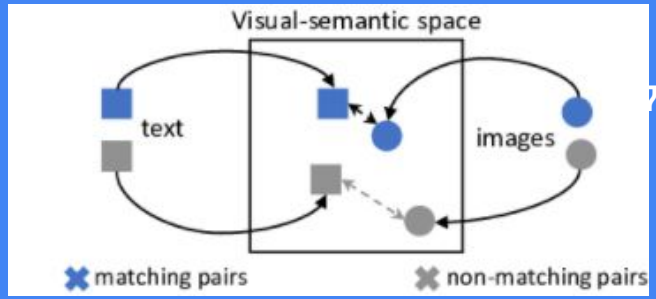
Joint Embedding (Text /Audio /Image)



אלמוג אמיגא ואריאל יחזקאל

Background

the idea of word embedding

Word embedding הוא מונח המשמש לייצוג מילים לניתוח טקסט, בדרך כלל בצורה של וקטור עם ערך מספרי המקודד את משמעות המילה כך שוקטורים שקרובים אחד לשני במרחב הוקטורי, המילים שהן מייצגות יהיו דומות מבחינה סמנטית.

Word2vec היא שם כולל למודל המשמש ליצירת Word embedding.

word2vec- טכניקה להפוך מילה אחת לוקטור מספרי אבל הוקטור הזה שומר

דימיון סמנטי לכן מילים שדומות יהיו נמצאות בוקטורים עם מרחק קרוב ומילים

שונות יהיו נמצאות במרחק רחוק יותר זה מזה.

Similar words: nearby embeddings

W2v("king") = [-3.168
-0.136
3.770
4.767
3.558
-4.168
0.464
2.034
3.411
...
0.866]

W2v("queen") = [-3.101
-0.057
3.800
4.862
3.632
-4.157
0.549
2.064
3.428
...
0.884]

D(W2v("king") – W2v("queen")) = 0.188

embedding

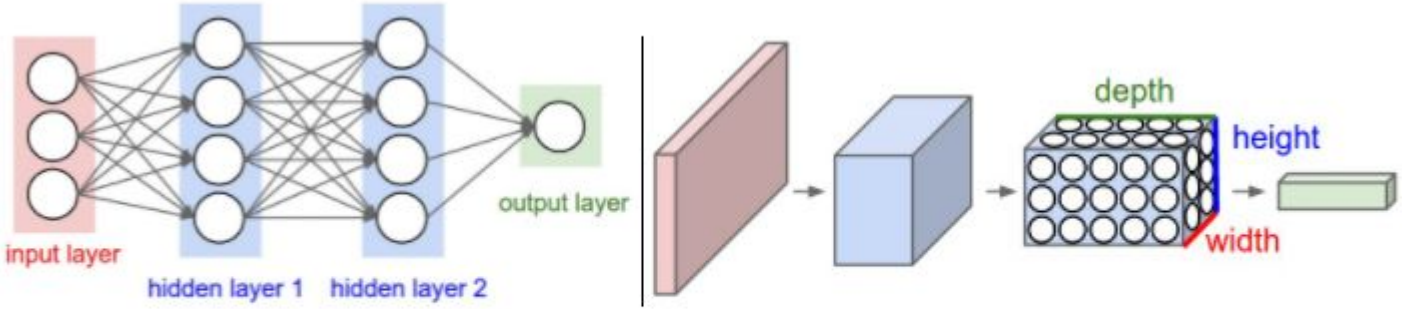
שיטה המשמשת לייצוג משתנים בדידים כוקטורים רציפים. בעצם מיפוי של משתנה קטגורי/בדיד לוקטור של מספרים רציפים.

NN embeddings שימושיים מכיוון שהם יכולים להפחית את הממד של משתנים קטגוריים ולייצג בצורה משמעותית בממד הקטן יותר.

הוקטורים הסופיים הם ייצוגים של המשתנה הבדיד כאשר למשתנים דומים, הוקטורים המייצגים אותם יהיו קרובים יותר זה לזה במרחב.

דוגמא להקטנת מימד

לתמונת הקלט יש ממדים 32x32x3 ולפלט הסופי עבורה יש ממדים 1x1x10



Left: A regular 3-layer Neural Network. Right: A ConvNet arranges its neurons in three dimensions (width, height, depth), as visualized in one of the layers. Every layer of a ConvNet transforms the 3D input volume to a 3D output volume of neuron activations. In this example, the red input layer holds the image, so its width and height would be the dimensions of the image, and the depth would be 3 (Red, Green, Blue channels).

Image Embedding

אם יש לנו 2 תמונות מאותו סוג של אובייקט, אנו יכולים לקבל 2 וקטורים בעלי מרחק קטן. באמצעות אלגוריתם KNN ששולף את הנקודה k הקרובה ביותר ממערכת נקודות, ובכך אנו יכולים לאחזר ביעילות תמונות סמוכות(דומות).

https://rom1504.github.io/image_embeddings/

מקרה שימוש פשוט של image embedding הוא אחזור מידע. עם קבוצה גדולה מספיק של image embedding, הוא מאפשר בנייה של יישומים מדהימים כגון:

- חיפוש צמח באמצעות תמונות של הפרח שלו, העלים שלו, וכו...
- חיפוש תמונה דומה בכל האינטרנט.
- מציאת מוצרים בעלי מראה דומה.

joint embedding

הגדרה - שילוב של לפחות שני סוגי דאטה לאותו מרחב הטמעה (embedding space), הנקרא joint space.

שילוב למרחב ששם הם יכולים להיות ברי השוואה - כלומר מוצגים באותו ה"שפה". לדוגמא: (טקסט, אודיו, תמונה ועוד).

embedding space - מרחב בעל מימד נמוך שאליו ניתן לתרגם אליו וקטורים בעלי ממדים גבוהים. embedding מקל על למידת מכונה בהינתן קלטים בעלים ממד גדול.

הרחבה של ההגדרה

אם יש לנו מודל המשלב בין שני קלטים שונים או יותר(מבחינה מרחבית), למשל בין קלט של טקסט וקלט של תמונה אז לאחר embedding על הקלט, הפלט של ה-embedding על כל קלט הוא וקטור המתאר אותו - כלומר יש לנו עכשיו שני מרחבי וקטורים המשתרעים על נתונים שלא ניתנים לצפייה ישירה אך הם כן מתארים את שני המודלים (במקרה שלנו - תמונה/טקסט). שני מרחבי הוקטורים האלו נקראים מרחבים סמויים (latent space). ולבסוף 2 המרחבים הסמויים משולבים ל-Joint Space ששם הם ברי השוואה - כלומר מדברים באותה ה"שפה" למרות שהגיעו ממרחבים שונים.

אפשר לחבר את המקורות השונים בשני מקומות ברשת -

1. ללמוד כל מקור בנפרד ואז בשכבה האחרונה או קרוב אליה לחבר את שתי ההחלטות של המודלים השונים.
2. ללמוד מההתחלה את הקשר בין שני המרחבים.

