security.googleblog.com /2023/11/improving-text-classification.html

Gmail se dote d'une nouvelle arme pour contrer les courriels indésirables

Améliorer la résilience et l'efficacité de la classification de texte avec RETVec

Elie Bursztein, Cybersécurité & Directrice de recherche en IA et
Marina Zhang, ingénieure logicielle

Google va-t-il mettre fin aux courriels de spams sur son service de messagerie Gmail? Gmail déploie une nouvelle technologie; RETVec, qui promet de nettoyer nos boîtes de réception des courriels indésirables.
Avec cette avancée, Google ouvre une nouvelle ère dans la lutte contre les courriels indésirables (spams), en utilisant une approche innovante basée sur l'intelligence artificielle.

Des systèmes tels que Gmail, YouTube et Google Play s'appuient sur des modèles de classification de texte pour identifier les contenus nuisibles, notamment les attaques de phishing, les commentaires inappropriés et les escroqueries.

Ces types de textes sont plus difficiles à classer pour les modèles d'apprentissage automatique, car les mauvais acteurs s'appuient sur des manipulations de texte contradictoires pour tenter activement d'échapper aux classificateurs.

Par exemple, ils utiliseront des homoglyphes, des caractères invisibles et du bourrage de mots clés pour contourner les défenses.

Pour rendre les classificateurs de texte plus robustes et efficaces, nous avons développé un nouveau vecteur de texte multilingue appelé RETVec (Resilient & Efficient Text Vectorizer) qui aide les modèles à atteindre des performances de classification de pointe et réduit considérablement les coûts de calcul.
Aujourd'hui, nous expliquons comment RETVec a été utilisé pour protéger les boîtes de réception Gmail.

Renforcer le classificateur de spam Gmail avec RETVec

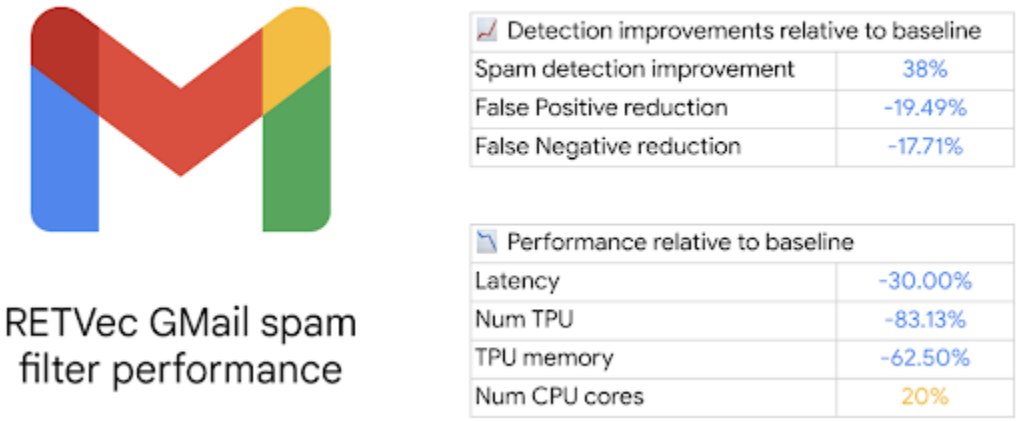

Figure 1. Améliorations du filtre anti-spam Gmail basé sur RETVec.

Au cours de l'année écoulée, nous avons testé RETVec de manière approfondie au sein de Google pour évaluer son utilité et avons constaté qu'il était très efficace pour les applications de sécurité et anti-abus.
En particulier, le remplacement de l'ancien vecteur de texte du classificateur de spam Gmail par RETVec nous a permis d'améliorer le taux de détection du spam par rapport à la référence de 38 % et de réduire le taux de faux positifs de 19,4 %.
De plus, l'utilisation de RETVec a réduit l'utilisation du TPU du modèle de 83 %, faisant du déploiement de RETVec l'une des plus grandes mises à niveau de défense de ces dernières années.
RETVec réalise ces améliorations en arborant un modèle d'intégration de mots très léger (~ 200 000 paramètres), nous permettant de réduire la taille du modèle Transformer à des performances égales ou supérieures, et en ayant la possibilité de diviser le calcul entre l'hôte et le TPU dans un réseau et une mémoire. manière efficace.

Avantages de RETVec

RETVec réalise ces améliorations en combinant un nouveau codeur de caractères très compact, un programme d'entraînement basé sur l'augmentation et l'utilisation de l'apprentissage métrique.
Les détails de l'architecture et les évaluations de référence sont disponibles dans notre article NeurIPS 2023 et nous RETVec open source sur Github.< /span>

Grâce à son architecture novatrice, RETVec fonctionne immédiatement sur toutes les langues et tous les caractères UTF-8 sans nécessiter de prétraitement du texte, ce qui en fait le candidat idéal pour les

déploiements de classification de texte sur appareil, sur le Web et à grande échelle.
Les modèles entraînés avec RETVec présentent une vitesse d'inférence plus rapide en raison de sa représentation compacte.
Avoir des modèles plus petits réduit les coûts de calcul et diminue la latence, ce qui est essentiel pour les applications à grande échelle et les modèles sur appareil.

