כתב יד, אחזור וקריאה בעברית ובערבית

דרור בן דוד ועידן בר

רקע על המוצר

החברה פיתחה יישום OCR המבוסס על ארכיטקטורה של Encoder-Decoder וביניהם מנגנוני Attention ו-BILSTM לקידוד Sequence.

פיתוח היישום נמשך למעלה משנה והוא כלל פיתוח יכולת "לסנטז" כתב בעברית ובערבית במגוון פונטים וסוגי הרעשות – "תפור" לסוגי ההרעשות שהלקוח פוגש. היכולת לאמן את המערכת נשענת על מחקרי עומק של חומרי הלקוח, כולל ניתוחים סטטיסטיים של תדירות הופעת מילים מסוימות, התמודדות עם כותרות, טבלאות, הערות שוליים, משפטים באנגלית שהשתרבבו לטקסט וכיוצא באלו.

היישום שולב במערכות הלקוח, כולל התאמת הפתרון למערכות ההפעלה של הלקוח ויכולותיו בתחום תמיכה בתשתיות. המעבר לסביבת הלקוח כלל ביצוע בדיקות על חומרים שלו, ביצוע Fine Tuning וכיוצא באלו.

על בסיס ממלאי התפקיד אצל הלקוח – המערכת נותנת ביצועים טובים יותר מכל מערכת אחרת שהלקוח הצליח למצוא בעולם (סד"ג 93% סיכויי הצלחה בפענוח מילה בערבית).

הקבוצה פיתחה גם יכולת איחזור סמנטית. יכולת זו נשענת על "אמבדינגס" המתקבלים מאימון על קורפוסים ענקיים שנמצאים ברשות החברה ואימון מטויב לקורפוס הרלוונטי ללקוח המסוים, כולל, ככל שנדרש, שימוש במאגרי מידע ייעודיים של הלקוח.

השילוב של יכולת איחזור סמנטית – והיכולת "למדוד מרחק" בין מסמך מסוים לאחר – במרחב המקודד, אפשרו לנו לבנות "מכונה/ יישום" שיודע לקבל מיליוני מסמכים, "לקרוא" אותם באמצעות OCR (בעברית ובערבית), "להבין אותם" סמנטית – ולסווג אותם לקטגוריות – על בסיס המרחק של המסמך המסוים, במרחב המקודד, אל המסמכים מאותה קטגוריה. כך למשל, "המכונה" יודעת לסווג חשבונית שצולמה עם טלפון סלולרי – לעומת צילום של תעודת זהות שבוצע באמצעות מכונת העתקה איכותית או חוות דעת משפטית, הערכה סרוקה של שמאי וכיו"ב.

יכולת זו משולבת, כשירות בענן, אצל לקוח שמנהל למעלה ממאה וחמישים מיליארד ש"ח בישראל.

צרוף היכולות שתוארו מאפשר גם יישומים כמו למשל קריאת חשבוניות, אימות פרטים שמולאו בכתב יד וכיוצא האלו.

בימים אלה, קבוצת המחקר בוחנת את היתרונות במעבר לרשתות מטיפוס Transformers.

בנוסף לתחומי העיסוק שתוארו למעלה – החברה פיתחה ומפתחת מגוון פתרונות בתחום "הבנת שפה טבעית", בעברית ובערבית ואולם נושא זה אינו בתכולת הפגישה הקרובה.

בפגישה הקרובה עם פורום SIGTRS נציג בקצרה את הפיתוחים השונים ואת האתגרים איתם צריך להתמודד כאשר מיישמים יכולות כאלה. נדבר על התרומה הפוטנציאלית של Transformers ועל התרומה העצומה, האפשרית, של יכולות כאלה – לשיפור השירות לאזרחים (בגופים ממשלתיים) ולשיפור התוצאות העסקיות ומנגנוני ROI, בגופים עסקיים.

בין השאר, נדבר גם על מודלים מתקדמים של "מו"פ משולב לקוחות" – שיטה שבה הפיתוח נעשה ע"י קבוצות משימה המורכבות הן מהחוקרים ומהפתחים שלנו והן מצוותי IT (ומו"פ – אם יש כאלה), של אנשי הלקוח - באתר הלקוח, תוך מינוף מקורות המידע שלו ובמתודולוגיית פיתוח "אג'ילית" שמאפשרת שחרור "בלוקים קטנים" אך תדירים, של יכולות – אל הסביבה התפעולית/ מבצעית של הלקוח.

ככל שהזמן ירשה, נדבר גם אל אתגרים בתחום השילוב בתהליכי לקוח, קימום שירותי ענן, התמודדות עם תלויות (Dependencies) ויכולות Scale-Up "חצי אוטומטיות", מה זה ארכיטקטורה תומכת DevOps ועוד.

רקע על החברה

חברת "מטריקס איי טי" היא חברה ציבורית, נסחרת בתל אביב ועובדים בה כ 11,000 איש.

"מטריקס דיפנס" היא חטיבה בתוך מטריקס וחברת בת בפני עצמה.

במטריקס דיפנס יש כ 600 עובדים והיא עוסקת בייזום וניהול פרויקטים מגוונים ובמו"פ, בעיקר בתחומי סייבר ובינה מלאכותית.

קבוצת "הבינה המלאכותית" נקראת "בוסטון", היא הוקמה לפני שש שנים ויש בה כיום ***כחמישים חוקרים***. הקבוצה מתמחה בתחום "למידה עמוקה"/ רשתות נוירונים רבודות והיא השלימה עד היום כשישים מחקרים בתחומים שונים. חלק מהמחקרים הפכו למוצרים המשולבים במערכי המיחשוב של הלקוחות וחלקם מונגשים כשירות בענן.

הקבוצה פיתחה, בין השאר, יכולות בתחום OCR ואיחזור, בעברית ובערבית, כולל בכתב יד ונכון להיום, האלגוריתמים של החברה "פותרים" למעלה ממיליון מסמכים בחודש אצל שני לקוחות עיקריים.

החברה עוסקת ברשתות נוירונים רבודות מסוגים שונים ובכללם CNN, RNN, GAN, AUTOENCODERS, TRANFORMERS ועוד.

פעילות החברה ממוקדת למתן מקסימום ערך ללקוחות – כשירות ומכיוון שכך, כל פרויקט מתחיל בסריקת כל מה שזמין בעולם, בדגש לאוניברסיטאות המובילות והחברות הגדולות, יצירת קשר עם צוות המפתחים ותיחור בין צוות פיתוח פנימי לפתרונות שמוצאים בעולם – Best Available Technologies.