in conjunction with the RGB features.

5.3. Multimodal classification

Analysing the results [Table 3], the added EMG features seem to have had a notable positive impact only in the case of the Tranformer, while in the case of MLP and LSTM the effect seems to be the opposite. This might be due to a higher importance of the sequential nature of the added features (EMG). There is inherently a trade-off in adding new features: the trade-off between the negative effect of the increase in data dimensionality and the positive effect of added information. If the sequential component of EMG is very important when it comes to action recognition, then it would make sense that the inclusion of this modality would produce a positive effect on an architecture like the Transformer and a negative effect on an architecture like MLP.
The outlier, in this case, would be the LSTM, but it must be noted that this model has been under-performing in all the experiments and might require a deeper architecture or more fine-tuning in order to show a similar pattern.
In general, it is hard to draw meaningful conclusions from the ActionNet experiments, given the size of the used subset, so further research would be needed.

	ACCURACY EMG %
LSTM + WAVE	16.67
CNN + SPECT.	16.67

Table 2. Results of the experiments done on ActionNet using only EMG features

		ACCURACY %					
		UNIFORM SAMPLING			DENSE SAMPLING		
		#frames=5	#frames=10	#frames=15	#frames=5	#frames=10	#frames=15
EK D1 RGB	**MLP**	49.20	53.33	54.02	53.10	46.67	47.59
	LSTM	43.22	51.26	51.26	45.75	37.47	37.24
	Transformer	53.33	56.32	56.09	55.41	46.67	49.43
EK D2 RGB	**MLP**	63.60	66.13	64.40	50.80	52.53	52.00
	LSTM	46.40	59.73	58.40	37.07	39.60	39.33
	Transformer	64.93	65.87	66.00	44.80	49.20	47.87
EK D3 RGB	**MLP**	65.20	68.69	57.80	64.58	55.65	55.65
	LSTM	48.56	61.09	57.80	50.41	44.35	44.76
	Transformer	65.71	70.53	70.43	62.73	50.72	50.00
ActionNet RGB	**MLP**	44.44	50	50	50	33.3	38.89
	LSTM	44.44	44.44	50	33.33	27.78	27.78
	Transformer	44.44	50	61.11	50	33.33	38.89
ActionNet MULTI	**MLP**	38.89	50	50	38.89	38.89	50.00
	LSTM	33.33	38.89	38.89	33.33	27.28	27.28
	Transformer	72.22	55.56	61.11	55.56	38.89	33.3

Table 3. Results of the transfer learning experiments on Epic Kitchens and ActionNet's data splits and of the multimodal training on ActionNet

6. Conclusions and Future Works

Action recognition for egocentric vision is an ever growing field, and in recent years multiple approaches to solving this task have been developed.

In the present work [8] the transfer learning capabilities of I3D model, applied on two different datasets, Epic Kitchens and ActionNet, have been tested through various experiments.
The results on the former confirm I3D's potential, especially when combined with a transformer classifier, which has reached, without any fine tuning, a 70% accuracy score at best. Unfortunately the ActionNet results require more caution because of the low number of samples. Further experiments could be done with the help of sampling-augmentation techniques, like bootstrapping or by extending the scope to the entirety of ActionNet and not just a subset. That being said, the transformer continues to be, also in this case, the best performing classifier analyzed.
The experiments regarding ActionNet's different EMG preprocessing and classification pipelines seem to have confirmed that the combined usage of CNN and spectrograms have superior potential with respect to the LSTM and EMG wave coupling, reaching the same accuracy with far less that a 10^{th} of the iterations.
For what concerns the possible gains obtained through the use of multimodal features, it appears that, in the case of a transformer classifier, this feature combination has shown its potential, almost always bettering the accuracy score, with respect to the solely use of RGB data, with the best gain being roughly 30%. Meanwhile, it seems that both the MLP and LSTM models do not benefit from this approach at all. This could be due to various reasons, which could be explored more in depth in future works.
Finally, a possible improvement to this results could be obtained by adding another modality, like sound or flow, while using a transformer architecture as classifier and by also adding a hyperparameters fine-tuning phase.

References

[1] Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, Feb. 2018. arXiv:1705.07750 [cs]. 2

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[4] Joseph DelPreto, Chao Liu, Yiyue Luo, Michael Foshey, Yunzhu Li, Antonio Torralba, Wojciech Matusik, and Daniela Rus. Actionsense: A multimodal dataset and record-

ing framework for human activities using wearable sensors in a kitchen environment. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 2

[5] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. 4

[6] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5491–5500, Seoul, Korea (South), Oct. 2019. IEEE. 2

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017. 4

[8] Ilaria Zerbini Laurentiu Apostol, Shahin Saeidpour. Github link: https://github.com/angrytako/deep-learning-egovision.git. 2023. 6

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 4

[10] Zijie J. Wang, Robert Turko, Omar Shaikh, Haekyu Park, Nilaksh Das, Fred Hohman, Minsuk Kahng, and Duen Horng Chau. CNN explainer: Learning convolutional neural networks with interactive visualization. *CoRR*, abs/2004.15004, 2020. 4