Joint Audio-Text Embedding

קצת רקע

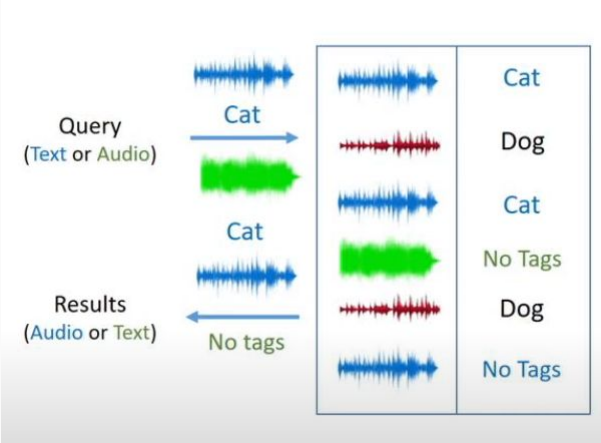
הרבה הקלטות מולטימדיה ושמע נתפסות כל יום. כל ההקלטות האלה נושאות מידע שאומר לנו מה קורה בתוכן. יש הקלטות שנתפסות ומשותפות ברשתות החברתיות למשל, כאשר לאנשים בדרך כלל יהיו כמה מילים או תיאורים לתאר מה יש בתוכן ההקלטה

אבל מצד שני יש גם הקלטות כמו אמזון אקו או אלקסה שבהם אף אחד לא לוקח את הזמן לתייג או להוסיף הערות לצלילים או איזה אודיו יש בתוכן ההקלטה

אז אם יש לנו את ההקלטות האלה קשה לעשות חיפוש ואחזור.

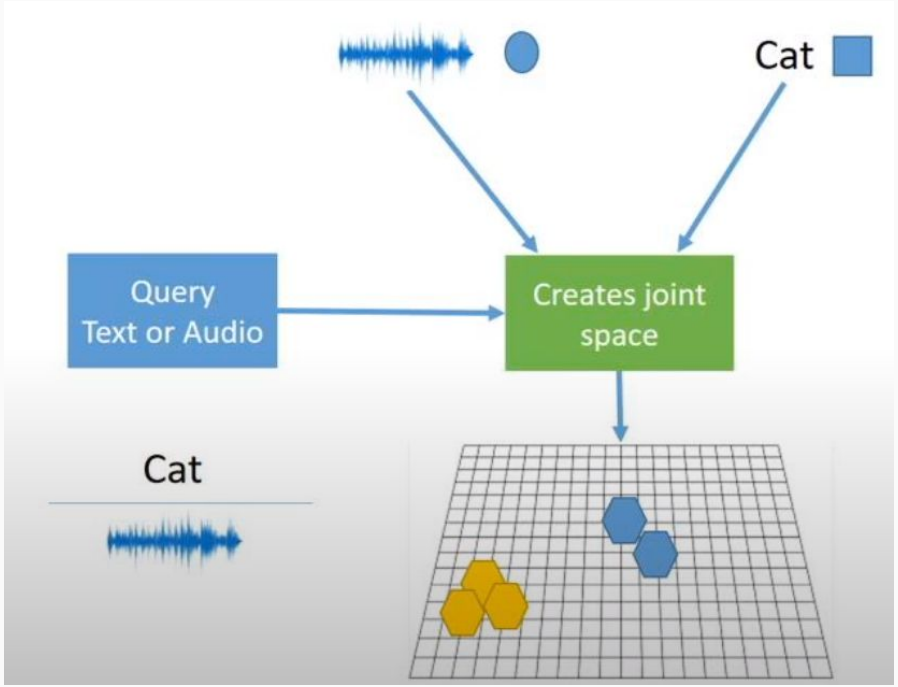
למה צריך?

יש פעמים שאין סיווג שקשור לשמע. לכן אנחנו לא יודעים מה התוכן של השמע או לא יודעים כיצד לפענח את רעש השמע עד שאנחנו שומעים אותו בהקשר מסוים. יש רעשים שקל לנו לשמוע ללא הקשר כמו נביחה של כלב. בנוסף, יש אנשים בעלי מוגבלויות שלא יכולים לשמוע את השמע וצריכים להבין את המתרחש.



המטרה שהjoint space ימצא נקודת אמצע לשמור

את הדימיון של הטקסט והאודיו באותו הזמן.


Cat
Query
Text or Audio
Creates joint
space
Cat

מטרות

1. למפות טקסט ושמע לאותו מרחב משותף. דבר זה יתרום לנו כי שמע וטקסט יכולים להיות ברי השוואה. כלומר, אם נשלח שאילתה עם טקסט או שמע מוקלט נוכל למפות את זה לאותו מרחב משותף ונמצא את השכנים שתואמים לשאילתה.
2. המרחב המשותף אמור לשמור על דימיון המשמעות (סמנטי) בין השמע לטקסט.

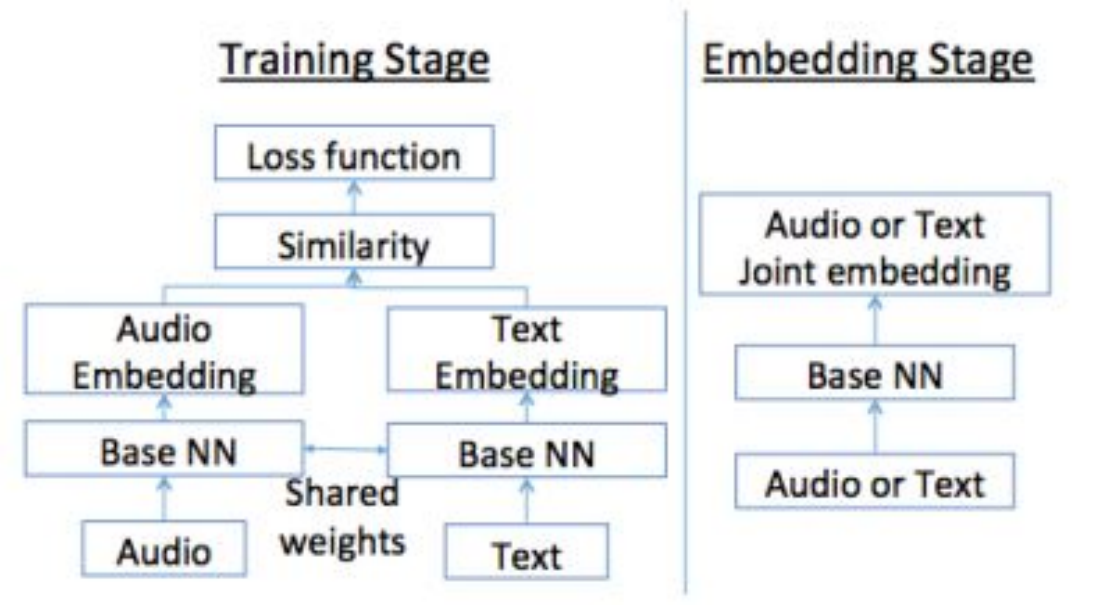
מודלים עצמאיים יכולים לנטות לשגיאות במיוחד עבור הקלטות שנוצרו על ידי משתמשים, הרועשות הן באודיו והן בתוויות הטקסט שמקושרות אליהן.

כדי להתגבר על מגבלה זו, הוצעה מסגרת שלומדת joint audio-text embedding, שבו ניתן למפות וקטורים מכל אחד מהמודלים ולהשוות אותם ישירות.

