MobileNet

ארכיטקטורת רשת עיצוב יעילה ליישומי רשתות ניידים

הקדמה

רשתות מורכבות ועמוקות משיגות דיוקים טובים משמעותית מרשתות קטנות, אבל דורשות הרבה מאוד כוח חישוב של מעבד וזמן ריצה ארוך.
מטרת המאמר היא להגיע לתוצאות כמה שיותר מדוייקות תחת אילוץ של זמן ריצה מהיר ובהתאמה מספר פעולות חישוב נמוך. זאת על מנת לאפשר ריצה של הרשת בטלפונים ניידים או ברכבים אוטונומיים שדורשים תגובה מיידית.
עבודות קודמות מתמקדות בהקטנת הרשת עצמה אך לא בהפחתת המהירות שלה, משתמשות בpretrained models ובknowledge distillation.
לארכיטקטורה המוצעת על ידי המאמר קוראים MobileNet והיא משתמשת בפעולות הבאות כדי להקטין עלויות חישוב:

1. Depthwise Separable Convolution
2. Width Multiplier: Thinner Models
3. Resolution Multiplier: Reduced Representation

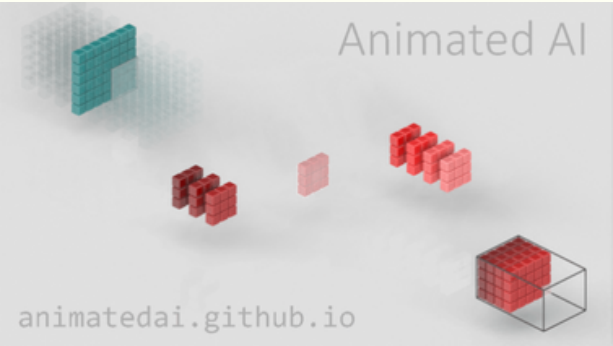
Depthwise Separable Convolution

נשים לב שעלות של קונובולוציה סטנדרטית היא:

$$\underbrace{D_f \cdot D_f}_{\text{Input channels dimension}} \cdot \underbrace{M}_{\text{number of input channels}} \cdot \underbrace{D_k \cdot D_k}_{\text{kernel dimensions}} \cdot \underbrace{N}_{\text{number of output channels}}$$

נרצה להמיר את פעולת הקונבולוציה הסטנדרטית לשתי פעולות שיבצעו אותו דבר בעלות מופחתת.

depthwise convolution:

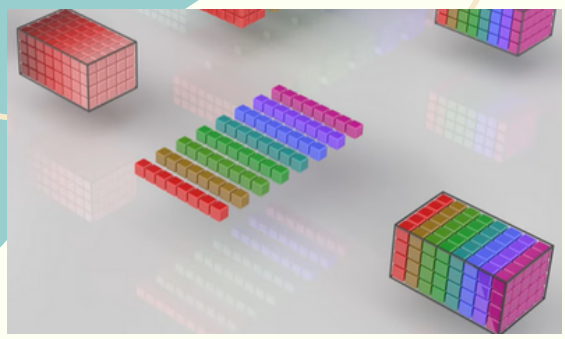


כל פילטר עובר על ערוץ אחד בלבד מהשכבה הקודמת ומוציא ערוץ אחד לשכבה הבאה.
עלות החישוב תהיה הבאה:

$$\underbrace{D_f \cdot D_f}_{\text{Input channels dimension}} \cdot \underbrace{M}_{\text{number of input channels}} \cdot \underbrace{D_k \cdot D_k}_{\text{kernel dimensions}}$$

לאחר פעולת depthwise convolution המידע לא מתערבב בין הערוצים ולכן נשתמש בpointwise convolution.

pointwise convolution:



העברת קרנלים בגודל 1x1 שמערבבים את המידע
של כל פיקסל מכל הערוצים בשכבה הקודמת.
עלות החישוב תהיה הבאה:

$$\underbrace{D_f \cdot D_f}_{\text{Input channels dimension}} \cdot \underbrace{M}_{\text{number of input channels}} \cdot \underbrace{N}_{\text{number of output channels}}$$

נשים לב שאכן עלות החישוב הכוללת בביצוע הפיצול לשתי הפעולות קטנה משמעותית מעלות החישוב של קונבולוציה סטנדרטית. נקרא לפעולה זו Depthwise Separable Convolution.

