

מעבדה תרגיל 2

נועה ליפשיץ 318262938 נעם שפריי 318296951

קישור לגיט: https://github.com/noalif/094295_hw2/tree/master

עיבוד מקדים לדאטה (Preprocessing):

בשלב זה בדקנו ידנית את הדאטה במטרה לאתר ולהסיר דוגמאות לא תקינות וגם להעביר דוגמאות שאינן משוייכות לתיקייה הנכונה (תיקון תיג שגוי). לאחר שהתיגים תוקנו ודוגמאות לא רלוונטיות נמחקו, שמנו לב שישנם תיגים בעלי מספר נמוך משמעותית מתיגים אחרים. בעקבות זאת החלטנו לשים לב, בשלב העשרת הנתונים, ליצור איזון בין המחלקות. השיטה שבחרנו היתה להכפיל כל תמונה מספר פעמים כך שבסופו של דבר החלוקה בין סט האימון והולידציה היה בקירוב 20-80. מכיוון שהדאטה סט גדל מאוד, חוסר האיזון ההתחלתי נהיה פחות משמעותי (מחלקות שנעות בין 600-1000 דגימות)

המחשה לדוגמאות שנמחקו מסט הנתונים:



עקרונות מנחים בבחירת אוגמנטציות להעשרה סט הנתונים:

כאשר עברנו על סט הנתונים שמנו לב למספר אוגמנטציות שבאו לידי ביטוי בדאטה עצמו, כמו למשל: translation, scaling, החסרת חלקים מהמספר ועוד. לכן, בחרנו להתחיל קודם עם האוגמנטציות שראינו בדאטה עצמו במטרה לשמר את ההתפלגות המקורית של הדוגמאות. במקביל, הוספנו אוגמנטציות נוספות כדי להפוך את הדאטה שלנו ליותר רובסטי כך שיוכל להכליל טוב יותר דוגמאות מסט מבחן שלא יגיע בהכרח מההתפלגות הרגילה של הנתונים (למשל shear - מתיחה ומריחה של התמונה).

בתחילה בדקנו עיבוי של הדאטה על ידי יצירת תמונות חדשות שעברו אוגמנטציה אחת בלבד ולאחר מכן התקדמנו לשילוב של אוגמנטציות.

חלק מהאוגמנטציות בוצעו על כל סט הנתונים וחלק הותאמו במיוחד עבור מחלקות מסוימות כדי לוודא שאוגמנטציות מסוימות לא משנות את הספרה, למשל שילוב של flip עם rotate שהופך את v_i ל- v_{iv} ולחלופין.

השתמשנו בחבילה albumentation בעלת api נוח לויזואליזציה: <https://demo.albumentations.ai>.
מה שאפשר לנו לראות איך האוגמנטציות והפרמטרים משפיעים על התמונה, ובחרנו אוגמנטציות ששומרות על כך שניתן יהיה לזהות את הספרה, אך משנות כמה שיותר את התמונה כדי להרחיב את הדאטה שלנו לדוגמאות חדשות משמעותיות.

תמונה להמחשה:

Select the interface mode

☐ Simple

☒ Professional

Select an image:

Upload my Image

Upload your image (jpg, jpeg, or png)

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG

Browse files

aec13410-ce5d-11eb-b317-...
1.8KB

Select transformation №1:

Downscale

Select transformation №2:

None

Params of the Downscale

scale_min & scale_max

0.25

0.01 0.99

interpolation

☒ 0

☐ 1

☐ 2

☐ 3

Demo of Albumentations








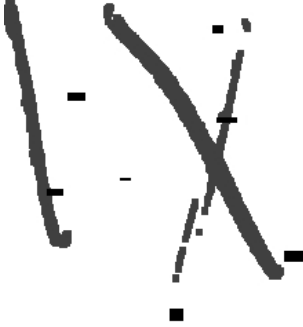


Original image



Transformed image

האוגמנטציות שבחרנו:

מתוך רשימה זו הגרלנו 7 פעמים שילוב של 3 אוגמנטציות שהופעלו על סט האימון המקורי, עבור כל תמונה.

תמונה	אוגמנטציה	תמונה	אוגמנטציה
	rotate - השתמשנו בסיבוב של 45- עד 45 מעלות על כל סוגי המספרים.		original
	translation - השתמשנו פה בסיבוב של 50- עד 50 מעלות, הזזה ושינוי מימד.		blur טשטוש של התמונה במטרה ליצור תמונות שונות אך עדיין נתנות לזיהוי כספרה המקורית
	grid distortion - במטרה שהמודל ילמד קצוות אלו ולא יתבלבל ויחשוב שזה חלק מהספרה.		dropout הורדת חלקי תמונה קטנים שעדיין משמרים את הזיהוי של הספרה. דאגנו שהחלקים החסרים לא יהיו גדולים מדי.
	down scale בדומה לblur, שינוי משמעותי לתמונה כך שעדיין אפשר לזהות את הספרה		gaussian noise הוספת רעש גטוסיאני. אוגמנטציה קלאסית שידוע שעוזרת להכללה.

