Il file demo.py contiene una linea di quello che devi fare.

Nella cartella data ci sono i dati che servono per far funzionare la rete:

* model.pth contiene i pesi della rete che fa VQA. Questa rete prende in ingresso le features dell’immagine e la domanda.
* dictionary.pkl è il dizionario delle parole che utilizza la rete. traduce ogni indice che identifica una parola nel corrispettivo embedding che la rete utilizza.
* dict\_ans.pkl è un file che contiene una lista di tutte le risposte che la rete può dare (ordinate) e un dizionario risposta→ indice.
* artpedia\_bottomup\_qa.json e artpedia\_visual\_qa\_cap.json sono due file json contenenti coppie di domanda risposta nella forma id:{domanda, risposta} dove id corrisponde all’id dell’immagine di Artpedia (queste coppie domanda-risposta sono relative alla gran parte delle domande di Artpedia ma non tutte).

Il primo passo è quello di costruire un dataloader contente domande di Artpedia, features delle immagini di Artpedia, risposte. IMPORTANTE le risposte delle domande spesso contengono più parole ci sono due strade per codificarle:

1. codificare ogni parola con la funzione tokenize del dictionary (come si fa per le domande).
2. mettere tutte le risposte in dizionario e assegnare a ogni risposta un numero (id) che viene usato nel dataloader. A quel punto per estrarre la riposta testuale basterà inserire l’id nel dizionario e prendere la riposta.

una volta creato il dataloader si cicla sulle batch (puoi usare anche batch di dimensione 1 tanto siamo in fase divalutazione se preferisci). Dentro il ciclo avviene la valutazione dell’accuracy. IMPORTANTE il modello di VQA mette in output un tensore con le probabilità di ogni risposta la risposta con la probabilità più alta è quella in output (argmax). Una volta che hai l’indice della risposta predetta potrai accedere alla risposta testuale tramite dict\_ans.

per valutare l’accuracy IMPORTANTE siccome il modello di VQA risponde con risposte composte quasi sempre da una singola parola (mentre le nostre risposte di ground truth sono composte da più parole) conviene valutare l’accuracy in questo modo:

* se la parola/parole della riposta predetta sono contenute anche nella risposta ti ground truth allora il modello risponde in modo corretto.