Classificação de intenções multimodal com modalidades incompletas usando propagação de embedding de texto

Victor Machado Gonzaga

Nils Murrugarra-Llerena

Ricardo Marcondes Marcacini

Universidade de São Paulo

machado.prx@usp.br

Objetivos

Neste trabalho, apresentamos um método de propagação de embedding de texto para classificação de intenção em documentos multimodais para lidar com modalidade textual incompleta. Propomos uma representação baseada em grafo para documentos de modelo multimodal. Cada vértice representa um documento multimodal e pode ter dois vetores de informação associados: (1) embeddings de imagem obtidos por uma rede residual pré-treinada [1] e (2) embeddings de texto obtidos por um modelo de linguagem BERT pré-treinado [2]. Assim, utilizamos um método transdutivo de aprendizagem de grafos para propagar embeddings de texto a partir dos vértices que contêm texto. A estrutura de grafos multimodal ajuda a aprender embeddings de texto para vértices que possuem apenas dados de imagem. Desse modo, um nó que possui apenas recursos visuais pode receber recursos textuais por meio de um processo que propaga iterativamente dados entre nós vizinhos. Após a propagação dos embeddings, todos os vértices terão modalidades completas e então qualquer método de classificação de documento multimodal pode ser usado. Realizamos uma avaliação experimental de classificação de intenção em um conjunto de dados de mídia social multimodal. Comparamos nosso método proposto com dois outros métodos: (1) um classificador de documentos usando apenas as imagens do post; e (2) um método de última geração que combina imagens e textos para classificação de intenções com modalidades completas. Obtivemos resultados competitivos, mesmo na presença de modalidades incompletas.

Métodos e Procedimentos

A abordagem proposta usa um kernel gaussiano como uma função não linear da distância euclidiana para calcular os pesos W do grafo e assim obter relações de vizinhança entre os documentos. Dados dois vértices xi e xj, o peso Wij é calculado de acordo com a Eq. 1, onde xi e xj são os respectivos vetores de características e sigma é um parâmetro de escala de comprimento. Uma matriz de afinidade normalizada S é calculada a partir de W, conforme definido na Eq. 2. D é uma matriz diagonal onde o elemento (i, i) é igual à soma da i-ésima linha de W.

$$Wij = \exp(-\frac{|xi - xj^2|}{2\,\sigma}) \quad (1)$$

$$S = D\hat{}(\frac{1}{2})WD\hat{}(\frac{1}{2}) \quad (2)$$

Agora, usamos a representação baseada em grafo gerada para propagar embeddings de texto entre vértices. Nosso método é inspirado

no aprendizado transdutivo baseado em grafos proposto por [3] para tarefas de classificação semi supervisionadas usando propagação de rótulos. Estendemos o método para propagação de embeddings de texto. O método visa obter uma nova representação textual F para todos os documentos multimodais, considerando tanto a estrutura do grafo (representado por matriz de afinidade S) quanto às características de texto existentes.

Resultados

O método proposto obtém métricas ACC e AUC maiores do que a linha de base (apenas imagens), o que indica que considerar algum nível de informação textual aumenta o desempenho de classificação da intenção do autor, mesmo com modalidades muito incompletas, conforme mostrado em Img + Txt-BERT (20%). Quando consideramos cenários com 60% e 80% de postagens com modalidade textual, o método proposto atinge resultados competitivos nas métricas ACC e AUC. Consideramos esses cenários os mais próximos dos aplicativos do mundo real envolvendo redes sociais baseadas em imagens, uma vez que uma porcentagem significativa de textos é descartada por serem muito curtos, não fornecidos ou sem sentido.

Método	ACC	AUC
Probabilidade / Classificador Aleatório	28.1	50.0
Img (Linha de base) [He et al. 2016]	42.9 (±0.0)	76.0 (±0.5)
Img+Txt-ELMo [Kruk et al. 2019]	56.7 (±0.0)	85.6 (±1.3)
Img+Txt-BERT (20%) [Nosso método]	44.0 (±0.02)	77.2 (±0.02)
Img+Txt-BERT (40%) [Nosso método]	46.9 (±0.02)	79.3 (±0.01)
Img+Txt-BERT (60%) [Nosso método]	51.1 (±0.02)	81.9 (±0.01)
Img+Txt-BERT (80%) [Nosso método]	54.5 (±0.01)	84.4 (±0.01)
Img+Txt-BERT (100%) [Nosso método]	58.6 (±0.01)	86.4 (±0.01)

Figura 1: Tabela de resultados do método

Conclusões

Propomos um método para classificar a intenção dos autores em conjuntos de dados multimodais com modalidades incompletas. Mostramos que nossa representação em grafo é uma estrutura promissora por combinar diferentes modalidades, nas quais geramos arestas por meio de características visuais e textuais. Também mostramos que a representação baseada em BERT é competitiva para a modalidade textual, mesmo em cenários com grande porcentagem de textos perdidos. Além disso, nosso método permite propagar as características iniciais do BERT considerando a topologia do grafo.

A propagação de características de texto mostrou-se útil para representar postagens que contêm apenas imagens (modalidade textual incompleta). Nossa propagação de características não deve ser vista como um gerador de legendas, mas um método para associar o conteúdo de incorporação semântica a imagens de acordo com as postagens vizinhas. Nesse cenário, notamos que os embeddings propagados representam tópicos gerais que são promissores para determinar a intenção do autor.

Referências Bibliográficas

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep ResidualLearning for Image Recognition. InConference on Computer Vision and PatternRecognition, (CVPR). IEEE

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding. InConference of the North American Chapter of the Association for ComputationalLinguistics: Human Language Technologies (NAACL-HLT). ACL.

[3] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and BernhardSchölkopf. 2003. Learning with local and global consistency.Advances in neuralinformation processing systems (NeurIPS)(2003).