Resumo: SPICE - Semantic Propositional Image Caption Evaluation

De Matheus Loiola Pinto Curado Silva

11 de dezembro de 2023

Introdução

O artigo diz que métricas automáticas de avaliação são primariamente sensíveis à sobreposição de n-gramas, o que não é necessário nem suficiente para tarefas de simulação de julgamento humano. Assim, o autor hipotetiza que o conteúdo da semântica proposicional é um componente importante para a avaliação da legenda humana, e propõe uma nova métrica automática de avaliação de legendas chamado SPICE, *Semantic Propositional Image Caption Evaluation*.

Avaliações abrangentes em uma variedade de modelos e conjuntos de dados indicam que o SPICE captura julgamentos humanos sobre legendas geradas por modelos melhores do que outras métricas automáticas. Além disso, o SPICE pode responder perguntas como "qual gerador de legendas entende melhor as cores?"e "Os geradores de legendas podem contar?"

No artigo, apresentam essa nova métrica de avaliação automática de legendas de imagens que mede a qualidade das legendas geradas através da análise de seu conteúdo semântico. O método se assemelha muito ao julgamento humano, oferecendo a vantagem adicional de que o desempenho de qualquer modelo pode ser analisado com mais detalhes do que com outros métricas automatizadas.

Eles estimamos a qualidade da legenda por transformar legendas candidatas e de referência em um representante semântico baseado em grafos e chamado de "scene graphs". Esse "scene graph"explicitamente encoda os objetos, atributos e relações achados nas legendas da imagem, abstraindo a maior parte lexicográfica e sintática da linguagem natural no processo.

Métrica SPICE

Dado uma legenda candidata c e um conjunto de legendas referência S associadas a imagem, o objetivo é computar uma pontuação que captura a similaridade entre C e S. Como o algoritmo explora a estrutura semântica das descrições da cena, ela dá preferência a substantivos, e por isso é melhor em avaliar legendas geradas por computador. O SPICE foca exclusivamente no significado semântico.

Ele utiliza o F-score, assim, sua pontuação está limitada entre 0 e 1. O SPICE mede o quão bem os geradores de legendas recuperam objetos, atributos e as relações entre eles. Uma preocupação potencial, então, é que a métrica poderia ser 'viciada', gerando legendas que representam apenas objetos, atributos e relações, ignorando outros aspectos importantes da gramática e da sintaxe. Como o SPICE negligencia a fluência, como acontece com as métricas de n-gramas, ele assume que as legendas são bem formadas.

Avaliação da métrica

em testes realizados em uma competição do MS COCO, apenas o SPICE recompensou legendas detalhadas, enquanto outras métricas penalizaram essas legendas, como visto abaixo:

Table 1. System-level Pearson's ρ correlation between evaluation metrics and human judgments for the 15 competition entries plus human captions in the 2015 COCO Captioning Challenge [6]. SPICE more accurately reflects human judgment overall (M1–M2), and across each dimension of quality (M3–M5, representing correctness, detailedness and saliency)

	M1		**M2**		**M3**		**M4**		**M5**	
	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value	ρ	p-value
Bleu-1	0.24	(0.369)	0.29	(0.271)	0.72	(0.002)	-0.54	(0.030)	0.44	(0.091)
Bleu-4	0.05	(0.862)	0.10	(0.703)	0.58	(0.018)	-0.63	(0.010)	0.30	(0.265)
ROUGE-L	0.15	(0.590)	0.20	(0.469)	0.65	(0.006)	-0.55	(0.030)	0.38	(0.142)
METEOR	0.53	(0.036)	0.57	(0.022)	0.86	(0.000)	-0.10	(0.710)	0.74	(0.001)
CIDEr	0.43	(0.097)	0.47	(0.070)	0.81	(0.000)	-0.21	(0.430)	0.65	(0.007)
SPICE-exact	0.84	(0.000)	0.86	(0.000)	0.90	(0.000)	0.39	(0.000)	0.95	(0.000)
SPICE	**0.88**	(0.000)	**0.89**	(0.000)	**0.89**	(0.000)	**0.46**	(0.070)	**0.97**	(0.000)

M1	Percentage of captions evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1–5 (incorrect - correct).
M4	Average detail of the captions from 1–5 (lacking details - very detailed).
M5	Percentage of captions that are similar to human description.

Em relação a percepção de cores e contagem, a métrica conseguiu avaliar bem a cor, quantidade e tamanho dos atributos nas legendas. Além disso, utilizando a correlação de Kendall's em nível de legenda, a métrica não se saiu muito diferentes das outras métricas criadas, porém, SPICE aproxima o julgamento humano melhor quando agregado com outras legendas.