We introduce the Action Transformer model for recognizing and localizing human actions in video clips. We repurpose a Transformer-style architecture to aggre