# PROCESSING MULTI-MODAL INPUTS USING LANGUAGE MODELS

## CROSS-REFERENCE TO RELATED APPLICATION

This application claims priority to Greek national patent application number GR 20210100374, filed on June 8, 2021. The disclosure of the prior application is considered part of and is incorporated by reference in the disclosure of this application.

## BACKGROUND

This specification relates to processing inputs that include images using neural networks.

Neural networks are machine learning models that employ one or more layers of nonlinear units to predict an output for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to the next layer in the network, e.g., the next hidden layer or the output layer. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

## SUMMARY

This specification describes a system implemented as computer programs on one or more computers in one or more locations that performs a multi-modal task using an image encoder neural network and a language model neural network. The language model neural network is pre-trained, e.g., is not trained during the training of the image encoder neural network.

In one aspect, one of the methods includes obtaining an input image; processing the input image using an image encoder neural network that has been trained to process the input image to generate an image sequence of image embedding vectors characterizing the input image; generating an input sequence of input embedding vectors that includes the image sequence of image embedding vectors; and processing the input sequence of input embedding vectors that includes the image sequence of image embedding vectors using a trained language model neural network to generate as output an output sequence of text tokens that represents a text response to the input sequence of input embedding vectors, wherein: each text token in the output sequence is selected from a vocabulary of text tokens, the trained language model neural network has been trained to perform a language modeling task on training input sequences of text embedding vectors that represent text tokens from the