שיטות קודמות

- גישה של חיפוש ואחזור על ידי שאילתות טקסט ואודיו, למרות שהגישה מהירה, היא פותחה בדרך בלתי תלויה היכן שהטקסט והאודיו אומנו בנפרד ואז בנקודה מסוימת השתלבו לתת ציון ולאחזר מידע. הגישה לא עבדה ללא אוצר מילים.
- חלל סמנטי עם טקסט ואז בחלל האקוסטי הוקטן המימד של האודיו ואז קישרו בין שני החללים האלו. קשה לעבור למידות גדולות כי צריך ליצור את הקשר הזה.

ארכיטקטורה של ה- SNN המשמשת ללמוד את המרחב המשותף ולחשב טבלאות משותפות של שמע וטקסט.



ארכיטקטורת הרשת

Siamse network- לוקח כקלט אודיו וטקסט בכניסות ויסווג את הזוגות כחיוביים או כשליליים.

חיוביים- אומר שהטקסט והאודיו תואמים לאותו הקלאס.

שליליים- אומר שהטקסט והאודיו תואמים לקלאס שונים.

הזוגות האלה יאמנו את הרשת ובסוף הרשת הloss function מנסה לגרום לכך ששתי שיטות שתואמות לאותו הקלאס יהיו קרובים יותר ואלה שלא תואמו יהיו רחוקים יותר.

embedding stage- לאחר אימון הרשת נוכל להזין טקסט או אודיו וליצור ייצוג משותף.

מכיוון שהיינו צריכים וקטור חישבנו את הממוצע של המטריצה שתאמה לכל אחד מההקלטות ואז יצרנו ממוצע וקטורים לאורך זמן הקלטת האודיו.

לאחר מכן משתמשים בNN פשוטה ואז יש לנו את הembedding שהוא מהאימון.

קוראים לרשת כך מכיוון ששני הצדדים הם אותו הדבר וחולקים אותם פרמטרים.

המטרה של ה loss המנוגד זה לכפות על זוגות שתאמו לאותו קלאס דומה להיות קרובים וכאלה שתאמו לclass שונה להיות רחוקים.

בחרו באלגוריתם KNN כי זה מבטיח מסגרת טובה יותר של הרעיון שיש שכונה של embedding ואז יהיה קל יותר להציג את תוצאות האחזור בהמשך. קל יותר לשלוף מידע.

Joint Video (audio) Embedding

Cocktail Party

בני אדם מסוגלים למקד את תשומת הלב השמיעתית שלהם על מקור צליל יחיד בסביבה רועשת, תוך התעלמות מרעשי רקע.

אפקט זה נקרא אפקט מסיבת הקוקטייל (cocktail party effect), אנו נציג מודל שפיתח יכולת זו מבחינה חישובית.

https://youtu.be/rVQVAPiJWKU?t=199

למה צריך?

משימת סיווג באמצעות שמע כקלט בלבד הינה מאתגרת ולא מבטיחה הצלחה.

מכיוון שהתכונות החזותיות (visual features) משמשות ל"מיקוד" השמע של האנשים הרצויים ולשיפור הפרדת הדיבור של כל מדבר המופיע בסרטון.

בעיות

1. הפרדת דיבור אוטומטית - הפרדת אות שמע קלט למקורות הדיבור האישיים שלו. על מנת להשיג פתרון סביר דרוש ידע מוקדם או תצורות מיקרופון מיוחדות.
2. בעיית הסיווג: אין דרך קלה לשייך כל מקור שמע מופרד למדבר המקביל שלו בסרטון.

שיטות קודמות

Canonical Correlation Analysis- היא דרך להסיק מידע ממטריצות. אם יש לנו שני וקטורים $X = (X1, ..., Xn)$ ו- $Y = (Y1, ..., Ym)$ של משתנים אקראיים, ויש קורלציות בין המשתנים, אז ניתוח Canonical Correlation ימצא שילובים לינאריים של X ו- Y שיש להם קורלציה מקסימאלית זה עם זה.

שיטות מבוססות Canonical Correlation Analysis אינן יציבות. בגלל עלות הזיכרון הגבוהה שלהן, עלות הזיכרון דורשת טעינה של כל הנתונים בזיכרון כדי לחשב את מטריצת נתוני המשתנים.

יתרונות

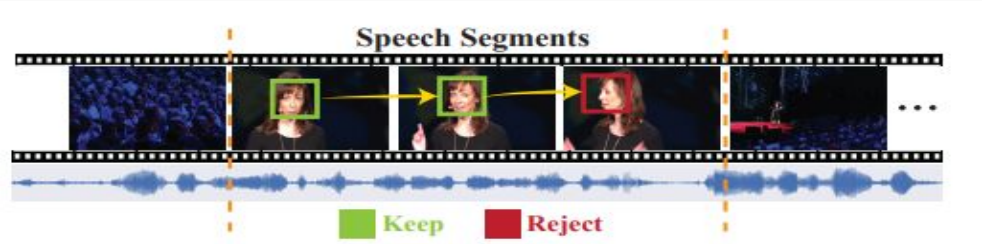
היתרונות של הגישה על פני שיטות שמע בלבד:

- תוצאות ההפרדה של המודל האודיו-וידיאו הן באיכות גבוהה יותר מאלו של מודל שמע בלבד.
- הגישה מתפקדת היטב בהינתן כמות מדברים מרובים יחד עם רעשי רקע, שום שיטת שמע בלבד לא נפתרה באופן מספק (ביחס למה שמצאנו באינטרנט).
- המודל פותר במשותף שתי בעיות בעיבוד דיבור: הפרדת דיבור והקצאת אות דיבור לפני כל מדבר בסרטון במקביל (עד כה טופלו בנפרד).
- המודל הוא בלתי תלוי בדוברים שבוידיאו, בגישות קודמות, המודלים היו תלויים בדוברים כלומר יש להכשיר מודל ייעודי בנפרד לכל בן אדם.
- המודל מסוגל להפריד ולשפר את הדוברים שיש בסרטון גם לגבי דוברים שאין בדאטה והמודל לא ראה מעולם לפני כן (בסט האימונים).
- המודל עובד גם על שפות שלא היו באימון המודל.

AVSpeech dataset

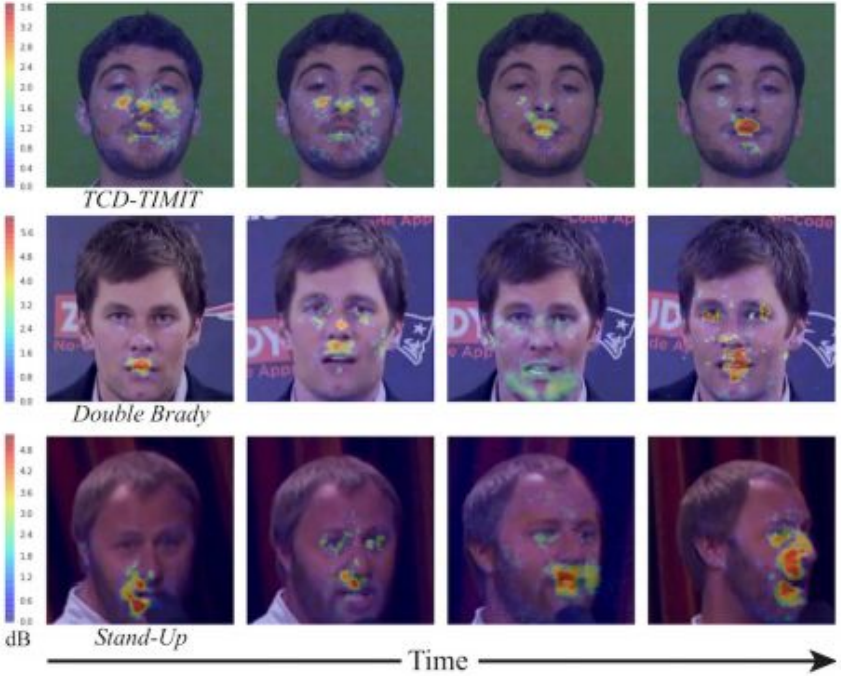
AVSpeech- מערך נתונים שמורכב מאלפי שעות של קטעי וידיאו מהאינטרנט. דוגמאות של תרחישים בעולם האמיתי אותו רוצים לבודד. מערך נתונים זה מגוון וגדול מספיק על מנת לתת תוצאות טובות.