Width Multiplier: Thinner Models

הפרמטר α מוגדר להיות בין 0 ל1 ומקטין את מספר ערוצי השכבה הקודמת ואת מספר ערוצי השכבה היוצאת פי α.
כך הסיבוכיות קטנה בערך פי α בריבוע.

Resolution Multiplier: Reduced Representation

הפרמטר ρ מוגדר להיות בין 0 ל1 ומקטין את רזולוציית השכבה הקודמת פי ρ.
כך הסיבוכיות קטנה בערך פי ρ בריבוע.

תוצאות

בשימוש בפעולת הקונבולוציה החסכונית וכן בתוספת שני הפרמטרים, כותבי המאמר הצליחו למצוא מודל שמייעל מאוד את המהירות והגודל של הרשת, כאשר אחוזי הדיוק דומים או קטנים מעט מביצועים של רשתות אחרות, אבל מספר פעולות החישוב קטן בצורה משמעותית מאוד.
כלומר הורדת המימדים נעשתה בצורה חכמה שתפגע לכל הפחות באחוזי הדיוק.

ShuffleNet

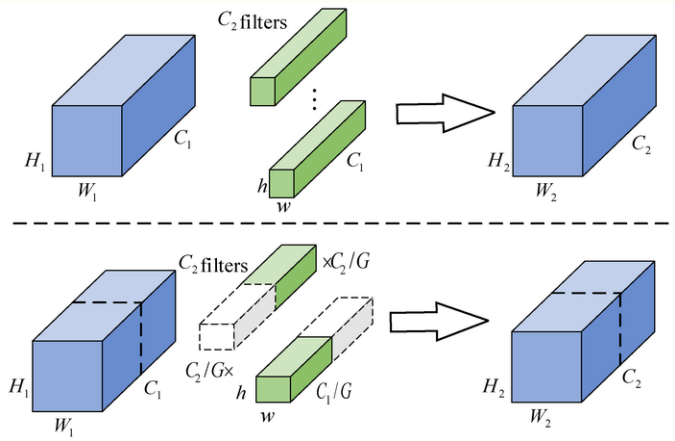
ארכיטקטורת רשת נוירונים יעילה למכשירים ניידים

הקדמה

רשתות מורכבות ועמוקות משיגות דיוקים טובים משמעותית מרשתות קטנות, אבל דורשות הרבה מאוד כוח חישוב של ביליונים של FLOPs.
בשל הצורך בהפעלת רשתות במכשירים ניידים, עיצובי מודלים יעילים תפסו תאוצה בשנים האחרונות. המודלים מתמקדים בהקטנת גודל הקלט והסרת קשרים בין ערוצים כדי לזרז את הרשת, וכן בהפחתת מספר הפילטרים ובכך גם הערוצים היוצאים משכבת הקונבולוציה כמו שנעשה בMobileNet.
המאמר שלנו מעוניין להגיע לתוצאות כמה שיותר מדוייקות בכוח חישובי מוגבל בעשרות עד מאות MFLOPs, ע"י ייצור של יותר ערוצים בשכבת הנוירונים הבאה באותו כוח חישובי.
ייצור של יותר feature map channels עוזר לקודד יותר מידע וכך הרשת תוכל לעבד את הנתונים במידה טובה יותר.
ייצור מוגבר של ערוצים בשכבת הפלט מתבצע ע"י 2 פעולות:

