ערב טוב חברים ושנה טובה, היום אנחנו עם המהדורה הראשונה של הפינה DeepNightLearners בשנה העברית החדשה עם סקירה קצרה של מאמר בתחום של הלמידה העמוקה.

היום בחרתי לסקור מאמר הנקרא Contrastive Representation Distillation שיצא בערך לפני שנה.

תחומי מאמר: המאמר מהשתמש בשיטה הנקראות Noise contrastive estimation (NCE) המבוסס על מידע הדדי (mutual information) השייכת לתחום למידת הייצוג (representation learning) בשביל הפקת ידע (knowledge distillation - KD) .

הסבר קצר על תחומי השיוך:

קודם כל הלמידת הייצוג זה תחום העוסק בשיטות להפקה ייצוגי מימד נמוך יעילים לדאטה בעלי מימד גבוה. ההנחה המהותית ב NCE הינה שייצוג חזק בהכרח יודע להפריד בין הדוגמא חיובית בהינתן הקונטקסט (הדוגמאות הקשורות או או אותה דוגמא עם אוגמנטציה) לבין דוגמא רנדומלית. בין השימושים של טכניקה זו אפשר להזכיר negative sampling שהשתמשו בו למשל ב- word2vec. במאמר שהציע infoNCE הוכח כי אם ככל הלוס של NCE קטן המידע הדדי בין הדוגמא במרחב המקורי לבין הייצוג של במרחב מימד נמוך עולה שזה כמובן מצביע על אובדן פחות אינפורמציה בין הדאטה לבין הייצוג קרי לייצוג פחות לוסי ויותר מייצג. חשוב לציין שהאימון מתבצע במרחב הייצוג לא במרחב המקורי כלומר הלוס מחושב על הייצוגים במרחב מימד נמוך. לוס NCE זה בעצם עושה הוא לוקח זוג דוגמאות קרובות והרבה דוגמאות רנדומליות ומנסה <u>למקסם את המנה בין דמיון של זוג הקרוב לסכום הדמיונות בינו לבין דוגמאות רנדומליות.</u>

KD זה תחום שהומצא עי" J. Hinton הענק בשלהי 2015. התחום עוסק בצמצום(דחיסה) מודלים גדולים וכבדים חישובית למודלים יותר קלים ונוחים יותר לאינפרנס. רוב השיטות של KD מתחלקות לשתי קבוצות:

1. אימון רשת סטודנט כך שהוא יחקה כמה שיותר טוב את הפלט הרך (לפני הסיגמואיד) של רשת (או רשתות המורה). בדרך כלל משתמשים בקרוס-אנטרופי בין ההסתברויות של רשת המורה ורשת הסטודנט (ההסתברויות מחושבות "תחת טשטוש מסוים הנקרא טמפרטורה", פשוט מחלקים באיזה קבוע את הכניסה לסיגמואיד לאיזה קבוע)

2. שיטות המנסות לגרום לרשת הסטודנט לחקות לא רק את היציאה של השכבה האחרונה של המורה אלא גם יציאות של שכבות ביניים. יש שם מגוון מאוד רחב של שיטות (יש רפרנסים רבים במאמר)

תקציר המאמר:

אז מה שהמאמר הנסקר מציע הוא לאמן רשת הסטודנט כך שהמידע ההדדי בינו לבין הייצוג של רשת המורה תהיה מקסימלית. כמו שאתם יודעים החישוב של המידע הדדי בין הרשתות הוא מאוד קשה, והמחברים מציעים להשתמש ב- NCE (יותר נכון ב infoNCE) בשביל למקסם את המידע הדדי. הרי כמו שהזכרתי ככל ש NCE לוס קטן המידע הדדי עולה (הם מוכיחים את החסם הזה בצורה רגורוזית).

אז מה עושים בתכלס, אתם שואלים? לוקחים דוגמא (חיובית) ומעבירים אותם דרך שתי הרשתות (מורה וסטודנט). אחר כך לוקחים N זוגות של דוגמאות רנדומליות ומעבירים אותן דרך שתי הרשתות גם כן ואז מנסים "להתאים" את NCE לוס למבנה הבאטץ' שלהם (דוגמא אחת חיובית והשאר שליליות - נוסחה 11 במאמר) כלומר לאמן מודל מאוד פשוט המשערך אותו. החישוב מתבצע בצורה מאוד דומה ל infoNCE. ובשלב האחרון מאמנים את המשקלים הרשת הסטודנט בשביל למקסם את מה שיצא בשלב האחרון. כמו הערך המתקבל בסוף מהווה חסם תחתון למידע ההדדי בין ייצוג הסטודנט לייצוג המורה כך שכל התהליך הזה מיועד להעלאתו.

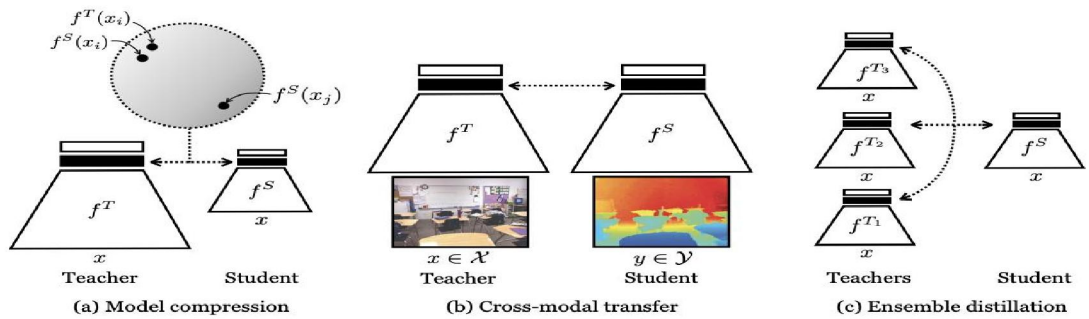

Figure 1: The three distillation settings we consider: (a) compressing a model, (b) transferring knowledge from one modality (e.g., RGB) to another (e.g., depth), (c) distilling an ensemble of nets into a single network. The constrastive objective encourages the teacher and student to map the same input to close representations (in some metric space), and different inputs to distant representations, as indicated in the shaded circle.

הלוס הסופי שלהם מורכב מהלוס המקורי של hinton המבוסס על קרוס אנטרופי בין הייצוגים של הסטודנט ושל המורה והלוס המשערך את המידע ההדדי שהסברתי בפסקה הקודמת

הישגי המאמר: הם מראים את היתרון של השיטה שלהם על מגוון שיטות SOTA עבור שלוש משימות הבאות:

1. דחיסת המודל (נבדקת כאן ירידה בביצועים של הסטודנט יחסית למורה)
2. Cross-modal transfer (לא יודע איך לתרגם את זה לעברית).

המטרה כאן לנסות לבנות רשת למשימה שיש לסטודנט הרבה פחות דאטה מתויג מאשר למורה

3. למידה מכמה מורים (לי זה נראה משימה מאוד לא טריויאלית)

הערה לגבי הביצועים:

מעניין שהשיטה השנייה מבחינת הביצועים הינה KD קלאסי של הינטון ולא כל שיטות שהומצאו ב 4 שנים האחרונות. זה קצת חשוד מוזר.

דאטה סטים: CIFAR-100, ImageNet, STL-10, TinyImageNet NYU-Depth V2

לינק למאמר :מאמר

לינק לקודcode :