בחירת הדאטה: תחילה, ניקוי חלקים שהפנים מטושטשות או שלא מצליחים לתפוס הבעת פנים טובה וכד', ניקוי הנאום לקטעים שיכללו דיבור נקי על ידי SNR.



SNR- הוא היחס בין עוצמת האות הרצוי לבין עוצמת הרעש הכולל. היחס מתאר את המידה שבה הרעש הבלתי רצוי הוא משמעותי או חזק, היחס מהווה אינדקציה עבור שימוש באות וניתוחו.

איך הרשת מצליחה לזהות?



הרשת משתמשת במודל של זיהוי פנים ובמודל של זיהוי קול ומשלבת את שניהם על מנת לקשר בין הדובר הראשי לקול שלו. לדוגמא בתמונה רואים זיהוי מתי הדובר מדבר על ידי heat map.

כצפוי בתמונה של האדם המדבר, הפה הוא הדבר שמעניין (וגם יהיה צבוע בהכי הרבה אדום) אך גילו כי גם העיניים או הלחיים תורמים.

ייחודיות נוספת

הרשת שימושית במיוחד גם עבור זיהוי דיבור אוטומטי (ASR). לדוגמא, הריצו את הרשת על סרטוני סטנד-אפ הנמצאים ביוטיוב, ה ASR של יוטיוב לא הצליח לתפוס קטעי דיבור אשר הרשת כן הצליחה לתפוס ובנוסף קטעים שה ASR של יוטיוב תפס בצורה לא מדויקת, הרשת תפסה בצורה מדויקת.

קלט ופלט של המודל

קלט: וידאו (כלומר כמה פריימים של תמונות עם אודיו) עם אחד או יותר אנשים שמדברים כאשר מה שאנחנו רוצים זה להפריד בין הקול של הבן אדם לרעשי הרקע.

פלט: פירוק של רצועת האודיו למסלול דיבור נקי, אחד לכל מי שמדבר בסרטון. דבר זה מאפשר לנו לאחר מכן לחבר קטעי וידאו שבהם הדיבור של אנשים ספציפיים משופר(מבחינה ווקאלית) כאשר שאר צלילי הרקע מדוכאים (כלומר ניתן לשמוע בבירור רק את מי שאנחנו רוצים).

כל מה שדרוש מהמשתמש זה לבחור איזה מבין הדוברים בסרטון הוא רוצה לשמוע.

STFT (Short-time-fourier-transform)

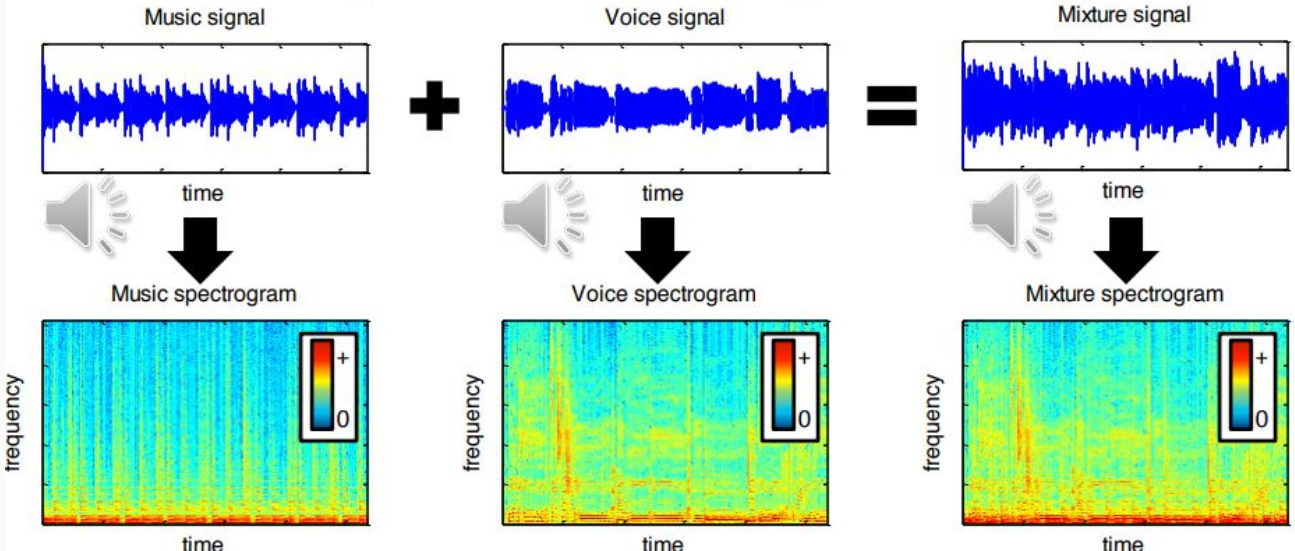
STFT זה התפלגות תדירות-זמן (time-frequency distributions).

בפועל, ההליך לחישוב STFT הוא לחלק אות זמן ארוך יותר למקטעים קצרים יותר באורך שווה ואז לחשב את טרנספורמציית פורייה בנפרד על כל קטע קצר יותר

