Multimodal sentiment analysis is a very actively growing field of research. A promising area of opportunity in this field is to improve the multimodal fusion mechanism. We present a novel feature fusion strategy that proceeds in a hierarchical fashion, first fusing the modalities two in two and only then fusing all three modalities. On multimodal sentiment analysis of individual utterances, our strategy outperforms conventional concatenation of features by 1%, which amounts to 5% reduction in error rate. On utterance-level multimodal sentiment analysis of multi-utterance video clips, for which current state-of-the-art techniques incorporate contextual information from other utterances of the same clip, our hierarchical fusion gives up to 2.4% (almost 10% error rate reduction) over currently used concatenation. The implementation of our method is publicly available in the form of open-source code.