多模态

# ViLT（vision-language transformer）