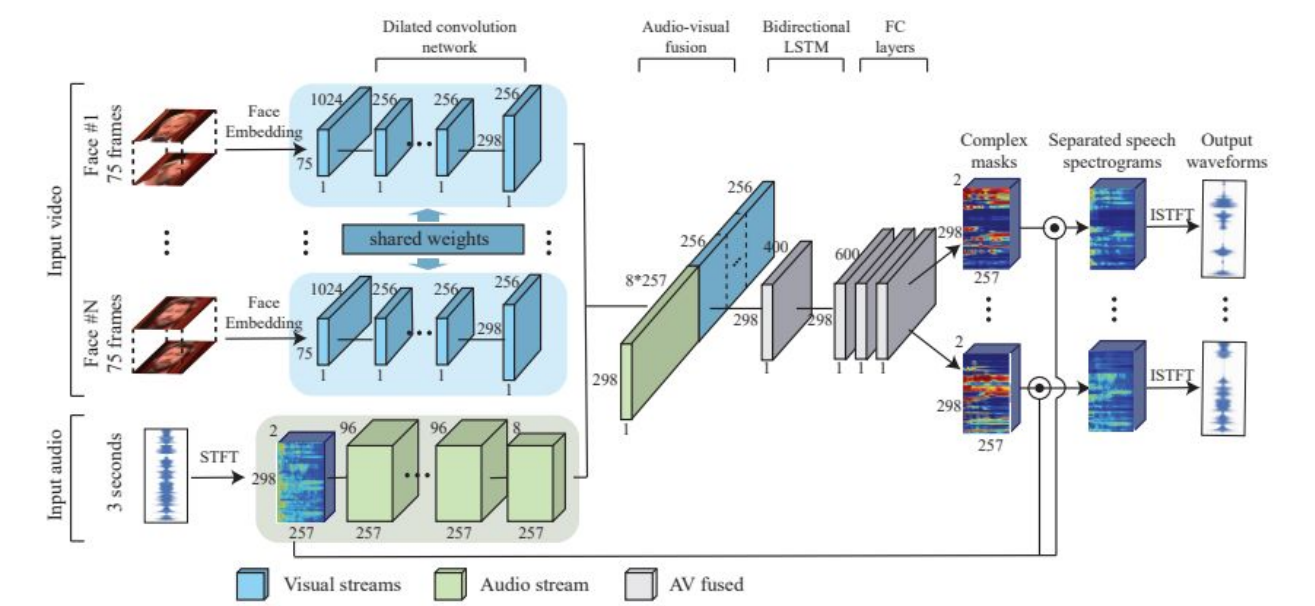
טרנספורמציית פורייה - כלי מרכזי באנליזה הרמונית שמפרק פונקציה לרכיבים מחזוריים ומבצע אנליזה מתמטית לפונקציה על ידי ניתוח רכיביה.

STFT with spectrogram

- Suppose we have a mixture of two sources: a music signal and a voice signal



ארכיטקטורת הרשת



תיאור בסיסי של הרשת

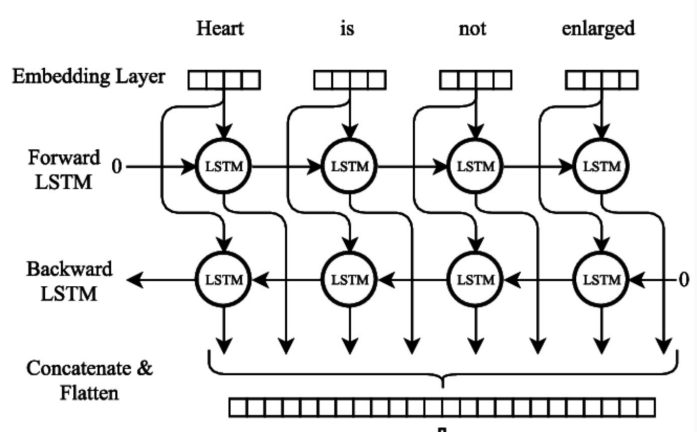
עיצוב ואימון מודל מבוסס רשת NN שלוקח את תערובת הצליל המוקלטת, יחד עם פרצופים מזוהים בכל פריים בסרטון כקלט, ומפצל את התערובת לזרמי שמע נפרדים לכל אדם שזוהה.

הזרמים הוויזואליים לוקחים כקלט תמונות של פרצופים שזיהה בכל פריים בסרטון, וזרם השמע לוקח כקלט את האודיו, המכיל תערובת של דיבור ורעש רקע. הזרמים החזותיים מחלצים טבילות פנים לכל תמונה ממוזערת באמצעות מודל מאומן לזיהוי פנים, ואז לומדים תכונה חזותית באמצעות CNN מורחב. זרם האודיו מחשב תחילה את ה- STFT של אות הקלט לקבלת ספקטרוגרמה(הצגה ויזואלית של עוצמת אות הקלט), ואז לומד ייצוג שמע באמצעות CNN מורחב. ייצוג אודיו-תמונה משותף נוצר לאחר מכן על ידי שרשור התכונות החזותיות והשמע שנלמדו, ועובד לאחר מכן באמצעות LSTM דו כיווני יחד עם שלוש שכבות מחוברות לחלוטין. הרשת מוציאה כפלט מסיכת ספקטרוגרמה מורכבת לכל אדם המדבר בסרטון, ומומרת חזרה לצורות גל לקבלת אות דיבור מבודד לכל אדם שמדבר בסרטון.

lstm and bidirectional lstm

מודל LSTM - שומר מידע מהקלט שכבר עבר דרכו.

מודל lstm דו כיווני - שומר מידע מהעבר לעתיד וגם מהעתיד לעבר ומה שמבדיל גישה זו מ LSTM חד-כיווני הוא שב- LSTM שפועל לאחור אתה שומר מידע רק מהעתיד ובאמצעות דו כיווני מסוגלים בכל נקודת זמן לשמור מידע מהעבר ומהעתיד.



Joint Video (text) Embedding

מטרה

המטרה היא לפתור את בעיית image-text retrieval, כלומר בהינתן סרטון, המודל יפלוט משפט שמתאר את המתרחש בסרטון.

לשיטות הלוקחות כקלט רק תמונה יהיה קשה להבדיל בין כלב נובח לכלב משחק, בגלל זה חשוב להשתמש גם בוידיאו בעל פעולות ואודיו.

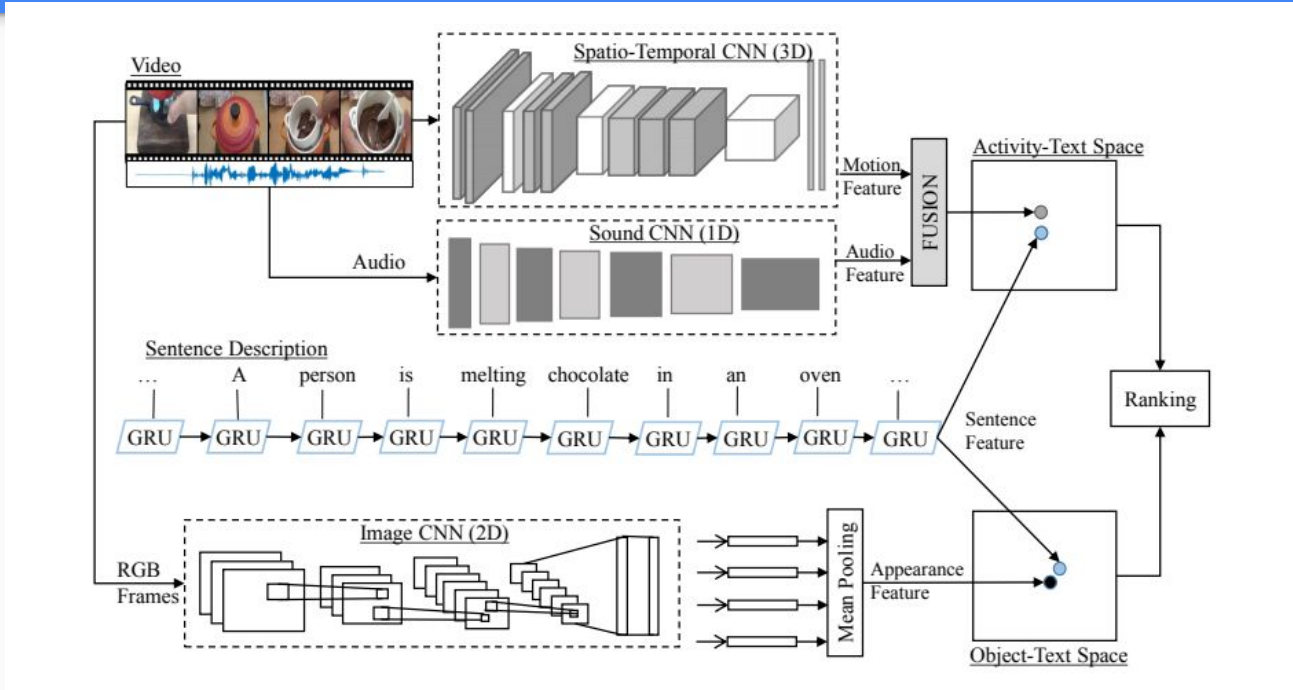
עלינו להבין שני היבטים עיקריים של כל סרטון:

1. האובייקטים הבולטים של הסרטון.
2. הפעולה והאירועים בסרטון.

