[0053] FIG. 4 is a schematic diagram of example components of one or more devices of FIG. 1 and/or FIG. 3, according to some non-limiting embodiments or aspects;

[0054] FIG. 5 is a schematic diagram of a vision transformer model, according to some non-limiting embodiments or aspects;

[0055] FIGS. 6A-6D are screenshots of a graphical user interface, according to some non-limiting embodiments or aspects;

[0056] FIGS. 7A-7F are diagrams of different pruning modes, according to some non-limiting embodiments or aspects;

[0057] FIGS. 8A and 8B are diagrams showing generation of attention strength vectors, according to some non-limiting embodiments or aspects;

[0058] FIG. 9 is a schematic diagram of attentions, according to some non-limiting embodiments or aspects; and

[0059] FIG. 10 is a diagram of attention patterns, according to some non-limiting embodiments or aspects.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0060] For purposes of the description hereinafter, the terms “end,” “upper,” “lower,” “right,” “left,” “vertical,” “horizontal,” “top,” “bottom,” “lateral,” “longitudinal,” and derivatives thereof shall relate to the embodiments as they are oriented in the drawing figures. However, it is to be understood that the embodiments may assume various alternative variations and step sequences, except where expressly specified to the contrary. It is also to be understood that the specific devices and processes illustrated in the attached drawings, and described in the following specification, are simply exemplary embodiments or aspects of the disclosed subject matter. Hence, specific dimensions and other physical characteristics related to the embodiments or aspects disclosed herein are not to be considered as limiting.

[0061] No aspect, component, element, structure, act, step, function, instruction, and/or the like used herein should be construed as critical or essential unless explicitly described as such. Also, as used herein, the articles “a” and “an” are intended to include one or more items and may be used interchangeably with “one or more” and “at least one.” Furthermore, as used herein, the term “set” is intended to include one or more items (e.g., related items, unrelated items, a combination of related and unrelated items, and/or the like) and may be used interchangeably with “one or more”

pruned layer output(s) for the respective attention head, the at least one respective pruned model output for the respective attention head, or any combination thereof.

[0091] In some non-limiting embodiments or aspects, determining the at least one importance metric may include determining at least one of: a probability change importance metric based on the at least one model output and the at least one respective pruned model output for the respective attention head, a Jensen-Shannon Divergence (JSD) importance metric based on the at least one model output and the at least one respective pruned model output for the respective attention head, a layer importance metric based on the at least one layer output and the at least one respective pruned layer output for the respective attention head, or any combination thereof. For example, the layer importance metric may include at least one of a class token-based layer importance metric, a patch token-based layer importance metric, or any combination thereof.

[0092] As shown in FIG. 2, at step 212, process 200 may include generating an attention strength vector for each attention head. For example, for each respective attention head of the plurality of attention heads, model analysis system 102 may generate a respective attention strength vector for the respective attention head based on an attention strength between each patch of the at least one sample and each other patch of the at least one sample.

[0093] As shown in FIG. 2, at step 214, process 200 may include determining at least one attention pattern. For example, model analysis system 102 may determine the attention pattern(s).

[0094] In some non-limiting embodiments or aspects, each respective attention head of the plurality of attention heads may be associated with a respective attention matrix (e.g., a plurality of attention matrices may include the respective attention matrix for each respective attention head of the plurality of attention heads). In some non-limiting embodiments or aspects, determining the attention patterns may include generating (e.g., by model analysis system 102) a respective embedding for each respective attention matrix of the plurality of attention matrices (e.g. a plurality of embeddings may include each respective embedding for each respective attention matrix of the plurality of attention matrices). The plurality of embeddings may be clustered (e.g., by model analysis system 102) to provide at least one cluster. Model analysis system 102 may determine at least one attention pattern based on the cluster(s).