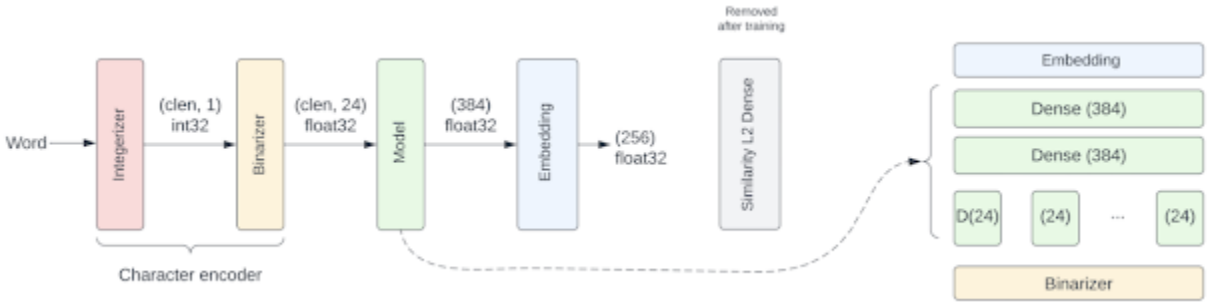

Figure 1. Schéma d'architecture RETVec.

Les modèles entraînés avec RETVec peuvent être convertis de manière transparente en TFLite pour les appareils mobiles et périphériques, grâce à une implémentation native dans TensorFlow Text.
Pour le déploiement du modèle d'application Web, nous fournissons une implémentation de couche TensorflowJS disponible sur Github et vous pouvez consulter une page Web de démonstration exécutant un modèle basé sur RETVec. .

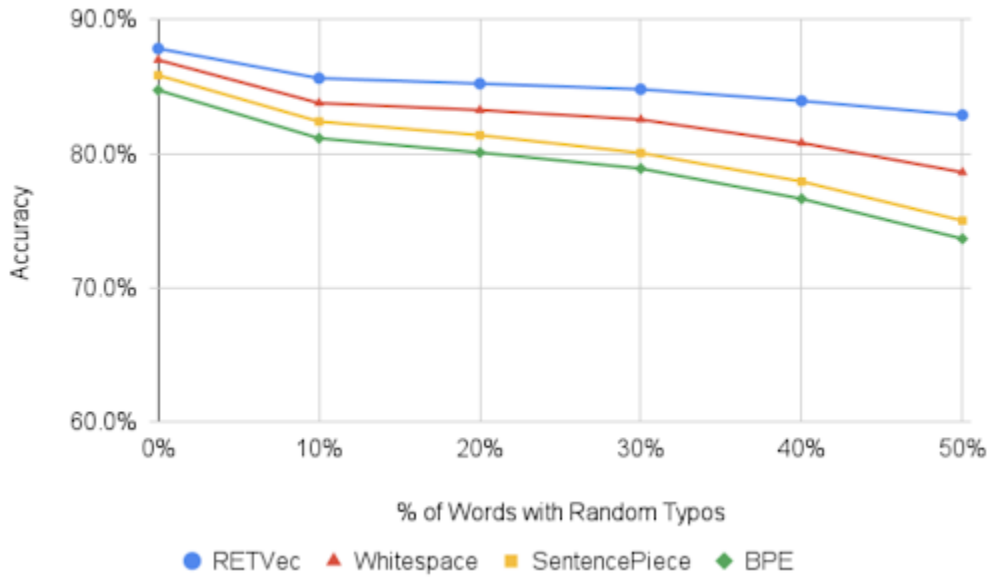

Figure 2. Résilience aux fautes de frappe des modèles de classification de texte entraînés à partir de zéro à l'aide de différents vectoriseurs.

RETVec est un nouveau vecteur de texte open sourcequi vous permet de créer des classificateurs de texte plus résilients et plus efficaces côté serveur et sur l'appareil.
Le filtre anti-spam Gmail l'utilise pour protéger les boîtes de réception Gmail contre les e-mails malveillants.

Si vous souhaitez utiliser RETVec pour vos propres cas d'utilisation ou recherches, nous avons créé un tutoriel pour vous aider à démarrer.
Si vous souhaitez utiliser RETVec pour vos propres cas d'utilisation ou recherches, nous avons créé un tutoriel pour vous aider à démarrer. une>

Cette recherche a été menée par Elie Bursztein, Marina Zhang, Owen Vallis, Xinyu Jia et Alexey Kurakin.
Nous tenons à remercier Gengxin Miao, Brunno Attorre, Venkat Sreepati, Lidor Avigad, Dan Givol, Rishabh Seth et Melvin Monténégro ainsi que tous les Googleurs qui ont contribué au projet.

Recherche et mise en page par:
Michel Cloutier
CIVBDL
20231214
"C'est ensemble qu'on avance"