יתרונות

- ההצלחה של אחזור טקסט הווידאו תלויה בהבנת וידאו חזקה יותר. שיטה זו מנסה להשיג את המטרה על ידי שימוש בתכונות ויזואליות שונות ושמע מסרטון.
- המסגרת המוצעת משתמשת בתכונות של פעולה, אובייקט, טקסט ואודיו. משתמשים ב-pairwise ranking loss כדי ללמוד טוב יותר את joint embedding.
- שיפור ברור בהשוואה לשיטות החדשות האחרות.
- הקלט הינו קטעי וידיאו באופן כללי לכן גילויים של פעילויות ואובייקטים יביאו ביצועים טובים יותר.

אריטקטורת הרשת



קלט של המודל

הקלט הוא וידיאו וטקסט, הטקסט הופך לפיצ'ר אחד והוידיאו הופך לשלושה פיצ'רים.

GRU -Gated recurrent units

GRU הוא כמו LSTM עם forget gate, אך יש לו פחות פרמטרים מאשר LSTM, מכיוון שהוא חסר output gate.

הוכח כי GRU משיג ביצועים טובים יותר במערכי נתונים קטנים יותר(משפט המתאר סרטון) ביחס ל LSTM.

פלט של המודל

בהינתן שאילתה (סרטון) מחשבים את ציון הדימיון של משפט השאילתה עם כל אחד מהסרטונים בדאטהסט בשני ה- joint embedding spaces. ומשתמשים בסכום ציוני דימיון עבור חישוב הranking (הפלט) הסופי.

אחזור מבוצע על ידי חיפוש בשכן הקרוב ביותר(KNN).

הסבר על שני המרחבים במודל

לצורך אחזור יעיל איתור האובייקטים והפעילויות מהסרטון חשוב מאוד לביצועים גבוהים יותר. מודל אחד לומד מרחב משותף (Activity-Text Space) בין תכונת טקסט לשילוב של תכונות פעילות ושמע. מודל אחר לומד מרחב משותף (Object-Text Space) בין תכונות טקסט לתכונות מראה חזותיות. כאן, מרחב Object-Text הוא המומחה לפתרון דו משמעותיות בין אובייקטים בסרטון, ואילו Activity-Text הוא המומחה לפתור דו משמעותיות בין פעולות / אירועים בסרטון.

features

Text Feature- שימוש בGRU בשביל קלט הטקסט. בעצם ניתן לראות שכל מילה בקלט הטקסט הינה GRU.

Object Feature- לצורך קידוד התמונה, משתמשים במודל CNN מאומן מראש. באופן ספציפי, משתמשים במודל ResNet-152 אשר מתמקדת בזיהוי אובייקטים בסרטון. אנו מחלצים תכונות תמונה ישירות מהשכבה הלפני אחרונה המחוברת לחלוטין. מתאימים את גודל התמונה ומזינים ל- CNN כקלט.

Activity Feature- משתמשים במודל CNN מאומן מראש. באופן ספציפי, משתמשים במערכת במודל RGB-I3D מאומן מראש אשר מתמקדת בזיהוי הפעילויות בסרטון.

Audio Feature- על ידי שיוך שמע נוכל לקבל רמזים חשובים לאירועים האמיתיים, שיעזרו לנו להסיר דו משמעותיות במקרים רבים. משתמשים ברשת SoundNet CNN. משתמשים רק אם יש אודיו בסרטון.

סיכום

וקטור מהמרחב הסמוי text-embedding ווקטור שמתאים לו מהמרחב הסמוי של video-embedding אמורים להיות קרובים אחד לשני במרחב המשותף.

משימת אחזור הטקסט-וידיאו מחזירה עבור כל שאילתת סרטון, רשימה מדורגת של תיאור הטקסט הסביר ביותר במערך הנתונים.

שני ההיבטים העיקריים של כל סרטון הם אובייקטים והפעולות בסרטון.

object-text embedding space- הוא המרחב המשותף אליו ממופים גם תכונות המראה וגם הטקסט(מקשר בין וידיאו ומשפטים המתמקדים באובייקטים)

activity-text embedding space- מתמקד בקישור תיאור וידאו ושפה המדגיש יותר את האירועים בסרטון, ממזג תכונות פעולה ושמע באמצעות שרשור וממפה את התכונה המשורשרת ותכונת הטקסט למרחב המשותף.

1

GT: A man is petting two dogs while holding a guitar.

ResNet: **(24)** a man is standing in front of a microphone holding a violin in one hand and a violin bow in the other.

I3D: (6) A couple of slow lorises are eating fruit.

Proposed: (1) A man pets a couple of dogs.

2

GT: A man is riding a motorcycle in the water at the edge of a beach.

ResNet: **(1)** A man is riding a bike across the waves by the beachside.

I3D: (6) A man on a motorcycle falls into a pool of mud.

Proposed: (1) A person is driving a motorcycle through waves on the shore.

3

GT: Someone wearing blue rubber gloves is slicing a tomato with a large knife.

ResNet: (58) A woman is chopping a red bell pepper into small pieces.

I3D: (18) A cat is eating a small wedge of watermelon.

Proposed: (2) A woman is chopping a red bell pepper into small pieces.

4

GT: A man and a woman are having a phone conversation.

ResNet: (9) A small man is drinking a large goblet of beer.

I3D: (6) The lady tried to wake up the man in costume.

Proposed: (2) The boy hugged the girl.

5

GT: A woman is riding a horse on an open ground.

ResNet: **(13)** A guy is riding a horse.

I3D: (1) The girl rode her brown horse.

Proposed: (1) The girl rode her brown horse.

6

GT: A man is drying off a woman with a towel.

ResNet: (2) Two women are wrestling each other.

I3D: (118) A young woman is putting stickers all over her face.

Proposed: (7) Women are dancing.

7

GT: A man slicing a bun in half with a knife appears to cut himself.

ResNet: (141) Man chops meat and puts it in a plate.

I3D: (7) A man is cutting vegetables.

Proposed: (3) A man slicing the roasted duck.

8

GT: A man pours a plate of shredded cheese in a pot of sauce.

ResNet: (4) Someone is mixing up chocolate batter in a bowl.

I3D: (8) Someone has picked up a handful of white substance from mixing bowl and squeezing it in a lump.

