2. Text to Image

(a) המודלים יהיו כדלקמן:

(1) Text Encoder: מודל שמקבל משפט ומחזיר וקטור מימד נמוך יותר שמייצג את המשמעות של המשפט ע"י RNN או Transformers

(2) Image Decoder: מקבל וקטור מימד נמוך כמו המימד שאליו דיקדנו את המשפט ומייצר ממנו תמונה (ע"י רשת CNN הפוכה)

ע"ש שימוש ב-transformers ורשתות CNN! RNN מודל יכול ללמוד לבנות משפט לוקטור שמדקדק את המשמעות ומודל יותר יכול ללמוד להפוך וקטור יותר קטן לתמונה בגודל שמתאים וכך אמור לבנות עוד מדגם לייצר אמוג'י מלקט.

(b) בגישה הכי מתאימה לנו בעזרת attention למד את הרשת ליצירת משפט המתקבל באופן שיהיה רלוונטי למיקום עליו מדובר בתמונה.

1. נוכל להשתמש ב attention ב Text encoder עלינו לייצר וקטורים שמייצגים את כל המילים של המשפט המתאימות.

2. נשתמש ב attention בכדי לדעת לשים שמייצג את המיקום (הרבוע) בוקטור ש ה encoder מייצר לפיו לפי כיצד בייצור התמונה ולהשפיע בהשמה על המיקום של הדקדוקים המתאימים במשפט.

כלומר לייצר באופן יותר עדין תמונה שמייצגת את המשפט.

Transformers networks.

בשכבות מבוססות שכבת attention, סדר ה tokens לא משנה את הניקוד
לכל token. שכבת attention מתייחסת אל
כל ה tokens כדי לבדוק את ההשפעה שלהם איתם על בחירת ולכן
שומרת תשומת לב ברמה הגלובלית.

בשכבות מבוססות שכבת FC היא נמדדת על כל הנתונים
באופן בלתי תלוי אחת משנייה ובנוסף ללא תלויות לסדר.

לכן הדרישה של שכבת attention ! FC מייצרת
Permutation invarient operator.

אם נוסיף Positional encoding, שמוסיפה את הניקוד על מיקום
ה token בתוך ה tokens באמצעות נגם המיקום בסדר
לתוך הקידוד,
נקבל שהדרישה של הפונקציה Permutation invarient
לא הושגה שהוסרה המבוצעת לגמרי בגלל את המיקום של
ה tokens ומכאן המיקום לא אבוד יותר.