לכל תמונה בחרנו לעשות שילוב של 3 טרנספורמציות מתוך השבע שהוצגו לעיל.
דוגמה לקומבינציה:

grid distortion + blur + down scale:



gaussian noise + translation + drop out



באחד מהניסיונות ההתחלתיים ראינו שהאוגמנטציה שהיא שילוב של shear + scale + rotate שיפרה את התוצאות, ולכן הוספנו אותה כעוד אוגמנטציה קבועה לכל התמונות.



חלוקת הדאטה לסט אימון וסט ולידציה:

תחילה הרצנו את המודל עם החלוקה שקיבלנו (תמונה אחת לכל תיוג) וקיבלנו תוצאות גבוהות (מעל 90 אחוז דיוק), זאת כמובן משום שהולידציה הייתה קטנה מאוד ולא ייצגה את המגוון הרחב האפשרי של המספרים. לאחר מכן ניסינו לחלק את הדאטה למספר אפשרויות אך לא ראינו הבדל משמעותי מספיק ביניהם ולכן בחרנו לרוב בחלוקה של 80-20 הסטנדרטית.

תוצאות:

בדקנו שני כיוונים כאשר חילקנו את הדאטה לסט אימון וולידציה, קודם ביצענו אוגמנטציות לכל האימון, ואז חילקנו אותו רנדומלית 80-20 לאימון וולידציה. בדרך זו הגענו לaccuracy של 0.929. אנו משערות שהגענו לדיוק גבוה זה מכיוון שסיכויים גבוהים שנלקחו לולידציה תמונות דומות יחסית לאלו שבסט האימון. זאת מכיוון שהאוגמנטציות דומות לתמונות המקוריות ולכן כמעט "הכפלנו" את התמונות ואז השתמשנו בהם לולידציה.

לכן החלטנו לחלק קודם את הדאטה לסט אימון וולידציה 80-20 ורק לאחר מכן לבצע את האוגמנטציות רק על סט האימון.

להלן תיאור הניסיונות:

ניסיון 1 - בדיקת תוצאות בסיס ללא אוגמנטציות:

חילקנו את התמונות 80-20 והרצנו את המודל 20 אפוקים.

sum train: 1593

sum val: 403

sum images: 1996

Best val Acc: 0.811414

ניסיון 2:

קודם עשינו אוגמנטציות ספציפיות, אח"כ חילקנו לאימון וולידציה

sum train: 6783

sum val: 1704

sum images: 8487

Best val Acc: 0.933099

- הסיבה לתוצאה הגבוהה מתוארת בפסקה הראשונה של התוצאות.

ניסיון 3:

בחרנו קומבינציה של 3 אוגמנטציות רנדומליות מתוך סט האוגמנטציות שתיארנו לעיל. חזרנו על תהליך הבחירה n פעמים לכל תמונה עד שקיבלנו כ- 10000 תמונות. (למשל, אם הגדלנו רק את סט האימון, ובאופן התחלתי מספר התמונות היה כ-2000, חילקנו את הדאטה ל-1000 תמונות לסט ולידציה ו-1000 תמונות לסט האימון והכפלנו כל תמונה בסט האימון 8 פעמים, כך שלבסוף יש כמעט 10,000)

בדיקה ראשונה - לחלק את הסט חצי חצי, ולחצי של האימון לעשות אוגמנטציה 8 פעמים לכל תמונה, כך שלבסוף הדאטה היה מחולק 10-90. הרצנו את המודל 15 אפוקים (ראינו התכנסות בזמן הזה)

sum train: 8955

sum val: 1001

sum images: 9956

Best val Acc: 0.888112

בדיקה שנייה - חילקנו את הדאטה 20-80 וביצענו לכל סט בנפרד את האוגמנטציות הרנדומליות. 4 פעמים לכל תמונה. הרצנו 10 אפוקים

sum train: 7965

sum val: 2015

sum images: 9980

Best val Acc: 0.756328

הרצנו שוב עם 20 אפוקים, הפעם הגענו ל- 0.97 על האימון

Best val Acc: 0.764764

- תוצאות אלה הרבה פחות טובות מאשר התוצאות הקודמות וגם מאשר התוצאות מהניסיון הראשון ללא האוגמנטציות. ההשערה שלנו היא שבעקבות ביצוע אוגמנטציות באופן רנדומלי על סט האימון ועל הולידציה התקבלו אוגמנטציות שונות בולידציה מאשר באימון (או שבאימון התקבלו אוגמנטציות יחסית דומות אך בכמות נמוכה מאוד ולא מספיק מייצגת את סט הולידציה) ולכן התוצאות הנמוכות. עם זאת, התוצאה הנמוכה התקבלה שלוש פעמים (פעם נוספת אחרי הוספה של עוד אוגמנטציה) ולכן אולי קיימת סיבה נוספת לכך שאם יש אוגמנטציות על סט הולידציה הביצועים נמוכים.

נסיון 4:

החלטנו להוסיף עוד שילוב של אוגמנטציה לכל תמונה מפני שראינו שהוא משפר תוצאות.
האוגמנטציה: shear + scale + rotate (בנוסף בחירת אוגמנטציות בצורה רנדומלית):

sum train: 7960

sum val: 1001

sum images: 8961

Best val Acc: 0.890110