Proposed: (2) A person mixes flour and water in a bowl.

9

GT: Several people are dancing on the patio.

ResNet: (44) A man persuades two ladies standing by the beach to come with him and then the three of them run to join some other people.

I3D: (1) People are dancing together near a house.

Proposed: (3) Many men and women are dancing in street.

NIC (Netural image caption) - Image-Text Retrieval without Joint Embedding

פתיח

המאמר הזה שקראנו היה פחות קשור לjoint embedding אלא שילב בין שני רשתות כלומר הוא יצא בגישה חדשנית של תמונה כ-input לתוך רשת CNN משם ישר עובר לרשת RNN ומשם יוצר משפט המתאר את התמונה שנכנסה כקלט. בעצם יש JOINT של הרשתות עצמם.

בעיות

1. רוב השיטות נתקלות בבעיות האימונים עם מערכי נתונים בקנה מידה קטן, המכסים מספר מצומצם של תמונות עם משפטים אמיתיים.
2. מאוד יקר ליצור מערך נתונים גדול על ידי ביאור מיליוני תמונות עם משפטים.

מטרה

מטרה: בהינתן תמונה, להחזיר משפט שמסווג את התמונה.

עבודות אחרונות הראו באופן משכנע כי רשתות RNN יכולות לייצר ייצוג של תמונת הקלט על ידי הטמעתה לווקטור באורך קבוע, ייצוג זה יכול לשמש על מנת להחזיר משפט שמסווג את התמונה.

במודל זה מציעים להוסיף לפני הRNN, רשת CNN, כך שהפלט שלה יכנס כקלט לRNN.

שיפור של מודלי עבר

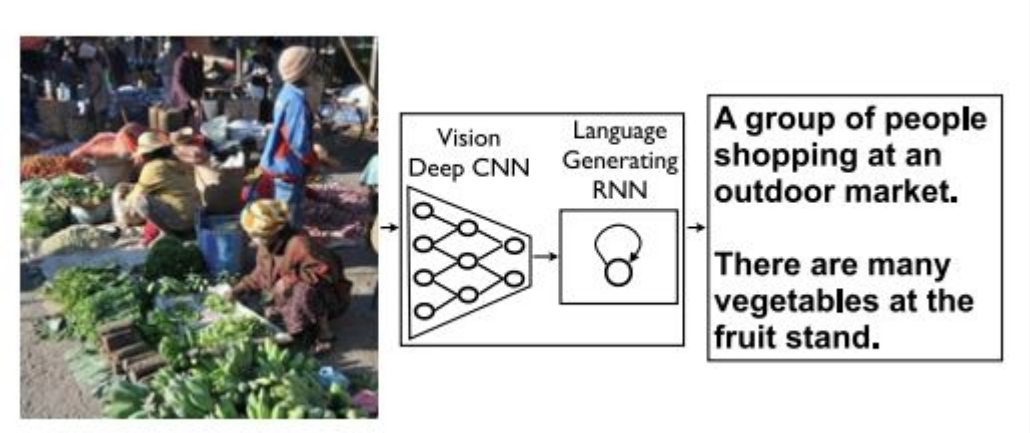
גישות עבר אינן יכולות לתאר חיבורים של אובייקטים שלא נראו בעבר למרות שהאובייקטים מופיעים בנתוני האימון בצורה נפרדת(כל אובייקט בנפרד), בנוסף הם נמנעים מטיפול בהערכת כמה התיאור שנוצר הוא אכן טוב ומדויק.

השוני הוא RNN חזק יותר, וסיפוק הקלט הוויזואלי למודל RNN ישירות מהCNN, מה שמאפשר ל- RNN לעקוב אחר האובייקטים.

יתרונות

- רשת נוירונים הניתנת לאימון מלא באמצעות gradient descent משופר.
- המודל משלב רשתות משנה חדשות למודלים של תמונה ושפה(ניצול של עוד נתונים)
- מניב תוצאות טובות יותר מגישות חדשות אחרות.

ארכיטקטורת הרשת



הסבר של הארכיטקטורה

NIC- מבוסס על רשת נוירונים המקבלת כקלט תמונה ומזין לרשת CNN את תמונה זו. את הוקטור המייצג את התמונה(הפלט של הCNN) המודל מזין לרשת RNN והפלט שמייצרת רשת הRNN הינו משפט שלם בשפה טבעית המתאר את התמונה.

סיכום

הצגנו את NIC, מערכת רשת עצבית מקצה לקצה שיכולה להציג תמונה באופן אוטומטי וליצור תיאור סביר בשפה טבעית. NIC מבוסס על CNN המקודדת תמונה לייצוג דחוס, ואחריה RNN ומייצרת משפט מקביל. המודל מאומן למקסם את הסבירות למשפט בהתחשב בתמונה. ניסויים במספר מערכי נתונים מראים את החוסן של NIC מבחינת תוצאות איכותיות (המשפטים שנוצרו הם סבירים מאוד).

Webly Supervised Joint Embedding for Cross-Modal Image-Text Retrieval

פתיח

על מנת ללמוד, צריך להיות לנו תחילה מאגר נתונים, עדיף שיהיה כמה שיותר גדול. אסיפת מידע ויצירת מאגר נתונים הינם דברים יקרים שלוקחים הרבה משאבים וזמן.

לכן נרצה להשתמש גם במידע זמין באינטרנט, למרות שהוא "רועש".

שלב זה חשוב מאוד כי ככל שהדאטהסט יהיה גדול יותר נוכל להשיג ביצועים טובים יותר.

שיטות קודמות

המטרה שלנו היא בהינתן טקסט, להחזיר תמונה שהטקסט מתאר. שיטות קודמות ניסו גם לפתור את אותה הבעיה, שיטה דומה היא NIC (שם בהינתן תמונה, המטרה היא להחזיר משפט המתאר את התמונה).

בעיות בשיטות קודמות:

- יצירת מערך נתונים גדול, עם זוגות של תמונות ומשפטים היא קשה ביותר ועתירת עבודה (ועקב כך שלב זה הוא כלכך חשוב והכרחי).
- בדרך כלל אפשרי שיהיה רק מספר מוגבל של משתמשים שיאמתו תמונות אימונים, מה שעלול להוביל למודל מוטה.

