ALBEF

- 相比ViLT，把image和text的特征提取器弄成了非对称的形式，image的特征提取器略大
- 在modality interaction之前利用CLIP的loss 训练了text和vision提取器的特征，使其更加一致
- 利用一个慢更新的动量模型提供pseudo-targets， 解决了wed data noisy的问题。就是网上的图片的文章描述可能不确切

作者：@ry664663 | 来自：知犀思维导图