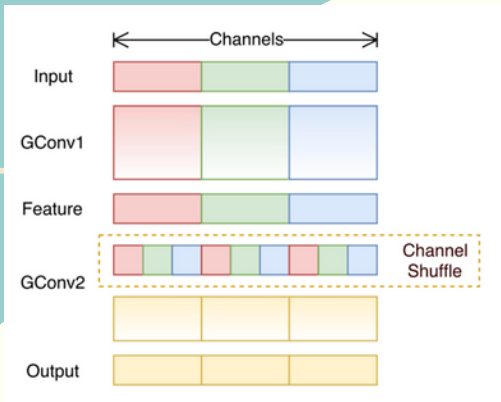
1. Group convolution
2. Channel shuffle

Group Convolution



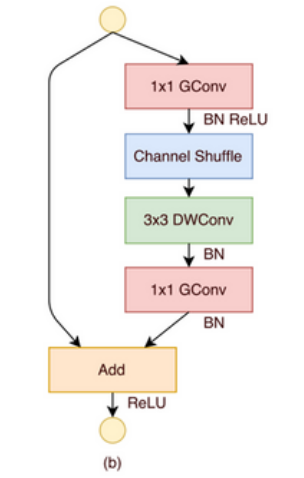
במקום שכל קרנל יעבור על כל השכבה הקודמת ברשת, השכבה הקודמת מפוצלת למספר קבוצות וכל קרנל עובר רק על אחת מהקבוצות, כלומר רק על חלק מהערוצים של השכבה הקודמת.
קונבולוציה בדרך כזו מפחיתה חפיפות במידע כיוון שכל פילטר מתאים את עצמו בצורה לוקאלית למספר ערוצים במקום ללמוד את כל הערוצים.
מכיוון שבצורה כזו כל פילטר מבצע פחות חישובים כי עובר על פחות ערוצים, נוכל באותו כוח חישוב נתון ומוגבל להעביר יותר פילטרים וככה להוציא יותר ערוצים לשכבה הבאה (כיוון שכל פילטר מוציא ערוץ לשכבה הבאה).
פילטרים נוספים וערוצים נוספים בשכבת הקלט מהווים עיבוד נתונים ולמידה משמעותית יותר.
נשים לב שנותרנו עם בעיה אחת – אין עירבוב של נתונים בין הקבוצות.
הפעולה הבאה פותרת לנו את הבעיה.

Channel Shuffle Operation



לאחר סיום פעולה של group convolution, נערבב את הערוצים כך שכל קבוצה בשכבת הקונבולוציה הבאה תחזיק קבוצת ערוצים שונים ממה שהחזיקה בקונבולוציה הקודמת. הפעולה פותרת את בעיית הנורמליזציה, שכן כך כל קבוצה תקבל נתונים גם מפיצ'רים שהיו בקבוצות אחרות. ככל הנראה כל הקרנלים יעבדו על כל הערוצים.

ארכיטקטורה של הרשת



הרשת בנויה משכבות קונבולוציה ובתוכן 3 שלבים של ערימות של יחידות של shuffleNet.
כל יחידה בנויה במבנה של bottleneck unit עם ResNet, כלומר ענף אחד מעביר את הנתונים משכבה לשכבה וענף שני מבצע Group convolution 1x1, לאחריו channel shuffle, לאחר מכן קונבולוציה רחבה יותר ואחריה שוב Group convolution 1x1.
דרך נוספת לבניית יחידה אחת היא תוספת של average pooling עם stride=2 .

תוצאות

מתוצאות המחקר אנחנו רואים השפעה גדולה יותר באחוזי הדיוק של רשתות קטנות בשימוש בShuffleNet.
לרשתות קטנות אין מספיק ערוצים לעיבוד המידע ולכן התוצאות שלהן פחות מדוייקות.
אך מכיוון שבהינתן תקציב חישובי, ShuffleNet יכולה להשתמש בfeature maps רחבות יותר ובכך לקודד מידע נוסף, השינוי באחוזי הדיוק יהיה משמעותי כי הרשת מצליחה לעבד את הנתונים בצורה טובה יותר משמעותית.
גם בהשוואה לMobileNet, רשת ShuffleNet מגיעה לאחוזי דיוק טובים יותר בכמעט 8% ברשתות קטנות מאוד.