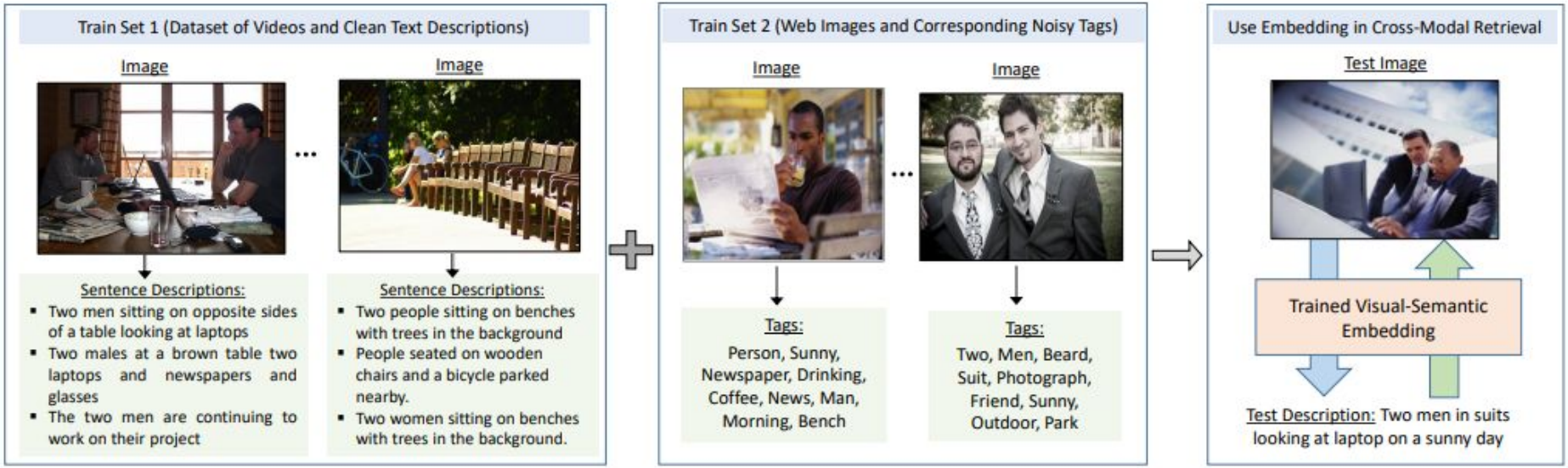
מטרה

המטרה של Webly Supervised Learning היא להשתמש במידע מהאינטרנט (Web) ובנוסף מידע מתויג.

במודל זה אנו רוצים להשתמש במידע מתויג ומידע מהאינטרנט על מנת לסווג תמונות, כלומר אנחנו משתמשים בשני דאטאסטס, אחד מתויג, כלומר הוא מכיל תמונות ומשפטים המתארים את התמונות, והשני נלקח מהאינטרנט, כלומר יש לנו תמונות וtags הקשורים לתמונה.

המכשול הגדול ביותר נובע מתגיות רועשות ומההבדל המהותי בין משפט לתגיות.

דוגמת למידה של הרשת



יתרונות

- לא צריך להשקיע הרבה כסף וזמן באסיפת המידע.
- יש שיפור ברור בביצועים במשימת האחזור.

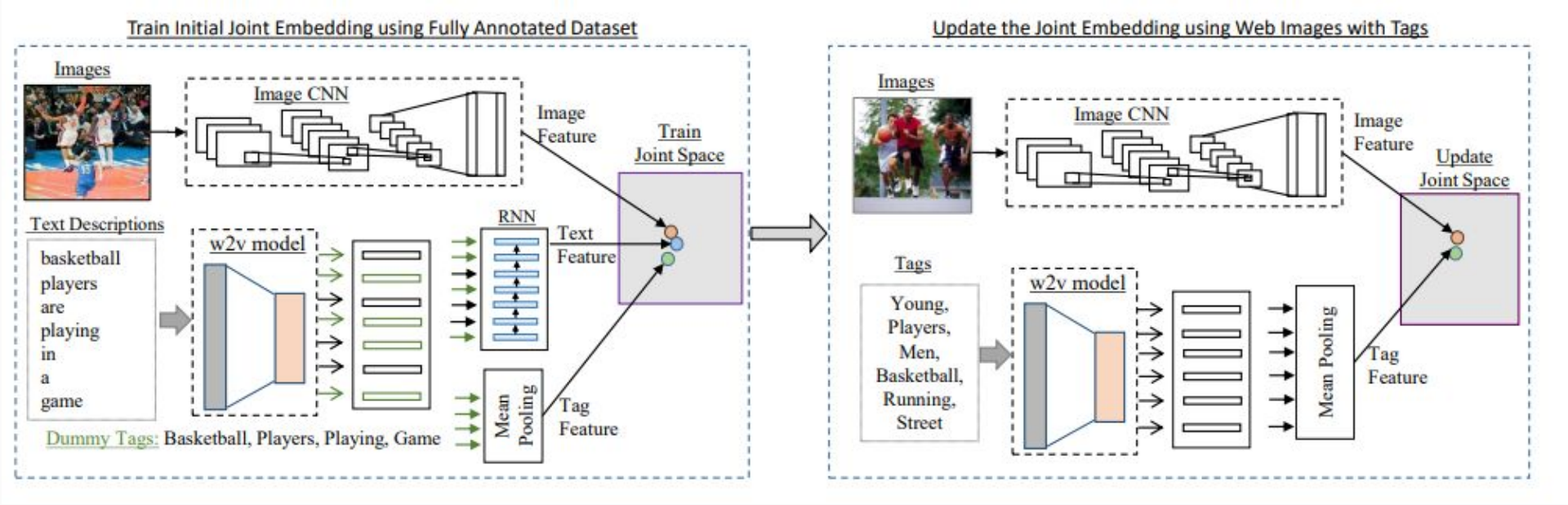
קשיים

ייצוג טקסט דומה ל, ועם זאת שונה מאוד, מייצוג התגיות. משפטים מיוצגים בדרך כלל על ידי Word2Vec עם מודל RNN.

לעומת זאת, לתגיות אין המשכיות כמו למשפט.

ולכן בחלק של הtags נשתמש רק בWord2Vec.

ארכיטקטורת הרשת



ייצוג של הקלט

ייצוג טקסט: לקידוד משפטים משתמשים בGRU. אשר שימשו לייצוג משפטים בעבודות רבות (ראינו דוגמא ברשת שלפני).

ייצוג תמונה: לקידוד התמונה אנו מאמצים מודל CNN עמוק שהוכשר במערך הנתונים של ImageNet. באופן ספציפי, אנו מתנסים במודל ResNet.

ייצוג תגיות: מייצרים ייצוג תכונות של תגיות על ידי Word2Vec של כל התגיות המשויכים לתמונה ואז מנרמלים לפי מספר התגיות.

הסבר על הרשת

ראשית, משתמשים במערך תמונות ובתיאורי המשפט שלהם כדי ללמוד ייצוג Image-Text.
לאחר מכן אנו מעדכנים את הייצוג המשותף באמצעות תמונות אינטרנט ותגים תואמים.

ההטבעה המאומנת משמשת במשימת איחזור Image-Text.

conclusion

- ענף חדשני אשר מתפתח עם השנים עוד ועוד.
- אלגוריתם KNN משותף לרוב המודלים (על מנת למצוא שכן קרוב במרחב המשותף).
- Joint embedding מניב תוצאות טובות

לסיכום, Joint embedding הוא מיוחד כי הוא משלב בין שני סוגים של דאטה לאותו מרחב הטמעה, ובכך הוא מחזיר תוצאות טובות יותר כי סוג דאטא אחד משלים את השני לדוגמא קטע אודיו אחד יכול להתפרש כשני סיווגים שונים, אך באמצעות טקסט תואם, נוכל לבטל את הדו משמעותיות.

ביבליוגרפיה

- https://arxiv.org/pdf/1804.03619.pdf
- https://dl.acm.org/doi/pdf/10.1145/3206025.3206064
- https://ieeexplore.ieee.org/document/7505636?denied=
- https://www.justinsalamon.com/uploads/4/3/9/4/4394963/cramer_looklistenlearnmore_icassp_2019.pdf
- https://arxiv.org/pdf/1808.07793.pdf

תודה על ההקשבה!